GazeMIL: Attention-Based Multiple Instance Learning for Interpretable Dyslexia Detection from Multi-task Eye-Tracking Data

Jayakanth Kunhoth, Moutaz Saleh(✉), Somaya Al-Maadeed, and Younes Akbari

Department of Computer Science and Engineering, Qatar University, Doha 2713, Qatar
{j.kunhoth,moutaz.saleh,s_alali,younes.akbari}@qu.edu.qa

Abstract. Automated dyslexia detection from eye-tracking data has predominantly relied on globally aggregated oculomotor statistics or image-based representations that discard the fine-grained structure of individual fixation events. Moreover, most methods evaluate on a single reading task, overlooking complementary cognitive demands across task types, and none provides word-level localization of where dyslexic patterns emerge. In this work, we introduce GazeMIL, an attention-based multiple instance learning framework that addresses these limitations. Rather than collapsing a reading trial into a fixed-size feature vector, GazeMIL treats each trial as a bag of fixation instances encoding spatial position, duration, saccadic context, and regression behavior. Task-specific gated-attention encoders learn to assign diagnostic importance weights to individual fixations, and the resulting bag embeddings from multiple cognitively distinct reading tasks are fused for subject-level classification. Because attention weights are intrinsic to the forward pass, they map directly to word and line positions on the reading stimulus, yielding interpretable heatmaps without post-hoc attribution. Evaluated on the ETDD70 benchmark under 5-fold stratified cross-validation, the best multi-task configuration achieves 90.0% accuracy, 0.914 AUC, and 0.899 F1 score, surpassing or matching existing methods on this benchmark. Attention heatmaps reveal that dyslexic readers exhibit concentrated high-attention fixations at specific word positions, while typically developing readers show broadly distributed attention. Dyslexic readers also show significantly lower attention entropy (4.08 ± 0.65 vs. 4.61 ± 0.54, $p = 0.0005$), providing a novel biomarker of reading regularity. To our knowledge, GazeMIL is the first application of attention-based multiple instance learning to eye-tracking data for dyslexia classification.

Keywords: Dyslexia Detection · Eye-Tracking · Multiple Instance Learning · Attention Mechanism · Multi-Task Learning · Explainable AI

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 20–37, 2027.
https://doi.org/10.1007/978-3-032-30488-9_2

1 Introduction

Developmental dyslexia is a prevalent neurodevelopmental condition that impairs fluent and accurate word recognition despite normal intellectual ability and adequate instructional exposure [1,2]. Delayed identification compounds the risk of academic underachievement and emotional distress [3,4], motivating the development of objective, scalable screening instruments. Eye-tracking technology offers a compelling pathway toward this goal: by recording fixation durations, saccade kinematics, and regression patterns at millisecond resolution, it captures the moment-by-moment unfolding of reading in a non-invasive manner [5,6]. Dyslexic readers consistently exhibit prolonged fixations, elevated regression rates, and reduced saccade amplitudes relative to typically developing peers [7,8], prompting a growing body of work on machine-learning-based classification from eye-tracking data [9–11].

Despite encouraging results, two key limitations persist. First, most prior methods either operate on globally aggregated feature vectors that condense an entire reading trial into summary statistics (e.g., mean fixation duration, total regression count) for traditional classifiers [7,12,13], or convert gaze data into image representations such as scanpath visualizations or fixation heatmaps for CNN-based classification [9,10]. In both paradigms, the fine-grained sequential and spatial structure of individual fixation events is lost: a single prolonged fixation on a morphologically complex word carries different diagnostic information than one caused by momentary distraction, yet neither approach preserves this distinction. Second, most studies evaluate on a single reading task or model multiple tasks independently [14,15], forfeiting the opportunity to capture both task-invariant and task-specific markers of reading difficulty across cognitively diverse contexts [16].

We address both limitations by casting eye-tracking-based dyslexia classification as a multiple instance learning (MIL) problem. Each reading trial is treated as a bag of fixation instances, each described by a compact 10-dimensional feature vector of spatial, temporal, and oculomotor attributes. Task-specific gated-attention MIL encoders learn to weight individual fixations by diagnostic relevance, and the resulting bag embeddings from three cognitively distinct reading tasks are fused for subject-level classification. Because attention weights are intrinsic to the forward pass, they map directly back to word- and line-level positions on the stimulus, yielding interpretable heatmaps without post-hoc attribution.

Our contributions are as follows:

- We introduce, to the best of our knowledge, the first MIL formulation for eye-tracking-based dyslexia detection, replacing globally aggregated features with fixation-level instance representations.
- We propose a multi-task gated-attention MIL architecture that fuses task-specific bag embeddings to capture complementary reading-difficulty markers across diverse tasks.

- We provide intrinsic explainability via attention-derived heatmaps and entropy analysis that reveal where in the text the model focuses its diagnostic attention.
- We evaluate on the ETDD70 dataset [12] and demonstrate competitive or superior classification performance relative to existing methods.

By shifting the unit of analysis from globally aggregated trial statistics to individual fixation events and by jointly modelling multiple reading tasks through learned attention, the proposed framework offers a more granular, interpretable, and generalizable approach to automated dyslexia screening. The remainder of this paper is organized as follows. Section 2 reviews prior work. Section 3 describes the dataset. Section 4 presents the proposed methodology. Section 5 reports experimental results. Section 6 provides a discussion, and Sect. 7 concludes the paper.

2 Related Work

Early work on eye-tracking-based dyslexia detection followed a common pipeline: extract oculomotor summary statistics from a single reading task, select informative features, and train a supervised classifier. Rello and Ballesteros [9] trained SVMs and decision trees on fixation and regression features from 97 Spanish-speaking participants. Nilsson Benfatto et al. [10] applied a linear SVM with recursive feature elimination to gaze data from 185 Swedish schoolchildren. Smyrnakis et al. [13] proposed the RADAR composite score combining fixation duration, saccade length, and regression rate, while Asvestopoulou et al. [11] applied LASSO-based selection to reduce the feature space. Prabha and Bhargavi [7] compared random forests, SVMs, and gradient-boosted classifiers on compact fixation and saccade features. A common limitation across these studies is the reliance on globally aggregated statistics that compress an entire reading trial into a fixed-length vector, with evaluation limited to a single cognitive task.

To reduce dependence on manual feature engineering, subsequent work applied deep learning to gaze signal representations. Nerušil et al. [14] trained a CNN on raw gaze coordinate trajectories, while Vajs et al. [17] extracted spatiotemporal features from multiple reading trials, though all trials involved the same reading activity. Svaricek et al. [18] introduced INSIGHT, converting fixation sequences into two-dimensional visualizations classified with a ResNet18 backbone and Grad-CAM explainability, evaluating across tasks and languages but modelling each task independently. Le et al. [19] proposed a time-series grid representation preserving spatial and temporal dynamics of gaze trajectories. While these methods capture richer structure than global statistics, they still operate on fixed-size representations that discard individual fixation-level information, and tasks are analysed in isolation.

Only a few studies have examined the influence of task type on classification. El Hmimdi et al. [15] evaluated four heterogeneous oculomotor tasks and observed substantial performance variation across task types, yet each task was modelled independently. Sedmidubsky et al. [12] released the ETDD70 dataset

with three cognitively distinct reading tasks and provided single-task baselines using over 400 features, but did not pursue cross-task fusion. Nguyen et al. [20] explored feature learning strategies on multi-task data, though analysis remained at the level of global feature vectors. Across the surveyed literature, no prior work has proposed a unified framework that jointly models fixation-level information from multiple cognitively distinct tasks through learned attention. The present work bridges this gap by introducing an attention-based MIL framework that retains fixation-level granularity, learns task-specific attention distributions, fuses multi-task bag embeddings for subject-level classification, and provides intrinsic interpretability by mapping attention weights back to positions on the reading stimulus.

3 Dataset

We evaluate the proposed framework on the ETDD70 dataset [12], the only publicly available eye-tracking dataset explicitly designed for multi-task dyslexia screening in children. Alternative datasets considered include the Swedish dataset by Nilsson Benfatto et al. [10], which contains recordings from 185 children but is limited to a single reading task, and the CopCo dataset [21], which comprises recordings from adult participants under a single-task design. Neither dataset supports the multi-task analysis central to our framework, making ETDD70 the most suitable choice for this study.

The dataset comprises recordings from 70 participants, 35 diagnosed with developmental dyslexia and 35 typically developing controls, aged 9 to 10 years, with balanced gender distribution across both groups. Dyslexia diagnoses were established prior to data collection through cooperation with psychological counseling centers, which facilitated the recruitment of pupils formally diagnosed with dyslexia. Non-dyslexic participants showing no symptoms of dyslexia were recruited through counseling facilities of selected elementary schools. The diagnostic tasks were designed based on standardized paper-based dyslexia screening instruments used in Czechia, transferred to a controlled digital format for eye-tracking measurement. Full details of the diagnostic procedure, screening instruments, and participant recruitment are reported in the original dataset paper [12].

Three cognitively distinct reading tasks are included, each targeting a different aspect of reading ability. Syllable reading presents a 9×10 grid of syllables probing sub-word phonological decoding ability in isolation from lexical or semantic context. Meaningful text reading presents a 7-line narrative paragraph at an appropriate reading level for the age group, targeting reading fluency, word recognition, and comprehension simultaneously. Pseudo-Text reading presents artificial sentences composed of phonologically plausible but semantically meaningless words, requiring pure phonetic decoding in the absence of semantic support. Together the three tasks capture complementary cognitive dimensions of reading, making their joint analysis well-suited to a multi-task learning framework.

Eye movements were recorded using a stationary SMI RED 250 remote eye tracker at a sampling rate of 250 Hz under controlled laboratory conditions. The dataset provides four file types per subject per task: raw binocular gaze streams, fixation events, saccade events, and a pre-extracted metrics file containing over 400 globally aggregated oculomotor features. The proposed framework operates exclusively on the fixation and saccade event files, constructing per-fixation instance vectors as described in Sect. 4.

Each fixation event is annotated with area-of-interest (AOI) labels at two levels of granularity: `aoi_line`, identifying the text line number ($l \in \{1, \ldots, 7\}$) on which the fixation landed, and `aoi_subline`, identifying the specific word position within that line ($w \in \{1, \ldots, W\}$). The average number of fixations per subject per task ranges from approximately 100 to 150, confirming that each reading trial produces a bag of sufficient size for attention-based aggregation. Figure 1 shows the reading stimuli presented across the three tasks.



Fig. 1. Reading stimuli used in the three tasks. Left: T1 (syllable reading), a 9×10 grid of isolated syllables. Top right: T2 (meaningful text), a 7-line narrative paragraph. Bottom right: T3 (pseudo-text), a 7-line passage of phonologically plausible but semantically meaningless words.

4 Methodology

Figure 2 presents an overview of the proposed multi-task attention-based multiple instance learning framework for dyslexia classification from eye-tracking data. The framework consists of four sequential components: (1) fixation instance construction from raw eye-tracking recordings, (2) instance-level feature extraction, (3) task-specific attention-based MIL encoding, and (4) multi-task fusion and classification. An auxiliary explainability module maps learned attention weights back to word-level and line-level positions on the reading stimulus.

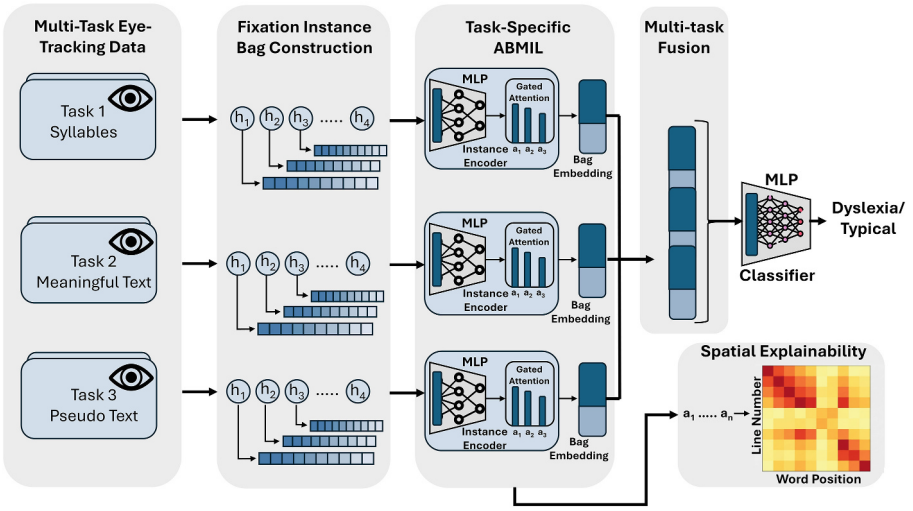

Fig. 2. Overview of the proposed GazeMIL framework. For each reading task, fixation instances are encoded by a task-specific MLP and aggregated via gated attention pooling into a fixed-size bag embedding. The task-specific bag embeddings are concatenated and passed to a classification head for binary prediction. Attention weights are mapped back to word and line positions on the reading stimulus for explainability.

4.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$ denote the dataset of N subjects, where $y_i \in \{0, 1\}$ is the subject-level label (0: typically developing, 1: dyslexic). Each subject participates in T reading tasks, $\mathcal{T} = \{T_1, T_2, T_3\}$, corresponding to syllable reading, meaningful text reading, and pseudo-text reading respectively.

For subject i and task t, the reading trial is represented as a *bag* of fixation instances:

$$\mathbf{X}_i^{(t)} = \{\mathbf{h}_{i,j}^{(t)}\}_{j=1}^{N_i^{(t)}} \tag{1}$$

where $\mathbf{h}_{i,j}^{(t)} \in \mathbb{R}^d$ denotes the j-th fixation instance vector and $N_i^{(t)}$ is the total number of fixations recorded for subject i during task t. In accordance with the MIL assumption, only the bag-level label y_i is observed during training; no instance-level annotations are available or required. The classification objective is to learn a mapping $f : \{\mathbf{X}_i^{(t)}\}_{t\in\mathcal{T}} \rightarrow y_i$ from multi-task fixation bags to subject-level dyslexia labels.

4.2 Instance Feature Extraction

Each fixation instance $\mathbf{h}_{i,j}^{(t)} \in \mathbb{R}^{10}$ is constructed by merging the fixation and saccade event files for the corresponding subject and task. The feature vector

encodes spatial, temporal, oculomotor, and text-structural properties of each individual fixation:

$$\mathbf{h}_{i,j}^{(t)} = \left[x_j,\ y_j,\ d_j,\ s_j^{\text{ampl}},\ s_j^{\text{vel}},\ r_j,\ l_j,\ w_j,\ \sigma_j^x,\ \sigma_j^y\right]^\top \tag{2}$$

where x_j and y_j denote the horizontal and vertical gaze position of fixation j in screen pixel coordinates; d_j is the fixation duration in milliseconds; s_j^{ampl} and s_j^{vel} are the amplitude and average velocity of the saccade immediately preceding fixation j; $r_j \in \{0, 1\}$ is a binary regression indicator, set to 1 if the horizontal gaze position decreased relative to the preceding fixation, indicating a backward eye movement; $l_j \in \{1, \ldots, L\}$ and $w_j \in \{1, \ldots, W\}$ denote the text line number and word position within the line, derived from the area-of-interest annotations provided in the dataset; and σ_j^x, σ_j^y represent the horizontal and vertical fixation dispersion.

4.3 Attention-Based Multiple Instance Learning

Instance Encoder. Each fixation instance $\mathbf{h}_{i,j}^{(t)}$ is independently encoded by a task-specific instance encoder $f_\theta^{(t)} : \mathbb{R}^{10} \to \mathbb{R}^E$, implemented as a two-layer multilayer perceptron (MLP) with ReLU activations and dropout regularization:

$$\mathbf{e}_{i,j}^{(t)} = f_\theta^{(t)}\left(\mathbf{h}_{i,j}^{(t)}\right) \in \mathbb{R}^E \tag{3}$$

where E denotes the embedding dimension. The set of encoded instances for subject i under task t is denoted $\mathbf{E}_i^{(t)} = \{\mathbf{e}_{i,j}^{(t)}\}_{j=1}^{N_i^{(t)}}$.

Gated Attention Pooling. To aggregate the variable-length set of instance embeddings into a fixed-size bag-level representation, we adopt the gated attention mechanism proposed by Ilse et al. [22]. For each instance embedding $\mathbf{e}_{i,j}^{(t)}$, a scalar attention weight $a_{i,j}^{(t)}$ is computed as:

$$a_{i,j}^{(t)} = \frac{\exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{U}\,\mathbf{e}_{i,j}^{(t)}\right) \odot \sigma\left(\mathbf{V}\,\mathbf{e}_{i,j}^{(t)}\right)\right)\right\}}{\sum_{k=1}^{N_i^{(t)}} \exp\left\{\mathbf{w}^\top \left(\tanh\left(\mathbf{U}\,\mathbf{e}_{i,k}^{(t)}\right) \odot \sigma\left(\mathbf{V}\,\mathbf{e}_{i,k}^{(t)}\right)\right)\right\}} \tag{4}$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{A \times E}$ are learnable weight matrices, $\mathbf{w} \in \mathbb{R}^A$ is a learnable weight vector, A denotes the attention dimension, $\tanh(\cdot)$ and $\sigma(\cdot)$ denote the hyperbolic tangent and sigmoid activation functions, respectively, and $\odot$ denotes element-wise multiplication. The softmax normalization ensures $\sum_{j=1}^{N_i^{(t)}} a_{i,j}^{(t)} = 1$.

The bag-level embedding for subject i under task t is then computed as the attention-weighted sum of instance embeddings:

$$\mathbf{z}_i^{(t)} = \sum_{j=1}^{N_i^{(t)}} a_{i,j}^{(t)}\, \mathbf{e}_{i,j}^{(t)} \in \mathbb{R}^E \tag{5}$$

4.4 Multi-task Fusion and Classification

Independent ABMIL encoders are applied to the fixation bags from each reading task, yielding task-specific bag embeddings $\mathbf{z}_i^{(T_1)}$, $\mathbf{z}_i^{(T_2)}$, and $\mathbf{z}_i^{(T_3)}$. These embeddings are concatenated to form a unified multi-task representation:

$$\mathbf{z}_i^{\text{fused}} = \mathbf{z}_i^{(T_1)} \,\|\, \mathbf{z}_i^{(T_2)} \,\|\, \mathbf{z}_i^{(T_3)} \in \mathbb{R}^{3E} \tag{6}$$

where $\|$ denotes vector concatenation. The fused representation is passed to a classification network $g_\phi : \mathbb{R}^{3E} \to \mathbb{R}$, implemented as a two-layer MLP, producing a scalar logit. The predicted probability of dyslexia is:

$$\hat{p}_i = \sigma\big(g_\phi\big(\mathbf{z}_i^{\text{fused}}\big)\big) \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The model is trained end-to-end by minimizing the binary cross-entropy loss over bag-level labels:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log \hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i)] \tag{8}$$

4.5 Explainability via Attention Mapping

A key property of the proposed framework is that the attention weights $\{a_{i,j}^{(t)}\}$ are produced as part of the model's forward pass and carry direct interpretive meaning, unlike post-hoc approximation methods such as Grad-CAM. Since each fixation instance $\mathbf{h}_{i,j}^{(t)}$ is annotated with its text line number l_j and word position w_j, the attention weights can be directly mapped back to positions on the reading stimulus.

Attention Heatmap Construction. For a given task t and subject group $\mathcal{G} \in \{\text{dyslexic}, \text{TD}\}$, the group-averaged attention weight at text position (l, w) is defined as:

$$\bar{a}^{(t)}(l, w) = \frac{1}{|\mathcal{G}|}\sum_{i \in \mathcal{G}} \frac{\sum_{j:\, l_j = l,\, w_j = w} \tilde{a}_{i,j}^{(t)}}{|\{j : l_j = l,\, w_j = w\}|} \tag{9}$$

where $\tilde{a}_{i,j}^{(t)}$ denotes the per-subject min-max normalized attention weight:

$$\tilde{a}_{i,j}^{(t)} = \frac{a_{i,j}^{(t)} - \min_k a_{i,k}^{(t)}}{\max_k a_{i,k}^{(t)} - \min_k a_{i,k}^{(t)} + \epsilon} \tag{10}$$

with ϵ being a small constant for numerical stability. Per-subject min–max normalisation ensures that no single subject dominates the group-level heatmap due to differences in absolute attention magnitude. However, it prevents the group heatmap from reflecting between-group differences in overall attention intensity, preserving only the relative spatial distribution of attention within each subject.

Attention Entropy. To quantify the regularity of reading attention patterns at the subject level, we define the attention entropy for subject i under task t as the Shannon entropy of the attention weight distribution. Since the attention weights $a_{i,j}^{(t)}$ are produced by the softmax operation in Eq. (4), they satisfy $\sum_j a_{i,j}^{(t)} = 1$ and $a_{i,j}^{(t)} \geq 0$, forming a valid probability distribution. The entropy is therefore defined directly over these weights:

$$H_i^{(t)} = -\sum_{j=1}^{N_i^{(t)}} a_{i,j}^{(t)} \log\left(a_{i,j}^{(t)}\right) \tag{11}$$

The mean attention entropy across tasks is computed as:

$$\bar{H}_i = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} H_i^{(t)} \tag{12}$$

Higher entropy indicates that attention is broadly and evenly distributed across fixations, reflecting smooth and systematic reading behavior. Lower entropy indicates that attention is concentrated on a small subset of fixations, reflecting erratic or struggling reading patterns. Group-level differences in $\bar{H}_i$ between dyslexic and typically developing readers are assessed using the non-parametric Mann-Whitney U test.

5 Evaluation and Results

All experiments follow a 5-fold stratified cross-validation protocol in which splits are performed at the subject level: all trials from a given subject are assigned exclusively to either the training or test fold, ensuring no subject appears in both partitions and preventing subject-level data leakage. For each fold, feature normalization is performed by fitting a `StandardScaler` on the training split and applying the learned statistics to both training and validation instances independently per task, preventing information leakage across folds.

Each task-specific instance encoder consists of a two-layer MLP mapping the 10-dimensional fixation input to a 64-dimensional hidden representation and then to a 32-dimensional embedding, with ReLU activations and dropout ($p = 0.3$) applied after the first layer. The gated attention module operates with an attention dimension of 16. The classification head maps the concatenated multi-task bag embeddings to a scalar logit through a two-layer MLP (hidden dimension 32, dropout $p = 0.3$). The model is trained end-to-end using the Adam optimizer with a learning rate of 10^{-3} and binary cross-entropy loss, for a maximum of 150 epochs with early stopping (patience = 20) based on validation loss. All experiments are implemented in PyTorch.

Classification performance is evaluated using accuracy, area under the ROC curve (AUC), F1 score, sensitivity, and specificity, each reported as mean $\pm$ standard deviation across the five folds.

5.1 Task Combination Ablation

To assess the contribution of individual reading tasks and their combinations, we evaluate all single-task and multi-task configurations. Table 1 reports the results.

Among single-task models, T2 (meaningful text) achieves the highest accuracy (78.6%) and AUC (0.902), suggesting that narrative reading, which engages lexical, syntactic, and semantic processing simultaneously, produces the most discriminative fixation patterns. T1 (syllable reading) and T3 (pseudo-text) perform comparably in accuracy but with higher variance, likely reflecting the more constrained cognitive demands of these tasks.

Multi-task fusion consistently improves performance over single-task models. The T2 + T3 combination achieves the best overall performance across all metrics (90.0% accuracy, 0.914 AUC, 0.899 F1), indicating strong complementarity between meaningful text reading and pseudo-text decoding. Notably, the full three-task combination (T1 + T2 + T3) achieves 85.7% accuracy, which, while superior to all single-task models, falls below the T2 + T3 configuration. This suggests that T1 (syllable grid reading), with its structurally different stimulus layout and more constrained reading pattern, may introduce representational noise that dilutes the discriminative signal captured by T2 and T3.

To assess whether the performance advantage of T2+T3 over T1+T2+T3 is statistically reliable, we conduct a one-sided paired Wilcoxon signed-rank test over the five fold-level scores, with the directional hypothesis that T2+T3 outperforms T1+T2+T3, motivated by the cognitive argument that T1's grid-based stimulus introduces qualitatively different scanning patterns that do not share the oculomotor signatures of continuous text reading. The test yields $W = 9.0$, $p = 0.41$ (accuracy) and $W = 6.0$, $p = 0.44$ (AUC), indicating no statistically significant difference at $\alpha = 0.05$. Given only five folds and 14 subjects per test fold, the test has inherently limited statistical power, and the observed mean difference (4.3% points in accuracy) should be interpreted as indicative. The preference for T2+T3 is further supported by its consistently lower variance (5.7 vs 6.4) and the cognitive complementarity between meaningful text and pseudo-text reading tasks.

Table 1. Ablation study over task combinations. T1: syllable reading, T2: meaningful text, T3: pseudo-text. Best result per metric is in **bold**.

Config	Acc. (%)	AUC	F1	Sens. (%)	Spec. (%)
T1 only	74.3 ± 10.7	0.869 ± 0.080	0.723 ± 0.134	71.4 ± 20.2	77.1 ± 11.4
T2 only	78.6 ± 13.6	0.902 ± 0.070	0.803 ± 0.109	82.9 ± 10.7	74.3 ± 26.2
T3 only	72.9 ± 14.6	0.829 ± 0.069	0.759 ± 0.094	80.0 ± 11.4	65.7 ± 35.7
T1 + T2	81.4 ± 7.3	0.878 ± 0.070	0.811 ± 0.079	82.9 ± 16.7	80.0 ± 14.6
T1 + T3	77.1 ± 7.0	0.869 ± 0.040	0.748 ± 0.106	71.4 ± 15.6	82.9 ± 5.7
T2 + T3	**90.0 ± 5.7**	**0.914 ± 0.062**	**0.899 ± 0.057**	**88.6 ± 5.7**	**91.4 ± 7.0**
T1 + T2 + T3	85.7 ± 6.4	0.894 ± 0.047	0.854 ± 0.068	85.7 ± 12.8	85.7 ± 9.0

Table 2. Ablation of MIL aggregation mechanisms. Classification accuracy (%) reported as mean ± std over 5-fold stratified cross-validation across all task configurations. The proposed gated attention results are reported in Table 1 for reference.

Config	Mean Pool	Max Pool	Std Attention
T1 only	72.9 ± 15.3	78.6 ± 7.8	78.6 ± 10.1
T4 only	81.4 ± 7.3	78.6 ± 10.1	81.4 ± 8.6
T5 only	77.1 ± 8.3	75.7 ± 9.7	77.1 ± 18.3
T1 + T4	81.4 ± 8.6	78.6 ± 9.0	82.9 ± 10.7
T1 + T5	68.6 ± 17.8	82.9 ± 7.3	74.3 ± 7.3
T4 + T5	74.3 ± 14.0	80.0 ± 9.5	82.9 ± 14.7
T1 + T4 + T5	81.4 ± 7.3	78.6 ± 12.8	80.0 ± 5.3

5.2 Ablation of MIL Aggregation Mechanisms

To evaluate the contribution of the proposed gated attention pooling, we compare it against three alternative MIL aggregation mechanisms: mean pooling, max pooling, and standard (non-gated) attention. In mean pooling, the bag-level embedding is computed as the arithmetic mean of all instance embeddings. Max pooling selects the maximum activation across instances. Standard attention uses a single-branch tanh mechanism without the sigmoid gate. All other components, the instance encoder, multi-task fusion, and classification head, remain identical across variants, and all experiments follow the same 5-fold stratified subject-level cross-validation protocol with identical random seeds.

Table 2 reports the classification accuracy across all task configurations for each aggregation variant. The proposed gated attention consistently achieves the highest accuracy on the best-performing T4+T5 configuration (90.0% ± 5.7), outperforming mean pooling (74.3% ± 14.0), max pooling (80.0% ± 9.5), and standard attention (82.9% ± 14.7) by margins of 15.7, 10.0, and 7.1% points respectively.

5.3 Comparison with Existing Methods

Table 3 compares the proposed GazeMIL framework against state-of-the-art methods evaluated on the ETDD70 dataset.

GazeMIL achieves 90.0% accuracy and 0.899 F1 score, surpassing or almost matching all published methods on this benchmark. The MLP baselines from Sedmidubsky et al. [12] achieve comparable accuracy on individual tasks (90.0–91.4%) but rely on 400+ globally aggregated statistical features per task and do not perform cross-task fusion. We followed same evaluation protocol as Sedmidubsky et al. [12], where all experiments employ 5-fold stratified subject-level cross-validation with random state 42.

The ResNet18-based INSIGHT method [18] and the time-series grid approach [19] both operate on single tasks with fixed-size representations. The LLM-based approach [20] performs multi-task fusion but requires 1,270 features and

Table 3. Comparison with published methods on the ETDD70 dataset. "Fusion" indicates whether multi-task information is combined.

Method	Fusion	Acc. (%)	F1
MLP (T1) [12]	No	91.43 ± 5.35	–
MLP (T2) [12]	No	90.00 ± 7.28	–
MLP (T3) [12]	No	88.57 ± 5.71	–
ResNet18 (T3) [18]	No	88.57 ± 8.57	–
TS-Grid + KNN [19]	No	85.7	0.86
LLM Few-shot [20]	Yes	80.6 ± 3.1	0.82 ± 0.06
GazeMIL (T2+T3)	Yes	**90.0 ± 5.7**	**0.899 ± 0.057**

achieves lower accuracy (80.6%). In contrast, GazeMIL operates on only 10 features per fixation instance, fuses information across tasks, and additionally provides intrinsic fixation-level interpretability that none of the compared methods offer.

5.4 Explainability Analysis

A key advantage of the proposed framework is that the learned attention weights provide intrinsic interpretability without requiring post-hoc attribution methods. In this section, we analyze the attention distributions produced by the best-performing T2 + T3 configuration through two complementary views: spatial attention heatmaps mapped to the reading stimulus, and attention entropy as a measure of reading regularity.

Attention Heatmaps. To examine where in the text the model concentrates its diagnostic attention, we visualize the group-averaged attention weights at two levels of granularity. For each task, a grid heatmap shows the mean attention weight at each line-by-word position, and a stimulus overlay maps attention-weighted fixation circles directly onto the reading stimulus. The grid heatmaps use absolute group-averaged attention values, while the stimulus overlays use per-subject min-max normalized weights to preserve relative within-subject patterns.

For the meaningful text task (Fig. 3), typically developing readers show high attention concentrated in the later lines (L5–L7) and early word positions (positions 1–3), with comparatively lower attention in the upper lines. Dyslexic readers show a different pattern: attention is elevated in the early-to-middle lines (L1–L3) with a prominent peak at specific word positions (e.g., position 7 in L3), and diminishes noticeably toward the later lines, particularly L5–L7. The stimulus overlay confirms this pattern, with TD fixation circles showing strong, uniformly sized circles in the lower portion of the text, while dyslexic fixation circles appear more variable in size across the stimulus.

For the pseudo-text task (Fig. 4), a distinct pattern emerges. TD readers show relatively uniform attention across lines, with moderate concentration in

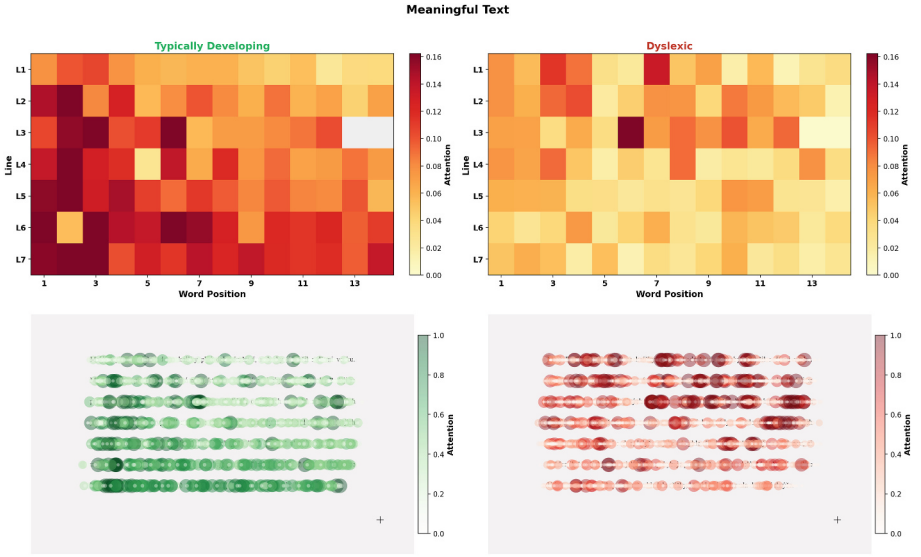

Fig. 3. Attention visualization for the meaningful text task. Top row: grid heatmap of mean attention weight per line and word position. Bottom row: attention-weighted fixation circles overlaid on the reading stimulus. Left column: typically developing readers. Right column: dyslexic readers. Grid heatmaps use absolute group-averaged values; stimulus overlays use per-subject min-max normalized weights.

the middle-to-later lines. Dyslexic readers, in contrast, exhibit markedly concentrated attention in the later lines (L5–L7), with the highest attention weights appearing in L6–L7 at both early and late word positions. The stimulus overlay for the dyslexic group shows visibly larger and darker fixation circles in the lower lines compared to the upper lines, indicating that the model identifies the strongest diagnostic signal in the later portions of the pseudo-text stimulus.

Attention Entropy. Beyond spatial localization, we quantify the overall regularity of each subject's reading attention using Shannon entropy computed over the attention weight distribution. Higher entropy indicates that attention is broadly and evenly distributed across fixations, while lower entropy indicates concentration on a smaller subset of fixations.

Figure 5 presents the distribution of mean attention entropy across subjects for both groups. Dyslexic readers show significantly lower attention entropy than typically developing peers (4.08 ± 0.65 vs. 4.61 ± 0.54, Mann-Whitney $U = 314.0$, $p = 0.0005$). The group-level separation is clearly visible in the boxplot, with minimal overlap between the two distributions. However an apparent outlier is observed in the typically developing group, likely reflecting individual variability in reading strategy or occasional lapses in sustained attention during the reading tasks.

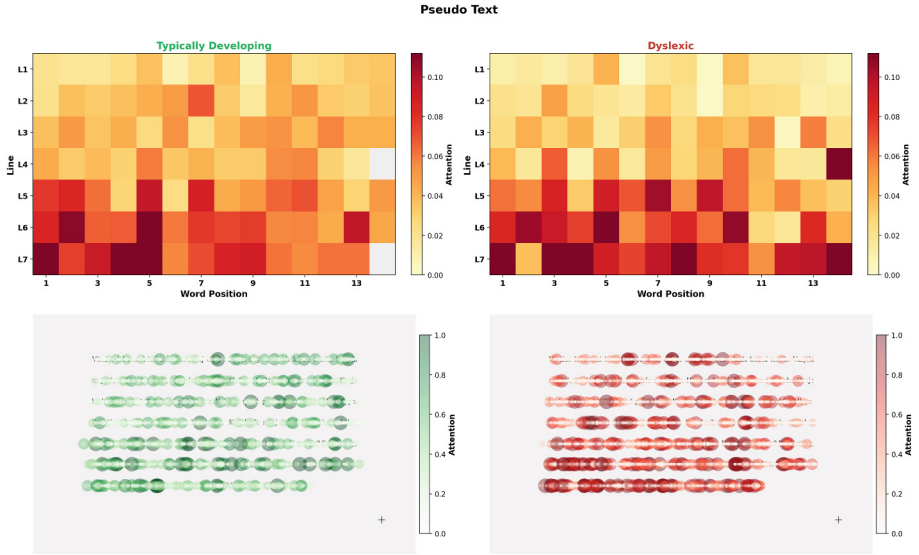

Fig. 4. Attention visualization for the pseudo-text task. Top row: grid heatmap of mean attention weight per line and word position. Bottom row: attention-weighted fixation circles overlaid on the reading stimulus. Left column: typically developing readers. Right column: dyslexic readers. Grid heatmaps use absolute group-averaged values; stimulus overlays use per-subject min-max normalized weights.

6 Discussion

The experimental results highlight several key observations regarding task complementarity, attention-based interpretability, and the trade-off between classification performance and explainability. The ablation study reveals that the combination of meaningful text (T2) and pseudo-text (T3) consistently outperforms all other configurations, including the full three-task model. This can be attributed to the complementary cognitive demands of these two tasks: meaningful text reading engages lexical access, syntactic parsing, and semantic integration simultaneously, while pseudo-text reading strips away semantic support and isolates phonological decoding ability. Together, these tasks capture both top-down language-driven processes and bottom-up decoding mechanisms, providing a richer diagnostic signal than either task alone. The inclusion of T1 (syllable grid reading) reduces performance, likely because its grid-based stimulus layout elicits qualitatively different scanning patterns that do not share the oculomotor signatures associated with reading difficulty in continuous text, introducing representational noise into the fused embedding.

The spatial attention patterns reveal qualitatively different reading strategies between groups, and these patterns vary meaningfully across tasks. In meaningful text, TD readers exhibit high attention in the later lines (L5–L7), suggesting that the model identifies sustained reading momentum as a distinguishing fea-

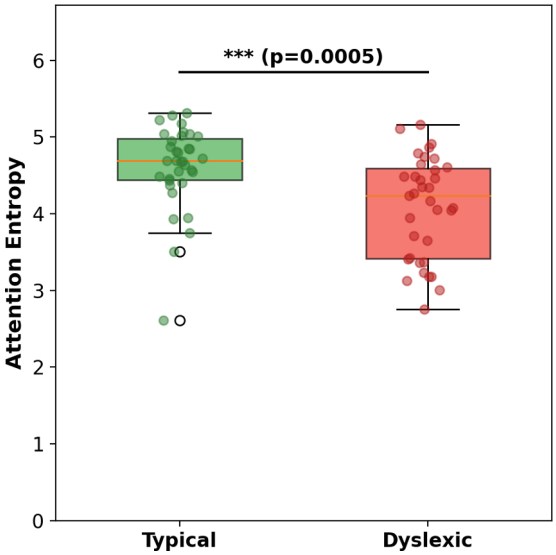

Fig. 5. Distribution of mean attention entropy for typically developing and dyslexic readers. Lower entropy reflects concentrated attention on fewer fixations. Statistical comparison via Mann-Whitney U test ($p = 0.0005$).

ture, while dyslexic readers show elevated attention in the early-to-middle lines (L1–L3) where initial decoding demands are highest. In pseudo-text, this pattern shifts: dyslexic readers show markedly concentrated attention in the later lines (L5–L7), where the absence of semantic context causes phonological decoding demands to accumulate progressively. This progressive degradation likely reflects the accumulation of phonological decoding demands in the absence of semantic support. The contrasting patterns across tasks further validate the multi-task design, as meaningful text and pseudo-text expose different aspects of reading difficulty that the model learns to exploit through task-specific attention encoders.

The significantly lower attention entropy observed in dyslexic readers ($p = 0.0005$) provides a quantitative characterization of reading regularity derived directly from the model's internal representation. High entropy indicates that many fixations contribute roughly equally to the bag-level representation, corresponding to smooth and regular reading, while low entropy indicates concentration on a small number of diagnostically salient fixations, reflecting intermittent periods of pronounced difficulty. This finding aligns with established observations that dyslexic readers produce more variable gaze patterns [7,8], but unlike conventional oculomotor statistics such as fixation duration variance, attention entropy captures this irregularity through learned diagnostic relevance, providing a complementary and potentially more sensitive measure.

A distinguishing property of the proposed framework is that its explainability is intrinsic to the model architecture. Post-hoc methods such as Grad-

CAM [18] generate attribution maps by perturbing or backpropagating through a trained model, producing explanations that approximate the decision process but are not guaranteed to reflect the model's actual internal computation. In contrast, the gated attention weights in the proposed MIL framework directly determine how individual fixation embeddings are aggregated into the bag-level representation, and in this computational sense constitute a faithful record of the aggregation process. While attention weights govern the aggregation process, we interpret them as indicators of relative fixation relevance within the model's learned representation rather than as causal explanations of diagnostic importance [23,24]. Combined with area-of-interest annotations, these weights produce spatially grounded explanations that are interpretable to clinicians and educators. Formal faithfulness evaluation through attention permutation or erase-and-evaluate analysis remains an avenue for future work.

GazeMIL achieves 90.0% accuracy, within 1.4% points of the best single-task MLP baseline (91.43%) reported by Sedmidubsky et al. [12]. This marginal difference should be interpreted in context: the MLP baseline relies on over 400 hand-crafted statistical features per task, evaluates each task independently without cross-task fusion, and provides no mechanism for identifying which fixation events or text positions contribute to the classification decision. GazeMIL surpasses all other published methods on this benchmark, including INSIGHT [18] (88.57%), the time-series grid method [19] (85.7%), and the LLM-based approach [20] (80.6%). Among these, only INSIGHT provides any form of explainability through post-hoc Grad-CAM, whereas GazeMIL offers intrinsic, fixation-level, stimulus-grounded attention maps alongside a statistically validated entropy-based biomarker.

Several limitations should be acknowledged. The ETDD70 dataset comprises 70 participants from a single site recorded in Czech, and generalizability to other languages, age groups, and recording conditions remains to be established. The current fusion strategy uses simple concatenation; more sophisticated mechanisms such as cross-task attention or learned task weighting may further improve performance. Additionally, the present analysis does not link attention patterns to specific linguistic properties of the stimulus words such as word frequency or morphological complexity, which would provide deeper cognitive insights. Finally, the evaluation is cross-sectional; longitudinal studies examining whether attention entropy changes in response to reading interventions could establish its clinical utility as a progress monitoring tool.

7 Conclusion

We presented GazeMIL, the first application of attention-based multiple instance learning to eye-tracking data for dyslexia classification. The framework replaces globally aggregated feature representations with fixation-level instance modelling, jointly processes multiple cognitively distinct reading tasks through task-specific gated-attention encoders, and provides intrinsic stimulus-grounded explainability without relying on post-hoc attribution methods. On the ETDD70

benchmark, GazeMIL achieves 90.0% accuracy and 0.914 AUC while operating on a compact per-fixation representation, demonstrating that preserving fixation-level granularity and exploiting multi-task complementarity offers a viable alternative to heavily engineered feature pipelines. The attention entropy measure introduced in this work provides a statistically significant ($p = 0.0005$) quantitative marker distinguishing dyslexic from typically developing reading patterns. Future work will target validation across languages and datasets, explore richer fusion strategies such as cross-task attention, correlate attention patterns with linguistic properties of stimulus words, and investigate longitudinal tracking of attention entropy to assess its utility for monitoring reading intervention progress.

Acknowledgment. This publication is supported by Qatar University High Impact Grant QUHI-CENG-25/26-665. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of Qatar University.

Data Availibility Statement. The ETDD70 dataset used in this study is publicly available at https://doi.org/10.5281/zenodo.13332134 [12]. The complete source code for the proposed method is available at https://github.com/Jayakanth21/GazeMIL.

Ethics Statement. This study uses the publicly available ETDD70 dataset [12]; all ethics approvals, parental consent, and GDPR compliance for eye-tracking data of minors were established by the original dataset authors [12].

References

1. Ahire, N., Awale, R., Patnaik, S., Wagh, A.: A comprehensive review of machine learning approaches for dyslexia diagnosis. Multimedia Tools Appl. **82**, 13557–13577 (2023)
2. Valdois, S.: The visual-attention span deficit in developmental dyslexia: review of evidence for a visual-attention-based deficit. Dyslexia **28**, 397–415 (2022)
3. Haft, S.L., Greiner de Magalhães, C., Hoeft, F.: A systematic review of the consequences of stigma and stereotype threat for individuals with specific learning disabilities. J. Learn. Disabil. **56**, 193–209 (2023)
4. Kunhoth, J., Al-Maadeed, S., Kunhoth, S., Akbari, Y., Saleh, M.: Automated systems for diagnosis of dysgraphia in children: a survey and novel framework. Int. J. Doc. Anal. Recogn. (IJDAR) **27**, 707–735 (2024)
5. Liu, Z., Yang, Z., Gu, Y., Liu, H., Wang, P.: The effectiveness of eye tracking in the diagnosis of cognitive disorders: a systematic review and meta-analysis. PLoS ONE **16**, e0254059 (2021)
6. Nguyen, T.K.C., Le, D.D., Le, T.H., Nguyen, T.C.H., Ngo, T.D.: The use of eye tracking in supporting individuals with dyslexia: a review. Disabil. Rehabil. Assist. Technol. **20**, 1183–1198 (2025)
7. Prabha, A.J., Bhargavi, R.: Predictive model for dyslexia from fixations and saccadic eye movement events. Comput. Methods Programs Biomed. **195**, 105538 (2020)

8. Raatikainen, P., Hautala, J., Loberg, O., Kärkkäinen, T., Leppänen, P., Nieminen, P.: Detection of developmental dyslexia with machine learning using eye movement data. Array **12**, 100087 (2021)
9. Rello, L., Ballesteros, M.: Detecting readers with dyslexia using machine learning with eye tracking measures. In: Proceedings of 12th International Web for All Conference, pp. 1–8 (2015)
10. Benfatto, M.N., et al.: Screening for dyslexia using eye tracking during reading. PLoS ONE **11**, e0165508 (2016)
11. Asvestopoulou, T., et al.: DyslexML: screening tool for dyslexia using machine learning. arXiv preprint arXiv:1903.06274 (2019)
12. Sedmidubsky, J., Dostalova, N., Svaricek, R., Culemann, W.: ETDD70: eye-tracking dataset for classification of dyslexia using AI-based methods. In: Proceedings of International Conference on Similarity Search and Applications, pp. 34–48 (2024)
13. Smyrnakis, I., et al.: RADAR: a novel fast-screening method for reading difficulties with special focus on dyslexia. PLoS ONE **12**, e0182597 (2017)
14. Nerušil, B., Polec, J., Škunda, J., Kačur, J.: Eye tracking based dyslexia detection using a holistic approach. Sci. Rep. **11**, 15687 (2021)
15. El Hmimdi, A.E., Ward, L.M., Palpanas, T., Kapoula, Z.: Predicting dyslexia and reading speed in adolescents from eye movements in reading and non-reading tasks: a machine learning approach. Brain Sci. **11**, 1337 (2021)
16. Toki, E.I.: Using eye-tracking to assess dyslexia: a systematic review of emerging evidence. Educ. Sci. **14**, 1256 (2024)
17. Vajs, I., Ković, V., Papić, T., Savić, A.M., Janković, M.M.: Spatiotemporal eye-tracking feature set for improved recognition of dyslexic reading patterns in children. Sensors **22**, 4900 (2022)
18. Svaricek, R., Dostalova, N., Sedmidubsky, J., Cernek, A.: Insight: combining fixation visualisations and residual neural networks for dyslexia classification from eye-tracking data. Dyslexia **31**, e1801 (2025)
19. Le, L., Nguyen, Q.T., Duong-Trung, N., Williams-King, D.: Time-series grid encoding of eye-tracking data for explainable AI in dyslexia detection. In: Proceedings of Symposium on Eye Tracking Research and Applications, pp. 1–3 (2025)
20. Nguyen, Q.-T., et al.: Learning disorder detection using eye tracking: are large language models better than machine learning?. In: Proceedings of Symposium on Eye Tracking Research and Applications, pp. 1–13 (2025)
21. Hollenstein, N., Barrett, M., Björnsdóttir, M.: The Copenhagen corpus of eye tracking recordings from natural reading of Danish texts. arXiv preprint arXiv:2204.13311 (2022)
22. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of International Conference on Machine Learning (ICML), pp. 2127–2136 (2018)
23. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of 2019 Conference on North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 3543–3556 (2019)
24. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of 2019 Conference Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 11–20 (2019)