

# Segment-Level Risk Discovery in Online Handwriting for Alzheimer’s Disease Detection

Changqing Gong, Huafeng Qin and Mounîm A. El-Yacoubi

**Abstract**—Online handwriting provides a non-invasive and low-cost behavioral biomarker for Alzheimer’s disease (AD) detection, as it reflects both cognitive planning and fine motor control. Existing handwriting-based AD detection methods usually rely on global trajectory features or whole-sample representations, which can be strongly affected by individual writing style, task-specific variation, and acquisition noise. In this paper, we propose NormPaST-Risk, a healthy-normative Paper-Air selective trajectory state-space risk network for interpretable AD detection from online handwriting. Instead of treating the entire trajectory as a single holistic representation, our method reformulates AD handwriting detection as local disease-relevant segment discovery. Specifically, a multi-scale temporal encoder captures stroke dynamics at different temporal resolutions, while a selective Paper-Air state-space encoder models long-range handwriting progression and distinguishes on-paper motor execution from in-air planning and transition behaviors. To explicitly characterize abnormal deviations, a healthy normative branch learns normal handwriting dynamics from healthy controls, and a task-aware multi-expert segment-risk module estimates segment-level AD risk calibrated by hidden-state changes and normative deviations. A weakly supervised segment-level objective further enables high-risk segment discovery without manual segment annotations. Experiments on the DARWIN benchmark demonstrate that the proposed framework achieves superior AD/HC classification performance compared with existing methods. Moreover, the discovered high-risk segments can be projected back to the original handwriting trajectory, providing interpretable evidence associated with AD-related handwriting variations.

**Index Terms**—Alzheimer’s disease, handwriting analysis, digital biomarkers, machine learning, computer-aided diagnosis.

## I. INTRODUCTION

Alzheimer’s disease (AD), the most common cause of dementia, is a progressive neurodegenerative disorder characterized by cognitive decline affecting memory, reasoning, and daily functioning [1]. Early identification is increasingly important for timely clinical intervention and disease management [2, 3]. However, established biomarker modalities, such as amyloid positron emission tomography (PET) and cerebrospinal fluid (CSF) assays, remain constrained in large-scale screening by their cost, limited accessibility, and/or invasiveness [4, 5]. Neuropsychological assessments, such as

the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA), are more accessible, but their performance can be influenced by cultural and educational factors, while repeated assessments may also be affected by practice effects and rater dependence [6, 7]. Therefore, there remains a strong need for accessible, objective, and low-cost approaches to support large-scale AD screening.

To this end, a variety of behavioral signals, including eye movements, speech, gait, and motor activity, have been investigated as potential digital markers of cognitive decline and dementia [8, 9]. Among these behavioral modalities, handwriting is particularly promising because it requires the coordinated engagement of cognitive, visuospatial, linguistic, and fine motor processes [10, 11]. With digital tablets, online handwriting tasks can capture fine-grained spatiotemporal dynamics, including pen trajectories, pressure, velocity, pauses, and in-air movements, providing an unobtrusive window into cognitive-motor function [12, 13]. Previous studies have reported alterations in handwriting kinematics and dynamics in individuals with AD and mild cognitive impairment, including changes in movement timing, fluency, pressure, and in-air behavior [11, 12, 13, 14].

Handwriting-based AD detection has traditionally relied on manually engineered kinematic, spatial, and pressure-related features combined with conventional machine learning classifiers [15, 16, 17]. With the development of deep learning, recent studies have increasingly explored learned handwriting representations, including sequential representation learning [12], convolutional neural networks [18], deep transfer learning [19], and multimodal Transformer architectures integrating 1D dynamic signals with 2D handwriting images [20]. Beyond task-level classification, recent research has also investigated more fine-grained handwriting characteristics, including the effects of word semantics and phonology on handwriting dynamics [14] and individual stroke-level representations for AD prediction [21]. These developments are consistent with broader evidence that AD-related changes can manifest across motor, visuospatial, and linguistic aspects of handwriting [11].

Building on these advances toward finer-grained handwriting analysis, an important question is whether disease-related handwriting evidence can be more effectively captured at the local trajectory level. Such evidence may not be uniformly distributed throughout an entire handwriting sequence, while global or task-level representations can be influenced by non-disease-related variations, including individual writing style, habitual pen usage, task-specific execution strategies, and acquisition noise. These factors may dilute or obscure subtle local abnormalities. Preserving local trajectory structure may

Changqing Gong and Mounîm A. El-Yacoubi are with Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France (e-mail: changqing.gong@telecom-sudparis.eu; mounim.el\_yacoubi@telecom-sudparis.eu).

Huafeng Qin is with the School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing 400067, China (e-mail: qinhuafengfeng@163.com).

Manuscript received xx; revised xx; accepted xx. This work was supported in part by xxx, and in part by xxx. We also would like to thank Telecom SudParis and Institut Polytechnique de Paris for their partial support (Corresponding author: Huafeng Qin, Mounîm A. El-Yacoubi.)

therefore provide a more direct way to identify disease-relevant patterns while reducing the influence of irrelevant global variations. Similar observations have also been reported in recent time-series studies, where modeling informative local patches or discriminative subsequences has shown advantages over relying exclusively on global representations [22, 23].

Learning such localized disease handwriting evidence, however, is challenging because only a single label is available for each complete handwriting trajectory, while no segment-level annotations indicate which local regions contain disease-related abnormalities. We therefore adopt weakly supervised segment-level risk learning, in which local segment risks are learned indirectly from trajectory-level labels through top- $k$  aggregation. This encourages the model to focus on a subset of informative handwriting regions that are more relevant to disease discrimination. To address these challenges, we propose NormPaST-Risk, a segment-level trajectory modeling framework for interpretable handwriting-based AD detection. The model first partitions online handwriting trajectories into state-consistent local segments that preserve Paper/Air boundaries, and a multi-scale temporal encoder captures short- and mid-range kinematic dynamics. A Selective Paper-Air state-space encoder further models long-range dependencies while explicitly incorporating pen-contact states to distinguish on-paper motor execution from in-air planning and transition behavior. A healthy normative branch learns normal handwriting dynamics from healthy controls, while a task-aware multi-expert segment-risk module estimates local disease risks and identifies diagnostically informative segments. Together, these components integrate local abnormality discovery with long-range trajectory modeling for interpretable AD detection. The main contributions of this work are summarized as follows:

- We introduce segment-level disease-risk modeling into online handwriting-based AD classification. Instead of treating all trajectory regions as equally informative, the proposed framework explicitly estimates segment-level disease relevance and uses the learned risks to adaptively weight local representations while retaining the complete trajectory context.
- We propose NormPaST-Risk, a novel trajectory model that integrates multi-scale temporal encoding, Paper-Air state modulation, selective state-space modeling, a healthy normative branch, and task-aware multi-expert segment-risk learning. The framework captures both local and long-range handwriting dynamics while enabling interpretable localization of disease-relevant segments.
- Extensive experiments demonstrate the effectiveness of NormPaST-Risk. The proposed method consistently outperforms representative machine learning and sequence-modeling baselines, while ablation and segment-deletion studies further validate the contributions of its key components and the discriminative value of the discovered high-risk handwriting segments.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents NormPaST-Risk, Section 4 reports the experiments, and Section 5 concludes the paper.

## II. RELATED WORK

Dynamic handwriting analysis has been widely investigated as a non-invasive means of characterizing cognitive and motor alterations associated with neurodegenerative disorders [24, 11]. Early handwriting-based AD detection mainly relied on manually engineered kinematic and spatial descriptors, such as writing duration, in-air and on-paper time, velocity, acceleration, pressure, jerk, and trajectory statistics, followed by conventional machine learning classifiers [15, 16]. The release of the DARWIN dataset further enabled systematic evaluation across multiple handwriting and drawing tasks [16]. Although these handcrafted representations are interpretable, they typically summarize an entire task and may therefore obscure subtle local abnormalities.

More recent studies have shifted toward learned representations of handwriting dynamics. El-Yacoubi et al. introduced sequential representation learning to capture temporal handwriting behavior beyond global kinematic statistics [12]. Cilia et al. encoded online handwriting dynamics into synthetic images for deep transfer learning [19], while Dao et al. directly modeled handwriting sequences using a 1D convolutional neural network [25]. Erdogmus and Kabakus further explored convolutional representations derived from handwriting signals [18]. More recently, attention-based and multimodal architectures have been investigated: Kang et al. employed self-attention across multiple handwriting tasks [26], whereas Gong et al. integrated 1D dynamic signals with 2D handwriting images using a hybrid Transformer [20]. Despite increasingly powerful representation learning, these methods primarily focus on task- or sample-level prediction.

Fine-grained handwriting dynamics have consequently received increasing attention. Online handwriting contains both on-paper and in-air movements, whose temporal and kinematic characteristics can provide complementary information about handwriting behavior [11, 13]. Cilia et al. further investigated how linguistic factors such as word semantics and phonology affect handwriting dynamics in AD [14]. More importantly, Nardone et al. analyzed individual handwriting strokes rather than compressing each task into a single aggregated feature vector, demonstrating the benefit of preserving fine-grained movement information [21]. Explainability studies have also attempted to identify discriminative temporal patterns in online handwriting; for example, Sweidan et al. interpreted CNN predictions and revealed localized movement characteristics associated with AD [27]. Nevertheless, existing fine-grained approaches mainly rely on predefined stroke units or post-hoc interpretation rather than directly learning disease relevance for local trajectory segments.

Online handwriting is also inherently sequential, requiring the modeling of both short-term motor variations and long-range writing dependencies. State-space models (SSMs) provide an efficient framework for long-sequence modeling. Structured and diagonal SSMs enable efficient modeling of long-range dependencies [28, 29, 30], while selective SSMs introduce input-dependent state propagation to adapt sequence dynamics to the current input [31]. However, SSM-based modeling remains largely unexplored for handwriting-based

AD detection. In particular, existing handwriting approaches typically summarize Paper/Air behavior through temporal or kinematic descriptors rather than explicitly using pen-contact state to modulate long-range sequence dynamics.

Overall, handwriting-based AD detection has progressed from handcrafted task-level descriptors to deep sequential, attention-based, multimodal, and stroke-level representations. However, three limitations remain. First, most methods perform prediction using global task representations or predefined stroke units rather than explicitly learning disease-relevant local segments from trajectory-level supervision. Second, Paper/Air information is generally represented through derived features rather than explicitly modeled as contextual state governing trajectory dynamics. Third, localized disease-risk discovery, long-range sequence modeling, and deviations from healthy handwriting dynamics have not yet been jointly modeled within a unified framework. These limitations motivate NormPaST-Risk, which constructs state-consistent Paper-Air segments, models long-range trajectory dynamics using a selective state-space encoder, learns healthy normative dynamics, and discovers high-risk local handwriting segments under weak supervision.

### III. PROPOSED APPROACH

Handwriting-based Alzheimer's disease (AD) detection is challenging because disease-related abnormalities may be sparse, localized, and task-dependent. A complete online handwriting trajectory contains substantial variations arising from individual writing style, task-specific execution, and acquisition noise, whereas discriminative AD-related evidence may appear only in short local segments, such as hesitation, unstable stroke execution, abnormal in-air movement, or irregular Paper-Air transitions. Therefore, instead of representing a handwriting sample solely by a global trajectory descriptor, we propose NormPaST-Risk, a healthy-normative Paper-Air selective trajectory state-space risk network for interpretable AD detection from online handwriting.

The proposed framework is motivated by four considerations. First, online handwriting is a long temporal sequence with multi-scale dynamics, requiring both short-range stroke dynamics and long-range writing progression to be modeled. Second, on-paper and in-air movements convey different behavioral information: Paper trajectories primarily reflect executed motor behavior, whereas Air trajectories may encode planning, hesitation, and cognitive-motor transitions. Third, local AD-related abnormalities can be more explicitly characterized by measuring deviations from healthy handwriting dynamics. Fourth, AD-related manifestations may vary across handwriting tasks and local trajectory segments, motivating task-aware multi-expert risk modeling to capture heterogeneous abnormal patterns and localize disease-relevant handwriting evidence. As shown in Fig.1, NormPaST-Risk consists of six components: (1) a multi-scale encoder, (2) a temporal segment aggregator, (3) a Selective Paper-Air SSM, (4) a healthy normative branch, (5) a task-aware multi-expert segment-risk module, and (6) a classification head with weakly supervised segment-risk learning.

#### A. Input Representation

Given an online handwriting sample, we denote the raw trajectory as

$$\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T_r}, \quad \mathbf{x}_t \in \mathbb{R}^F, \quad (1)$$

where  $T_r$  is the raw sequence length and  $F$  is the feature dimension. We define the raw valid mask, paper mask, and air mask as  $\mathbf{m}, \mathbf{m}^p, \mathbf{m}^a \in \{0, 1\}^{T_r}$ , where  $m_t = 1$  indicates a valid raw point,  $m_t^p = 1$  indicates a pen-down point, and  $m_t^a = 1$  indicates a pen-up point. For every valid point, the paper and air states are mutually exclusive.

#### B. Multi-Scale Trajectory Encoder

AD-related handwriting abnormalities may occur at different temporal scales. For example, short receptive fields can capture local stroke instability, while longer receptive fields can capture hesitation and movement organization. Therefore, we first encode the raw trajectory with a multi-scale temporal encoder.

Each raw point is projected into a latent space and augmented with positional encoding:

$$\mathbf{h}_t^0 = \phi_{\text{in}}(\mathbf{x}_t) + \mathbf{p}_t, \quad (2)$$

where  $\phi_{\text{in}}(\cdot)$  is a linear projection and  $\mathbf{p}_t$  is the positional encoding. Then,  $Q$  temporal convolutional streams with different kernel sizes and dilation rates are applied:

$$\mathbf{H}^q = \mathcal{E}_q(\mathbf{H}^0), \quad q = 1, \dots, Q, \quad (3)$$

where  $\mathbf{H}^q = \{\mathbf{h}_t^q\}_{t=1}^{T_r}$  and  $\mathcal{E}_q(\cdot)$  denotes the  $q$ -th temporal stream. Each stream preserves the original sequence length.

To adaptively fuse different scales at each raw point, we compute scale weights:

$$a_{t,q} = \frac{\exp(s(\mathbf{h}_t^q))}{\sum_{l=1}^Q \exp(s(\mathbf{h}_t^l))}, \quad (4)$$

where  $s(\cdot)$  is a learnable scoring function. The fused point-level representation is

$$\bar{\mathbf{h}}_t = \sum_{q=1}^Q a_{t,q} \mathbf{h}_t^q. \quad (5)$$

This operation produces a point-wise sequence

$$\bar{\mathbf{H}} = \{\bar{\mathbf{h}}_t\}_{t=1}^{T_r}, \quad \bar{\mathbf{h}}_t \in \mathbb{R}^D. \quad (6)$$

#### C. Temporal Segment Aggregation

Directly feeding all raw points into the sequence encoder can be inefficient and sensitive to point-level noise. However, naive fixed-stride downsampling may merge pen-down and pen-up states into the same token, which weakens the meaning of paper-air modeling. To address this issue, we introduce a state-consistent temporal segment aggregator.

Let  $\gamma$  be the maximum number of raw points represented by one segment token. The aggregator partitions the valid raw trajectory into segment intervals:

$$\mathcal{I} = \{I_j\}_{j=1}^{T_f}, \quad I_j = [s_j, e_j), \quad (7)$$

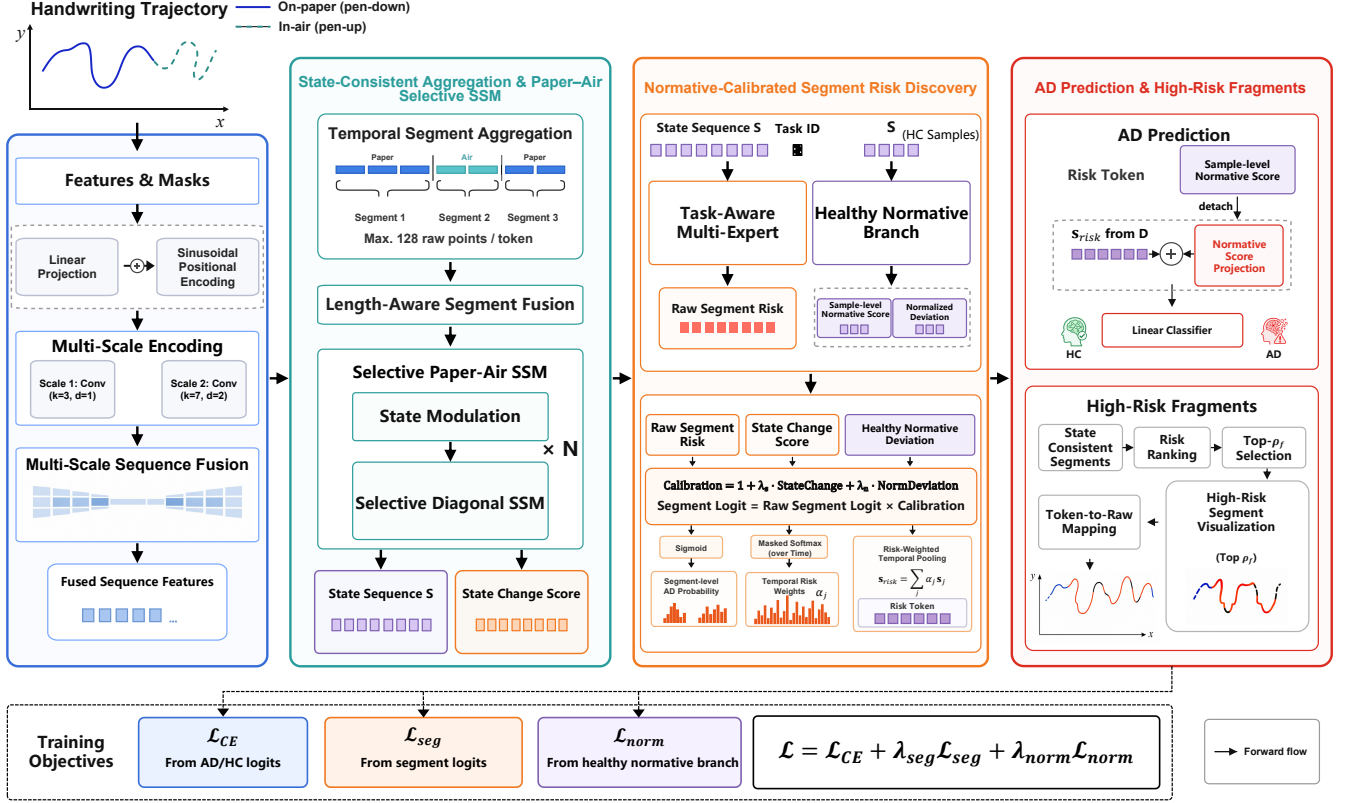


Fig. 1. Overall architecture of NormPaST-Risk. The framework consists of online handwriting representation, multi-scale trajectory encoding, state-aware segmentation and Paper-Air selective state-space modeling, normative-calibrated segment-risk discovery, and AD prediction.

where  $T_f$  is the number of segment tokens. Each interval satisfies

$$1 \leq \ell_j = e_j - s_j \leq \gamma. \quad (8)$$

A Paper/Air state transition always terminates the current interval. Therefore, all raw points inside the same interval share the same pen state.

The raw trajectory is aggregated using mean pooling:

$$\bar{\mathbf{x}}_j = \frac{1}{\ell_j} \sum_{t=s_j}^{e_j-1} \mathbf{x}_t. \quad (9)$$

The fused point-level features are aggregated using the same segment layout:

$$\boldsymbol{\mu}_j = \frac{1}{\ell_j} \sum_{t=s_j}^{e_j-1} \bar{\mathbf{h}}_t. \quad (10)$$

The same intervals are also used to aggregate scale weights and to construct segment-level valid, paper, and air masks. Thus, every downstream token is aligned with a true raw-point interval and belongs to exactly one Paper/Air state.

#### D. Length-Aware Segment Token Fusion

Mean pooling provides a stable gradient path for all raw points inside a segment, but it can dilute salient local responses. Therefore, we additionally compute a channel-wise max-pooled token:

$$\boldsymbol{\nu}_j = \max_{t \in I_j} \bar{\mathbf{h}}_t, \quad (11)$$

where the maximum is taken independently for each channel.

The mean token and max token are combined by a learnable channel-wise convex mixture:

$$\mathbf{f}_j = \boldsymbol{\mu}_j + \mathbf{w} \odot (\boldsymbol{\nu}_j - \boldsymbol{\mu}_j), \quad \mathbf{w} = \sigma(\boldsymbol{\theta}), \quad (12)$$

where  $\boldsymbol{\theta} \in \mathbb{R}^D$  is a learnable parameter,  $\mathbf{w} \in (0, 1)^D$ , and  $\odot$  denotes element-wise multiplication. This design keeps the stable mean-pooling path while giving salient responses an additional length-independent path.

Because segments can have different true lengths, especially when a segment is terminated early by a Paper/Air transition, we encode the duration of each segment using two descriptors:

$$\ell_j^{\text{lin}} = \frac{\ell_j}{\gamma}, \quad \ell_j^{\text{log}} = \frac{\log(1 + \ell_j)}{\log(1 + \gamma)}. \quad (13)$$

The final segment token is

$$\mathbf{z}_j = \mathbf{f}_j + \phi_\ell([\ell_j^{\text{lin}}, \ell_j^{\text{log}}]), \quad (14)$$

where  $\phi_\ell(\cdot)$  is a learnable duration projection. The resulting segment-token sequence is

$$\mathbf{Z} = \{\mathbf{z}_j\}_{j=1}^{T_f}, \quad \mathbf{z}_j \in \mathbb{R}^D. \quad (15)$$

#### E. Selective Paper-Air State-Space Encoder

Dynamic handwriting inherently couples motor execution with higher-level cognitive processes such as planning and

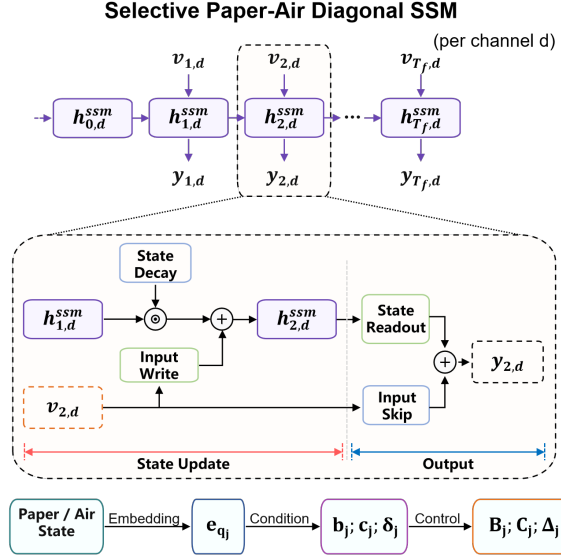


Fig. 2. Illustration of the selective diagonal SSM.

coordination, while on-paper and in-air movements exhibit distinct behavioral characteristics and have shown different sensitivity to cognitive impairment [24, 32]. This motivates us to treat the observed Paper/Air state as contextual information rather than as an ordinary kinematic feature. Inspired by feature-wise conditional modulation [33], we use layer-specific Paper-Air modulation to adapt intermediate trajectory representations to the observed pen state at different abstraction levels. The resulting state-aware sequence is subsequently modeled by selective state-space dynamics. Structured and diagonal SSMs have demonstrated strong capability for long-range sequence modeling [28, 29, 30], while selective SSMs further enable input-dependent state propagation through dynamically parameterized state-space components [31]. These observations motivate a hierarchical encoder in which Paper/Air state modulates the representation space and selective SSMs adapt the temporal state evolution.

The segment-token sequence still contains long-range temporal dependencies. To model them efficiently, we use a Selective Paper-Air state-space encoder composed of  $L$  stacked selective SSM blocks. At each layer  $l$ , the current token representation is adaptively modulated according to its observed Paper/Air state:

$$\tilde{\mathbf{h}}_j^l = (\mathbf{h}_j^{l-1} + \mathbf{e}_{q_j}^l) \odot (0.75 + 0.5\mathbf{g}_j^l), \quad (16)$$

where  $\mathbf{h}_j^{(0)} = \mathbf{z}_j$ ,  $q_j$  denotes the observed Paper/Air state,  $\mathbf{e}_{q_j}^l \in \mathbb{R}^D$  is its layer-specific learnable latent embedding, and  $\mathbf{g}_j^l \in (0, 1)^D$  is a layer-specific Paper-Air modulation gate predicted from the concatenation of  $\mathbf{h}_j^{l-1}$  and  $\mathbf{e}_{q_j}^l$ . Although the observed state  $q_j$  remains unchanged across layers, its latent representation and modulation effect are learned independently at different representation levels.

Inside the  $l$ -th selective diagonal SSM block, a gated input token is computed as

$$\mathbf{v}_j^l = \mathbf{o}_j^l \odot \sigma(\mathbf{g}_j^{s,l}), \quad [\mathbf{o}_j^l; \mathbf{g}_j^{s,l}] = \phi_{\text{in}}^{s,l}(\text{LN}(\tilde{\mathbf{h}}_j^l)). \quad (17)$$

where  $\phi_{\text{in}}^{s,l}$  denotes a learnable linear input projection in the  $l$ -th selective SSM layer.

The current token also generates channel-wise selection factors:

$$[\mathbf{b}_j^l; \mathbf{c}_j^l; \delta_j^l] = \phi_{\text{sel}}^l(\text{LN}(\tilde{\mathbf{h}}_j^l)), \quad (18)$$

where  $\mathbf{b}_j^l, \mathbf{c}_j^l, \delta_j^l \in \mathbb{R}^D$ ,  $\phi_{\text{sel}}^l$  denotes a learnable linear projection in the  $l$ -th selective SSM layer. The selection factors are transformed as

$$\mathbf{b}_j^{+,l} = 2\sigma(\mathbf{b}_j^l), \quad \mathbf{c}_j^{+,l} = 2\sigma(\mathbf{c}_j^l), \quad (19)$$

$$\Delta_j^l = \text{Softplus}(\eta_{\Delta}^l + 0.5 \tanh(\delta_j^l)), \quad (20)$$

where  $\eta_{\Delta}^l \in \mathbb{R}^D$  is a learnable channel-wise base parameter for the SSM step size. The diagonal transition parameter is constrained to be stable:

$$\mathbf{A}^l = -\exp(\mathbf{A}_{\log}^l). \quad (21)$$

For channel  $d$ , the SSM state is  $\mathbf{h}_{j,d}^{\text{ssm},l} \in \mathbb{R}^{N_s}$ , where  $N_s$  is the state dimension. The selective diagonal recurrence is

$$\mathbf{h}_{j,d}^{\text{ssm},l} = \exp(\mathbf{A}_d^l \Delta_{j,d}^l) \odot \mathbf{h}_{j-1,d}^{\text{ssm},l} + (\mathbf{b}_{j,d}^{+,l} \mathbf{B}_d^l) v_{j,d}^l, \quad (22)$$

$$y_{j,d}^l = (\mathbf{c}_{j,d}^{+,l} \mathbf{C}_d^l)^\top \mathbf{h}_{j,d}^{\text{ssm},l} + D_d^l v_{j,d}^l, \quad (23)$$

where  $\mathbf{A}_d^l, \mathbf{B}_d^l, \mathbf{C}_d^l \in \mathbb{R}^{N_s}$  are channel-wise SSM parameters, and  $D_d^l$  is a learnable direct feedthrough parameter.

Finally, the channel-wise SSM responses are aggregated and projected to obtain the SSM output  $\mathbf{y}_j^l \in \mathbb{R}^D$ . The resulting output is then combined with the input of the current layer through a residual connection followed by layer normalization. The operations of the  $l$ -th Selective Paper-Air diagonal SSM layer can be summarized as

$$\mathbf{h}_j^l = \text{LN}(\mathbf{h}_j^{l-1} + \mathcal{D}_{\text{SSM}}^l(\mathcal{M}_{\text{PA}}^l(\mathbf{h}_j^{l-1}, q_j))). \quad (24)$$

Here,  $\mathcal{M}_{\text{PA}}^l$  represents the Paper-Air modulation, while  $\mathcal{D}_{\text{SSM}}^l$  represents the selective diagonal SSM transformation, including state propagation, input writing, state readout, and output projection.

After the final layer, we obtain the context-aware state sequence

$$\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^{T_f}, \quad \mathbf{s}_j = \mathbf{h}_j^L \in \mathbb{R}^D. \quad (25)$$

To capture abrupt hidden-state changes, we compute a normalized state-change score:

$$\chi_j = \frac{\|\mathbf{s}_j - \mathbf{s}_{j-1}\|_2}{\max(\max_k \|\mathbf{s}_k - \mathbf{s}_{k-1}\|_2, \epsilon)}, \quad j \geq 2, \quad (26)$$

This score is used only as a calibration factor for learned segment risk, rather than as an independent disease indicator.

### F. Healthy Normative Branch

To explicitly characterize healthy handwriting patterns, we introduce a healthy normative branch, as shown in Fig.3 (a). Given the contextual state sequence  $\mathbf{S}$ , the branch estimates the subsequent aggregated trajectory token as

$$\hat{\mathbf{x}}_{j+1} = f_{\text{norm}}(\mathbf{s}_j), \quad j = 1, \dots, T_f - 1. \quad (27)$$

Here,  $\mathbf{s}_j$  is the context-aware state representation of the  $j$ -th segment, and the estimation target is the subsequent aggregated trajectory token  $\bar{\mathbf{x}}_{j+1}$ . To ensure that  $f_{\text{norm}}$  captures healthy handwriting dynamics, its prediction objective is optimized exclusively using HC samples during training.

For valid consecutive segments, the normative deviation is computed as

$$d_{j+1} = \frac{1}{F} \|\hat{\mathbf{x}}_{j+1} - \bar{\mathbf{x}}_{j+1}\|_2^2, \quad (28)$$

where  $F$  denotes the dimensionality of the trajectory feature vector. To align the deviation sequence with the segment-token sequence, we set  $d_1 = 0$ .

The segment-wise deviation is normalized within each sample as

$$\tilde{d}_j = \frac{d_j}{\max(\max d_k, \epsilon)}, \quad (29)$$

where  $\epsilon > 0$  ensures numerical stability.

The sample-level normative score is computed from the unnormalized deviation:

$$d_{\text{score}} = \frac{\sum_{j=1}^{T_f-1} m_j^{\text{step}} d_{j+1}}{\sum_{j=1}^{T_f-1} m_j^{\text{step}}}, \quad (30)$$

where  $m_j^{\text{step}} = 1$  means that both tokens  $j$  and  $j+1$  are valid.

The normative prediction objective is optimized only using healthy control samples. Therefore, the normative predictor learns segment-level trajectory regularities characteristic of healthy handwriting without directly fitting AD samples. During inference, deviations from these healthy regularities provide auxiliary abnormality cues for segment-risk calibration and subject-level classification.

### G. Task-Aware Multi-Expert Segment-Risk Modeling

Different handwriting tasks may emphasize different cognitive and motor functions. Therefore, we use a task-aware multi-expert module to estimate token-level AD risk.

As shown in Fig.3 (b), each expert produces a segment-risk logit:

$$r_j^e = f_e(\mathbf{s}_j), \quad e = 1, \dots, E, \quad (31)$$

where  $E$  is the number of experts. A global state vector is obtained by masked mean pooling:

$$\mathbf{s}_g = \frac{\sum_{j=1}^{T_f} \bar{m}_j \mathbf{s}_j}{\sum_{j=1}^{T_f} \bar{m}_j}, \quad (32)$$

where  $\bar{m}_j$  is the segment-level valid mask. The router input is the concatenation of the global state and the task embedding:

$$\mathbf{z}_r = [\mathbf{s}_g; \mathbf{e}_\tau], \quad (33)$$

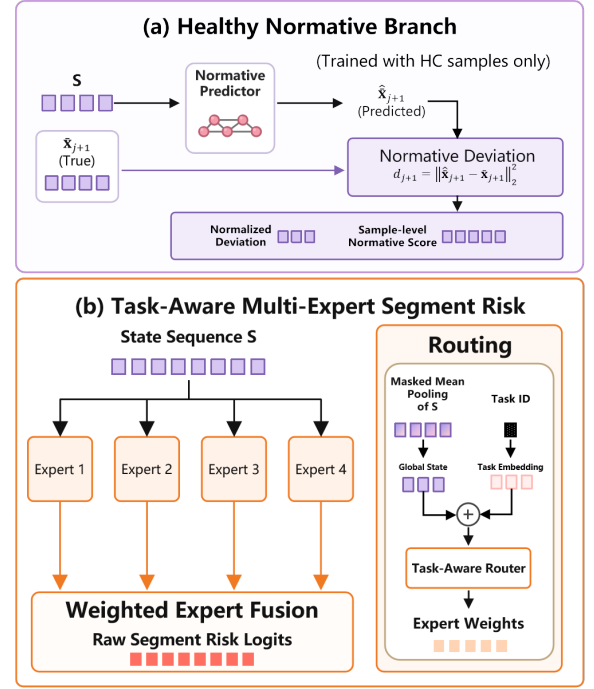


Fig. 3. Illustration of (a) the Healthy Normative Branch and (b) the Task-Aware Multi-Expert Segment Risk module.

where  $\tau$  denotes the task identity. The router produces expert logits:

$$\rho = \frac{R(\mathbf{z}_r)}{\tau_r}, \quad (34)$$

where  $\tau_r$  is the router temperature. The router logits are converted into normalized expert weights using a softmax function:

$$\pi_e = \frac{\exp(\rho_e)}{\sum_{l=1}^E \exp(\rho_l)}, \quad e = 1, \dots, E. \quad (35)$$

The raw segment-risk logit is

$$r_j = \sum_{e=1}^E \pi_e r_j^e. \quad (36)$$

The final segment-risk logit is calibrated by hidden-state change and healthy-normative deviation:

$$\tilde{r}_j = r_j \left( 1 + \lambda_c \chi_j + \lambda_n \tilde{d}_j \right), \quad (37)$$

where  $\lambda_c$  and  $\lambda_n$  control the contributions of state-change and normative-deviation calibration, respectively. The normative deviation is detached in this calibration path, so that AD labels do not directly optimize the healthy-only normative predictor through the segment-risk objective.

The segment-level AD probability is

$$p_j = \sigma(\tilde{r}_j). \quad (38)$$

The temporal risk weights are computed by masked softmax:

$$\alpha_j = \frac{\bar{m}_j \exp(\tilde{r}_j)}{\sum_{k=1}^{T_f} \bar{m}_k \exp(\tilde{r}_k)}. \quad (39)$$

The handwriting-level risk representation is obtained by risk-weighted pooling:

$$\mathbf{s}_{\text{risk}} = \sum_{j=1}^{T_f} \alpha_j \mathbf{s}_j. \quad (40)$$

#### H. Final Classification

To further incorporate the overall deviation from healthy handwriting patterns, the sample-level normative score is projected into the same feature space and added to  $\mathbf{s}_{\text{risk}}$ :

$$\mathbf{s}'_{\text{risk}} = \mathbf{s}_{\text{risk}} + \phi_n(\log(1 + d_{\text{score}})), \quad (41)$$

where  $\phi_n$  is a learnable projection. The normative score is detached before this operation to prevent the classification objective from directly optimizing the healthy normative predictor.

Finally, the AD/HC prediction is

$$\hat{y} = \text{Softmax}(\mathbf{W}_c \mathbf{s}'_{\text{risk}} + \mathbf{b}_c). \quad (42)$$

#### I. Training Objective

The model is trained with a compact objective:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{norm}} \mathcal{L}_{\text{norm}}. \quad (43)$$

where  $\lambda_{\text{seg}} = 0.1$  and  $\lambda_{\text{norm}} = 0.2$ . These weights keep the auxiliary segment-level and normative objectives effective without overwhelming the primary classification loss, thereby maintaining a balanced and stable optimization.

The sample-level classification loss is the standard cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{i=1}^B \log \hat{y}_{i,y_i}. \quad (44)$$

Since no manual segment-level annotations are available, we use weakly supervised segment-level multiple-instance learning. For sample  $i$ , let  $L_i = \sum_{j=1}^{T_f} \tilde{m}_{i,j}$  be the number of valid segment tokens. The number of selected top-risk tokens is

$$K_i = \max(1, \lceil \eta L_i \rceil), \quad (45)$$

where  $\eta$  is the top- $k$  ratio. The sample-level logit is

$$\bar{r}_i = \frac{1}{K_i} \sum_{j \in \text{Top}K_i} \tilde{r}_{i,j}. \quad (46)$$

The segment MIL loss is

$$\mathcal{L}_{\text{seg}} = -\frac{1}{B} \sum_{i=1}^B [y_i \log \sigma(\bar{r}_i) + (1 - y_i) \log (1 - \sigma(\bar{r}_i))]. \quad (47)$$

This loss encourages AD samples to contain at least some high-risk segments, while suppressing strong AD-like segment responses in HC samples.

The healthy normative loss is applied only to healthy control samples:

$$\mathcal{L}_{\text{norm}} = \frac{\sum_{i \in \mathcal{B}_{HC}} \sum_{j=1}^{T_f-1} m_{i,j}^{\text{step } \frac{1}{F}} \|\hat{\mathbf{x}}_{i,j+1} - \bar{\mathbf{x}}_{i,j+1}\|_2^2}{\sum_{i \in \mathcal{B}_{HC}} \sum_{j=1}^{T_f-1} m_{i,j}^{\text{step}}}, \quad (48)$$

where  $\mathcal{B}_{HC}$  is the set of healthy control samples in the batch. If no healthy-control sample is present in a mini-batch,  $\mathcal{L}_{\text{norm}}$  is set to zero. AD samples do not contribute to this loss, so the normative predictor learns healthy segment-level trajectory regularities rather than disease-specific patterns.

## IV. EXPERIMENTS

#### A. Dataset

We used the DARWIN-RAW dataset [16, 19], a public online handwriting benchmark for Alzheimer's disease (AD) detection. The dataset contains recordings from 174 participants, including 89 AD patients and 85 healthy controls (HC). Each participant completed 25 handwriting tasks designed for early AD detection, covering memory and dictation (M), copying (C), and graphic (G) tasks. The handwriting data were collected using a Wacom Bamboo tablet at a sampling rate of 200 Hz. In our experiments, each participant-task recording is treated as one trajectory sample with a subject-level AD/HC label. From the raw signals, we derive six dynamic kinematic features, including velocity, acceleration, jerk, curvature, pressure variation, and angular velocity, which are used as the input sequence of our model.

#### B. Experimental Setup

We evaluated the proposed method using five-fold cross-validation on the DARWIN-RAW dataset. In each fold, all samples from the same participant were assigned to the same split to avoid subject information leakage. Each participant-task recording was treated as one trajectory sample, and all 25 handwriting tasks were used for joint training and evaluation.

For a fair comparison, all baseline models were evaluated under the same subject-level splits, input features, and evaluation protocol as the proposed method. We compared our method with classical machine learning and deep sequence modeling baselines, Random Forest (RF)[34], CNN-1D(AD)[25], BiLSTM[35], Reservoir Computing(AD)[36], Transformer[37], Mamba[31], HSDA-MS(AD)[20]. During training, we used the AdamW optimizer with an initial learning rate of  $1 \times 10^{-4}$  and a weight decay of  $3 \times 10^{-4}$ . The batch size was set to 8, and the maximum number of training epochs was set to 100. A cosine annealing learning rate scheduler was adopted, with the minimum learning rate set to 1% of the initial learning rate. Early stopping was applied when the validation performance did not improve for 10 consecutive epochs. The validation set is randomly split from the training data, accounting for 10% of the full dataset. All experiments were implemented in PyTorch and conducted on NVIDIA<sup>TM</sup> GPUs.

We use balanced accuracy as the primary metric to account for potential class imbalance between AD and HC participants. We further report Precision-AD, F1-AD, sensitivity, and specificity for a comprehensive evaluation. We additionally report the area under the receiver operating characteristic curve (AUC) to evaluate the model's discriminative ability.

The main architectural configurations of the proposed NormPaST-Risk model are summarized in Table I.

TABLE I  
ARCHITECTURAL SETTINGS OF THE PROPOSED NORMPaST-RISK MODEL.

| Parameter                             | Value      |
|---------------------------------------|------------|
| Input Dimension                       | 6          |
| Hidden Dimension                      | 96         |
| Multi-scale Kernel Sizes              | [3, 7]     |
| Segment Token Stride                  | 128 points |
| Number of SSM Layers                  | 4          |
| SSM State Dimension                   | 16         |
| Segment-risk Experts                  | 8          |
| Expert Hidden Dimension               | 48         |
| Router Hidden Dimension               | 96         |
| State-change Weight ( $\lambda_c$ )   | 0.5        |
| Normative Risk Weight ( $\lambda_n$ ) | 0.3        |
| Dropout Rate                          | 0.3        |

### C. Comparison with Representative Baselines

To evaluate the effectiveness of the proposed NormPaST-Risk, we compare it with a diverse set of representative methods, including a classical machine learning model (Random Forest), recurrent and convolutional neural networks (BiLSTM and CNN-1D), reservoir computing, a Transformer-based model, the selective state-space model Mamba, and the handwriting-oriented HSDA-MS. All methods are evaluated under the same experimental protocol, and the results are reported as mean  $\pm$  standard deviation (%) across five folds. The best result for each metric is highlighted in bold, while the second-best result is underlined.

As shown in Table II, NormPaST-Risk achieves the best performance on five of the six evaluation metrics, including an AUC of 95.84%, a balanced accuracy of 92.55%, a Precision-AD of 92.43%, an F1-AD score of 92.74%, and a specificity of 91.77%. Compared with the second-best result for each metric, the proposed method improves balanced accuracy, Precision-AD, F1-AD, and specificity by 3.50, 6.30, 2.31, and 9.42 percentage points, respectively, while also achieving a slightly higher AUC.

It is worth noting that the highest sensitivity does not necessarily correspond to the best overall classification performance. Mamba achieves the highest sensitivity of 97.71%, while its specificity is comparatively lower at 79.93%. Similarly, Transformer obtains a sensitivity of 95.49% with a specificity of 82.35%, suggesting a trade-off between detecting AD participants and correctly identifying healthy controls. This pattern is also reflected in their Precision-AD values of 84.59% and 85.39%, respectively, which are not correspondingly high despite their high sensitivity. HSDA-MS shows a similar pattern, achieving 95.75% sensitivity and 82.35% specificity, indicating that the performance between the two classes remains somewhat unbalanced.

In contrast, NormPaST-Risk achieves 93.33% sensitivity together with 91.77% specificity, resulting in the highest balanced accuracy of 92.55%. This more balanced sensitivity-specificity trade-off suggests that the proposed method can effectively identify AD participants while maintaining reliable recognition of healthy controls. Such balanced discrimination is desirable for AD detection, as excessive false-positive predictions may limit the practical reliability of a classifica-

tion system. These results suggest that relying mainly on a global sequence representation may capture discriminative but less disease-specific variations, whereas identifying localized abnormal handwriting segments can reduce the influence of irrelevant trajectory portions and provide more robust AD-related handwriting evidence.

The performance differences among the sequence models further demonstrate the effectiveness of the proposed architecture. Although BiLSTM, CNN-1D, Reservoir, Transformer, and Mamba are capable of modeling temporal handwriting signals to different extents, their F1-AD scores remain between 77.78% and 90.43%, which are lower than the 92.74% achieved by NormPaST-Risk. Moreover, NormPaST-Risk achieves the highest Precision-AD of 92.43%, indicating that the proposed model not only identifies AD participants more effectively but also produces fewer false-positive AD predictions. These results suggest that generic temporal modeling alone is insufficient to fully characterize the AD-related abnormalities embedded in online handwriting trajectories.

A key advantage of NormPaST-Risk is that it incorporates handwriting-specific structural priors rather than treating the trajectory as a homogeneous sequence. First, the multi-scale temporal encoder captures handwriting dynamics at different temporal resolutions, enabling the model to characterize both short-term kinematic variations and longer motion patterns. This provides richer local representations than conventional CNN-1D or generic sequence models relying on a single temporal modeling mechanism.

Second, NormPaST-Risk explicitly distinguishes on-paper and in-air behaviors through Paper-Air state-aware modeling. These two states characterize different aspects of the handwriting process: on-paper trajectories primarily reflect writing execution, whereas in-air movements are associated with spatial transitions and movement preparation. By preserving these heterogeneous states rather than modeling them as an undifferentiated sequence, the proposed framework maintains state-specific dynamics and captures variations across writing execution and transitional behaviors. This provides a more handwriting-specific representation than generic sequence architectures such as BiLSTM, Transformer, and Mamba.

Third, the selective state-space encoder complements local multi-scale representations by capturing long-range dependencies across the handwriting sequence. Rather than relying exclusively on local convolution or recurrent accumulation, it enables interactions between temporally distant segments and characterizes changes in handwriting dynamics over longer temporal ranges. This is particularly useful when informative abnormalities occur intermittently rather than continuously throughout a trajectory.

In addition, the Healthy Normative Branch provides an explicit reference for normal handwriting dynamics. Rather than relying solely on direct AD-HC discrimination, it models regular trajectory patterns from HC samples and quantifies deviations from the learned healthy dynamics. This normative deviation provides complementary information to the discriminative representation and encourages the model to focus on handwriting patterns that deviate from normal progression rather than relying only on subject or task variation.

TABLE II

COMPARISON WITH REPRESENTATIVE BASELINE METHODS FOR ONLINE HANDWRITING-BASED AD DETECTION. RESULTS ARE REPORTED AS MEAN  $\pm$  STANDARD DEVIATION (%) OVER FIVE FOLDS. THE BEST RESULT IS SHOWN IN BOLD AND THE SECOND-BEST RESULT IS UNDERLINED.

| Model       | BAcc                               | Precision-AD                       | F1-AD                              | Sensitivity                        | Specificity                         | AUC                                |
|-------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-------------------------------------|------------------------------------|
| CNN-1D      | 74.80 $\pm$ 8.28                   | 71.21 $\pm$ 5.92                   | 77.78 $\pm$ 9.22                   | 86.67 $\pm$ 16.01                  | 62.94 $\pm$ 9.67                    | 88.39 $\pm$ 9.94                   |
| BiLSTM      | 76.90 $\pm$ 5.47                   | 77.80 $\pm$ 8.65                   | 78.05 $\pm$ 4.08                   | 79.67 $\pm$ 9.66                   | 74.12 $\pm$ 16.43                   | 84.81 $\pm$ 4.08                   |
| RF          | 77.88 $\pm$ 10.73                  | 78.24 $\pm$ 14.45                  | 79.71 $\pm$ 10.00                  | 83.99 $\pm$ 14.80                  | 71.76 $\pm$ 22.55                   | 85.50 $\pm$ 7.19                   |
| Reservoir   | 81.16 $\pm$ 5.00                   | 76.56 $\pm$ 3.54                   | 83.57 $\pm$ 4.99                   | 92.09 $\pm$ 7.53                   | 70.22 $\pm$ 4.24                    | 90.26 $\pm$ 4.07                   |
| Transformer | 88.92 $\pm$ 6.59                   | 85.39 $\pm$ 8.25                   | 89.96 $\pm$ 6.26                   | 95.49 $\pm$ 7.26                   | <u>82.35 <math>\pm</math> 10.19</u> | 92.73 $\pm$ 5.72                   |
| Mamba       | 88.82 $\pm$ 6.61                   | 84.59 $\pm$ 9.31                   | <u>90.43 <math>\pm</math> 5.31</u> | <b>97.71 <math>\pm</math> 3.13</b> | <u>79.93 <math>\pm</math> 13.45</u> | 93.62 $\pm$ 4.21                   |
| HSDA-MS     | <u>89.05 <math>\pm</math> 6.08</u> | <u>86.13 <math>\pm</math> 9.72</u> | <u>90.36 <math>\pm</math> 4.56</u> | <u>95.75 <math>\pm</math> 2.84</u> | <u>82.35 <math>\pm</math> 14.41</u> | <u>93.88 <math>\pm</math> 4.15</u> |
| <b>Ours</b> | <b>92.55 <math>\pm</math> 4.34</b> | <b>92.43 <math>\pm</math> 5.72</b> | <b>92.74 <math>\pm</math> 4.23</b> | 93.33 $\pm$ 6.09                   | <b>91.77 <math>\pm</math> 6.71</b>  | <b>95.84 <math>\pm</math> 2.90</b> |

More importantly, NormPaST-Risk does not rely solely on a global trajectory representation for classification. The multi-expert segment-risk module estimates AD-related predictive evidence at the segment level and emphasizes locally informative regions. By incorporating normative deviation into segment-risk calibration, the model prioritizes discriminative segments while accounting for their deviation from learned healthy dynamics, thereby reducing the influence of less informative trajectory portions.

Overall, NormPaST-Risk combines multi-scale dynamics, Paper-Air state modeling, healthy normative deviation, and segment-level risk discovery to capture handwriting-specific AD evidence more effectively than generic temporal models, resulting in more discriminative and balanced participant-level predictions.

#### D. Ablation Study

To investigate the contribution of each component in NormPaST-Risk, we conduct a series of ablation experiments under the same five-fold evaluation protocol. For each variant, only the corresponding component is removed or replaced, while the remaining network architecture and training settings are kept unchanged. The ablation results in Table III demonstrate that all major components contribute to the overall performance of NormPaST-Risk. Removing the segment-risk module causes the largest degradation in BAcc and F1-AD, decreasing them from 92.55% and 92.74% to 88.56% and 88.46%, respectively. Sensitivity also drops from 93.33% to 87.71%. These results confirm that explicitly identifying and aggregating disease-relevant local segments is more effective than treating all trajectory regions equally.

Removing Paper-Air modulation reduces BAcc to 89.67% and sensitivity to 88.76%, indicating that explicitly distinguishing on-paper motor execution from in-air planning behavior provides useful state-dependent information for AD recognition. Similarly, removing the Selective SSM decreases BAcc to 89.74% and produces the largest reduction in sensitivity, from 93.33% to 85.36%, despite achieving a higher specificity of 94.12%. This suggests that long-range state-space modeling is particularly important for identifying AD participants, while its removal biases the model toward more conservative HC predictions.

TABLE III  
ABLATION STUDY OF THE PROPOSED NORMPAST-RISK MODEL.

| Variant                      | BAcc         | F1-AD        | Sens.        | Spec.        |
|------------------------------|--------------|--------------|--------------|--------------|
| w/o Segment-Risk             | 88.56        | 88.46        | 87.71        | 89.41        |
| w/o Paper-Air Modulation     | 89.67        | 89.59        | 88.76        | 90.59        |
| w/o Selective SSM            | 89.74        | 89.33        | 85.36        | <b>94.12</b> |
| w/o Multi-scale Encoder      | 90.23        | 90.46        | 91.04        | 89.41        |
| w/o Healthy Normative Branch | 90.95        | 90.10        | 89.93        | 91.97        |
| <b>Full NormPaST-Risk</b>    | <b>92.55</b> | <b>92.74</b> | <b>93.33</b> | 91.77        |

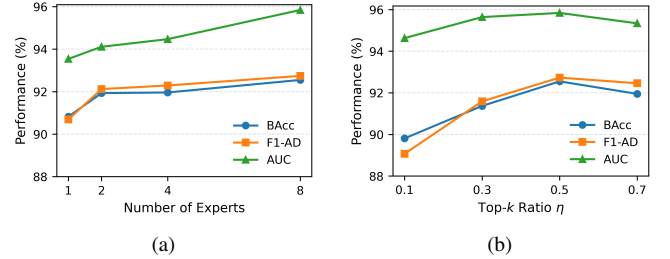


Fig. 4. Performance sensitivity: (a) the number of experts and (b) the top- $k$  ratio  $\eta$ .

The multi-scale encoder also provides consistent improvements across all evaluation metrics. Without multi-scale temporal modeling, BAcc and F1-AD decrease to 90.23% and 90.46%, respectively, demonstrating the benefit of capturing handwriting dynamics at different temporal scales. Finally, removing the Healthy Normative Branch reduces BAcc from 92.55% to 90.95% and sensitivity from 93.33% to 89.93%. Although this degradation is smaller than that caused by removing the other major components, it confirms that deviations from healthy handwriting dynamics provide complementary information for calibrating segment-level risk.

To further examine the robustness of NormPaST-Risk to key design choices in the segment-risk module, we conduct parameter sensitivity analyses on the number of risk experts and the segment top- $k$  ratio. The former determines the capacity of the multi-expert module to model heterogeneous abnormal handwriting patterns, while the latter controls the proportion of high-risk segments selected for weakly supervised segment-risk learning.

TABLE IV  
PARTICIPANT-LEVEL SEGMENT-RISK STATISTICS FOR HC AND AD GROUPS.

| Aggregation | HC                | AD                | $\Delta$ Risk | Cohen's $d$ |
|-------------|-------------------|-------------------|---------------|-------------|
| Mean        | 0.377 $\pm$ 0.014 | 0.435 $\pm$ 0.036 | 0.058         | 2.07        |
| Max         | 0.466 $\pm$ 0.024 | 0.561 $\pm$ 0.055 | 0.096         | 2.24        |
| Top 50%     | 0.411 $\pm$ 0.017 | 0.485 $\pm$ 0.046 | 0.074         | 2.12        |

TABLE V  
PERFORMANCE UNDER SEGMENT-LEVEL DELETION.  $\Delta$  DENOTES THE PERFORMANCE DECREASE RELATIVE TO THE ORIGINAL PREDICTION.

| Strategy | Ratio | BAcc (%) | $\Delta$ BAcc | AUC (%) | $\Delta$ AUC |
|----------|-------|----------|---------------|---------|--------------|
| Original | –     | 92.55    | –             | 95.84   | –            |
| Low-risk | 0.1   | 91.34    | 1.21          | 95.77   | 0.07         |
| Random   | 0.1   | 91.01    | 1.54          | 95.75   | 0.09         |
| Top-risk | 0.1   | 85.20    | <b>7.35</b>   | 93.88   | <b>1.96</b>  |
| Low-risk | 0.2   | 90.75    | 1.80          | 95.78   | 0.06         |
| Random   | 0.2   | 90.67    | 1.88          | 95.65   | 0.19         |
| Top-risk | 0.2   | 83.01    | <b>9.54</b>   | 92.36   | <b>3.48</b>  |
| Low-risk | 0.3   | 90.36    | 2.19          | 95.51   | 0.33         |
| Random   | 0.3   | 90.37    | 2.18          | 95.35   | 0.49         |
| Top-risk | 0.3   | 83.01    | <b>9.54</b>   | 91.98   | <b>3.86</b>  |
| Low-risk | 0.4   | 90.03    | 2.52          | 95.45   | 0.39         |
| Random   | 0.4   | 89.66    | 2.89          | 95.34   | 0.50         |
| Top-risk | 0.4   | 82.45    | <b>10.10</b>  | 91.30   | <b>4.54</b>  |
| Low-risk | 0.5   | 89.44    | 3.11          | 95.39   | 0.45         |
| Random   | 0.5   | 89.37    | 3.18          | 95.03   | 0.81         |
| Top-risk | 0.5   | 79.71    | <b>12.84</b>  | 89.62   | <b>6.22</b>  |

Increasing the number of experts generally improves the performance of the segment-risk module. Compared with a single expert, using multiple experts provides clear gains in balanced accuracy, F1-AD, and AUC, suggesting that different experts can capture complementary AD-related handwriting patterns. As shown in Fig.4 (a), the 8-expert configuration achieves the best overall performance. In contrast, the improvement from four to eight experts is relatively moderate, indicating that increasing model capacity beyond a certain point yields diminishing returns. These results support the use of multiple risk experts for modeling heterogeneous AD-related handwriting abnormalities.

The segment top- $k$  ratio controls the proportion of high-risk segments selected for weakly supervised segment-level learning. As shown in Fig.4 (b), as the ratio increases from 0.1 to 0.5, the performance consistently improves, indicating that AD-related abnormalities are distributed across multiple informative handwriting segments rather than being restricted to only a few isolated regions. The performance slightly decreases at 0.7, suggesting that including more lower-ranked segments may dilute discriminative AD-related evidence. A moderate top- $k$  ratio therefore helps retain sufficient abnormal segments while suppressing less informative regions that may contain variations associated with individual writing style, task-specific differences, or acquisition noise. This selective aggregation enables the model to focus more consistently on disease-relevant local patterns rather than irrelevant trajectory variability.

### E. Interpretability and Faithfulness Analysis

To examine whether the learned segment-risk scores differentiate HC and AD participants, we compare participant-level calibrated risks using mean, maximum, and top-50% aggregation, as shown in Table IV. AD participants consistently exhibit higher risk scores than HC participants across all three aggregation strategies. The absolute differences between the two groups are 0.058, 0.096, and 0.074 for mean, maximum, and top-50% aggregation, respectively.

Although these absolute differences are numerically modest on the  $[0, 1]$  risk scale, the corresponding Cohen's  $d$  values are all greater than 2.0, indicating large standardized differences between the HC and AD groups. Cohen's  $d$  was used to quantify the effect size, with values of 0.2, 0.5, and 0.8 conventionally interpreted as small, medium, and large effects, respectively [38]. In particular, maximum and top-50% aggregation yield larger group differences than mean aggregation, suggesting that emphasizing high-risk segments enhances the distinction between HC and AD participants. These findings support the segment-level risk modeling strategy and suggest that AD-related predictive evidence is more strongly concentrated in selected high-risk segments rather than being uniformly distributed throughout the trajectory. To further assess the contribution of these high-risk segments to the final prediction, we subsequently perform a segment-level deletion analysis.

To evaluate the faithfulness of the learned segment-risk scores, we perform a post-hoc deletion experiment without model retraining. At inference time, valid segments are ranked by their risk scores, and the highest-risk, lowest-risk, or randomly selected segments are masked from segment-risk pooling at ratios  $\rho_{\text{del}} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ . Random deletion is repeated and averaged, while all model parameters and evaluation settings are kept fixed. Results are reported as five-fold participant-level averages.

As shown in Table V, removing high-risk segments consistently causes substantially larger performance degradation than removing random or low-risk segments. When only 10% of the highest-risk segments are removed, BAcc and AUC decrease by 7.35 and 1.96 percentage points, respectively. In comparison, removing the same proportion of low-risk segments results in only a 1.21 point decrease in BAcc and almost no change in AUC. As the deletion ratio increases, the overall degradation becomes more pronounced. At a deletion ratio of 0.5, removing high-risk segments leads to decreases of 12.84 percentage points in BAcc and 6.22 percentage points in AUC, whereas low-risk deletion results in only a 3.11 point decrease in BAcc and negligible AUC degradation.

These results indicate that the learned segment-risk scores effectively rank handwriting segments according to their contribution to AD discrimination. The substantially larger degradation caused by removing high-risk segments suggests that the localized high-risk regions are closely associated with the discriminative handwriting evidence used by the model, thereby supporting the faithfulness of the proposed segment-level risk localization.

To qualitatively assess model interpretability, we visualize

one representative HC participant and one representative AD participant on a subset of handwriting tasks, while both participants completed all 25 tasks. Similar or visually redundant tasks are omitted for clarity. Each task is presented with three views: the original trajectory (solid lines: on-paper; dashed lines: in-air), the highest-risk segment selected by the model, and the full trajectory risk heatmap.

As shown in Fig. 5, clear differences can be observed between the two participants across multiple tasks. For

the HC participant, most trajectories exhibit relatively low risk, with only sparse local high-risk regions. In contrast, the AD participant shows more frequent and pronounced high-risk regions across different task types. Importantly, these regions are localized rather than uniformly distributed over the entire trajectory, indicating that the model relies on specific handwriting portions rather than assigning globally elevated risk to the whole sample.

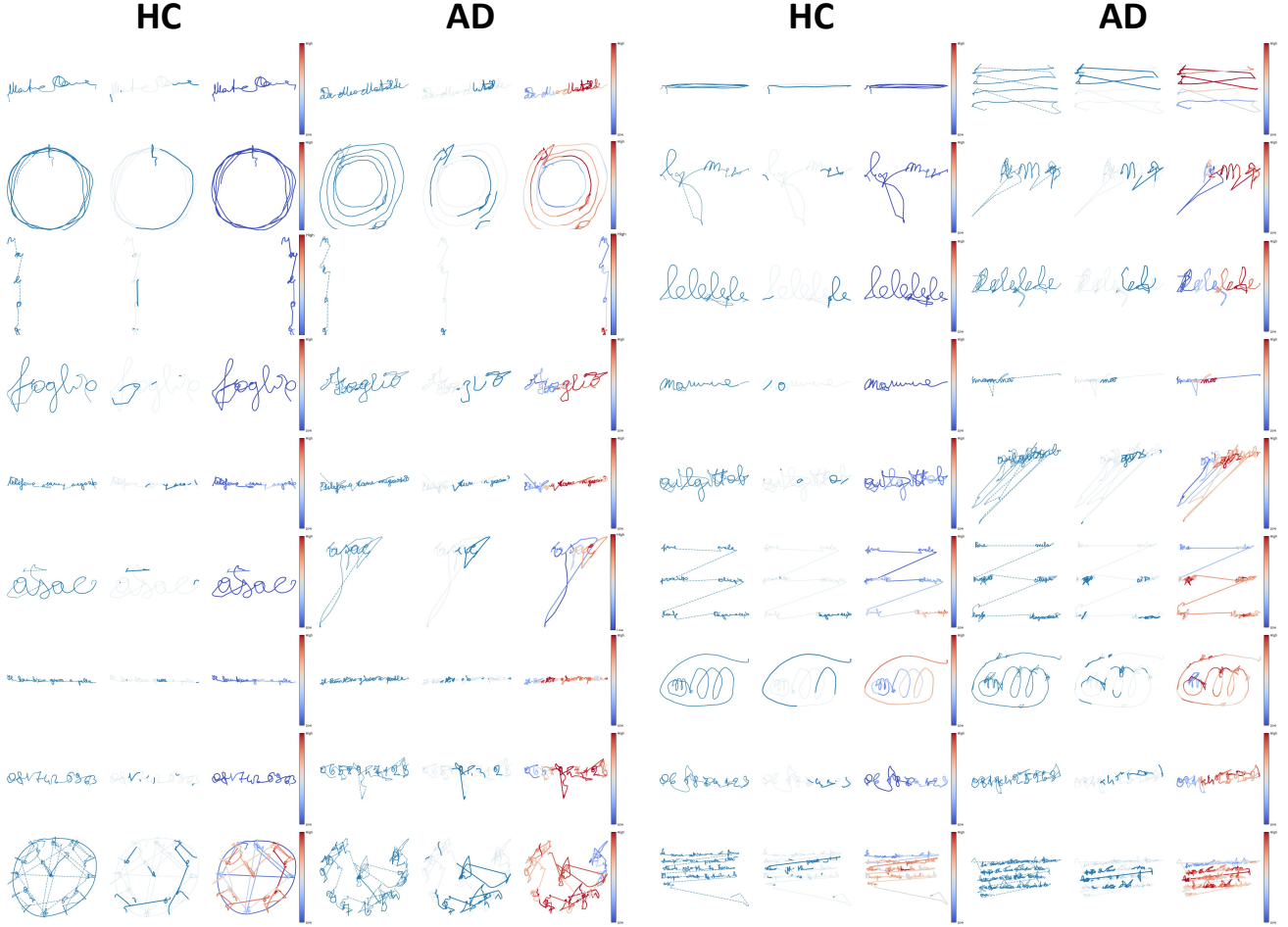


Fig. 5. Qualitative visualization of segment-level risk for one representative HC participant and one representative AD participant across selected handwriting tasks. Similar tasks are omitted for clarity. For each task, three subplots are shown: the original trajectory (solid lines indicate on-paper writing and dashed lines indicate in-air movement), the highest-risk segment selected by the model, and the trajectory-level risk heatmap, where warmer colors indicate higher predicted AD-related risk.

Despite substantial differences in trajectory shape and task content, localized high-risk regions repeatedly appear in the AD participant, whereas the HC participant generally maintains lower risk across the selected tasks. This suggests that the learned segment-risk representation can identify localized disease-related evidence across heterogeneous handwriting tasks, rather than being restricted to a single trajectory pattern.

High-risk regions are often located around portions with relatively complex local dynamics, such as curved strokes, direction changes, stroke transitions, and Paper-Air switch-

ing. However, these patterns should be interpreted as model-derived associations rather than direct evidence of specific cognitive or motor impairments. The localized risk distribution further suggests that the model identifies disease-related evidence from selected trajectory segments rather than from uniformly abnormal behavior throughout the sample.

The selected highest-risk segments generally coincide with locally elevated regions in the full risk maps. More importantly, this qualitative observation is consistent with the segment-deletion experiment, where removing high-risk segments results in a substantially larger performance decrease

than removing low-risk or randomly selected segments. This provides additional evidence that the identified regions are relevant to the model’s AD predictions.

Overall, the visualization suggests that NormPaST-Risk can localize high-risk handwriting segments that contribute to participant-level AD prediction. The contrast between HC and AD risk patterns supports the interpretability of the proposed segment-level risk discovery mechanism. These visualizations should nevertheless be regarded as qualitative model evidence rather than direct clinical biomarkers.

## V. CONCLUSION

In this work, we proposed NormPaST-Risk, a trajectory-based framework for online handwriting-based AD detection. The model integrates multi-scale temporal encoding, Paper-Air state-aware modeling, selective state-space modeling, healthy normative learning, and task-aware multi-expert segment-risk estimation to capture both global handwriting dynamics and localized AD-related abnormalities. In addition to participant-level prediction, the learned segment risks enable high-risk segments to be projected back onto the original handwriting trajectory, providing interpretable handwriting evidence for AD detection.

Experiments demonstrate that NormPaST-Risk achieves strong and balanced classification performance, reaching 92.55% BAcc and 95.84% AUC at the participant level. Ablation studies verify the contributions of the main components, while segment-deletion experiments show that removing high-risk segments causes substantially larger performance degradation than removing low-risk or randomly selected segments, supporting the reliability of the discovered local evidence. Overall, these results suggest that explicitly modeling Paper-Air dynamics and localized abnormal handwriting patterns provides an effective alternative to relying solely on global trajectory representations.

Several limitations remain. First, the current segment construction uses a fixed length of 128 trajectory points. Although this provides a consistent temporal unit for segment-level modeling, the resulting boundaries do not necessarily align with semantic handwriting units such as complete strokes, letters, or words. Consequently, a meaningful writing unit may be split across multiple segments, potentially weakening semantic continuity and limiting the interpretability of the discovered high-risk regions. Future work could explore adaptive or stroke-aware segmentation to better preserve meaningful handwriting structures. Second, the current evaluation is mainly based on a single online handwriting dataset, and the robustness of the proposed framework across different acquisition devices, handwriting tasks, and populations still requires further validation. Third, the segment-risk module is learned under weak supervision without manual segment-level annotations, meaning that the identified high-risk regions should be interpreted as model-derived evidence rather than direct clinical biomarkers. Finally, the healthy normative branch relies on HC samples to learn normal handwriting dynamics, which may be sensitive to the diversity and representativeness of the healthy reference population. Future work will therefore

focus on adaptive segmentation, cross-dataset validation, more robust normative modeling, and clinical verification of the discovered high-risk handwriting patterns.

## REFERENCES

- [1] D. S. Knopman, H. Amieva, R. C. Petersen, G. Ch  telat, D. M. Holtzman, B. T. Hyman, R. A. Nixon, and D. T. Jones, “Alzheimer disease,” *Nature reviews Disease primers*, vol. 7, no. 1, p. 33, 2021.
- [2] J. M. Long and D. M. Holtzman, “Alzheimer disease: an update on pathobiology and treatment strategies,” *Cell*, vol. 179, no. 2, pp. 312–339, 2019.
- [3] C. R. Jack Jr, J. S. Andrews, T. G. Beach, T. Buracchio, B. Dunn, A. Graf, O. Hansson, C. Ho, W. Jagust, E. McDade *et al.*, “Revised criteria for diagnosis and staging of alzheimer’s disease: Alzheimer’s association workgroup,” *Alzheimer’s & Dementia*, vol. 20, no. 8, pp. 5143–5169, 2024.
- [4] H. Hampel, S. E. O’Bryant, J. L. Molinuevo, H. Zetterberg, C. L. Masters, S. Lista, S. J. Kiddle, R. Batrla, and K. Blennow, “Blood-based biomarkers for alzheimer disease: mapping the road to the clinic,” *Nature Reviews Neurology*, vol. 14, no. 11, pp. 639–652, 2018.
- [5] T. K. Karikari, N. J. Ashton, G. Brinkmalm, W. S. Brum, A. L. Benedet, L. Montoliu-Gaya, J. Lantero-Rodriguez, T. A. Pascoal, M. Suarez-Calvet, P. Rosa-Neto *et al.*, “Blood phospho-tau in alzheimer disease: analysis, interpretation, and clinical utility,” *Nature Reviews Neurology*, vol. 18, no. 7, pp. 400–418, 2022.
- [6] K. K. Tsoi, J. Y. Chan, H. W. Hirai, S. Y. Wong, and T. C. Kwok, “Cognitive tests to detect dementia: a systematic review and meta-analysis,” *JAMA internal medicine*, vol. 175, no. 9, pp. 1450–1458, 2015.
- [7] L. C. Kourtis, O. B. Regele, J. M. Wright, and G. B. Jones, “Digital biomarkers for alzheimer’s disease: the mobile/wearable devices opportunity,” *NPJ digital medicine*, vol. 2, no. 1, p. 9, 2019.
- [8] W. Qi, X. Zhu, B. Wang, Y. Shi, C. Dong, S. Shen, J. Li, K. Zhang, Y. He, M. Zhao *et al.*, “Alzheimer’s disease digital biomarkers multidimensional landscape and ai model scoping review,” *npj Digital Medicine*, vol. 8, no. 1, p. 366, 2025.
- [9] G.   pukaityt  , C. Newton, and D. Chan, “Early detection of diseases causing dementia using digital navigation and gait measures: A systematic review of evidence,” *Alzheimer’s & Dementia*, vol. 20, no. 4, pp. 3054–3073, 2024.
- [10] J. H. Yan, S. Rountree, P. Massman, R. S. Doody, and H. Li, “Alzheimer’s disease and mild cognitive impairment deteriorate fine movement control,” *Journal of Psychiatric Research*, vol. 42, no. 14, pp. 1203–1212, 2008.
- [11] C. P. Fernandes, G. Montalvo, M. Caligiuri, M. Pertsinakis, and J. Guimaraes, “Handwriting changes in alzheimer’s disease: a systematic review,” *Journal of Alzheimer’s Disease*, vol. 96, no. 1, pp. 1–11, 2023.
- [12] M. A. El-Yacoubi, S. Garcia-Salicetti, C. Kahindo, A.-S. Rigaud, and V. Cristancho-Lacroix, “From aging to early-

- stage alzheimer's: uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning," *Pattern Recognition*, vol. 86, pp. 112–133, 2019.
- [13] J. Kawa, A. Bednorz, P. Stepień, J. Derejczyk, and M. Bugdol, "Spatial and dynamical handwriting analysis in mild cognitive impairment," *Computers in Biology and Medicine*, vol. 82, pp. 21–28, 2017.
- [14] N. D. Cilia, C. De Stefano, F. Fontanella, and S. M. Siniscalchi, "How word semantics and phonology affect handwriting of alzheimer's patients: a machine learning based analysis," *Computers in Biology and Medicine*, vol. 169, p. 107891, 2024.
- [15] P. Ghaderyan, A. Abbasi, and S. Saber, "A new algorithm for kinematic analysis of handwriting data; towards a reliable handwriting-based tool for early detection of alzheimer's disease," *Expert Systems with Applications*, vol. 114, pp. 428–440, 2018.
- [16] N. D. Cilia, G. De Gregorio, C. De Stefano, F. Fontanella, A. Marcelli, and A. Parziale, "Diagnosing alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking," *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104822, 2022.
- [17] J. Garre-Olmo, M. Faúndez-Zanuy, K. López-de Ipiña, L. Calvó-Perxas, and O. Turró-Garriga, "Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls," *Current Alzheimer Research*, vol. 14, no. 9, pp. 960–968, 2017.
- [18] P. Erdogmus and A. T. Kabakus, "The promise of convolutional neural networks for the early diagnosis of the alzheimer's disease," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106254, 2023.
- [19] N. D. Cilia, T. D'Alessandro, C. De Stefano, F. Fontanella, and M. Molinara, "From online handwriting to synthetic images for alzheimer's disease detection using a deep transfer learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4243–4254, 2021.
- [20] C. Gong, H. Qin, and M. A. El-Yacoubi, "Hybrid transformer for early alzheimer's detection: Integration of handwriting-based 2d images and 1d signal features," *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [21] E. Nardone, C. De Stefano, N. D. Cilia, and F. Fontanella, "Handwriting strokes as biomarkers for alzheimer's disease prediction: a novel machine learning approach," *Computers in Biology and Medicine*, vol. 190, p. 110039, 2025.
- [22] J. Park and S. Kang, "Paano: patch-based representation learning for time-series anomaly detection," *arXiv preprint arXiv:2602.01359*, 2026.
- [23] Z. Liu, Y. Wang, B. Li, J. Zheng, E. Eldele, M. Wu, and Q. Ma, "A unified shape-aware foundation model for time series classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, pp. 23972–23980.
- [24] D. Impedovo and G. Pirlo, "Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective," *IEEE reviews in biomedical engineering*, vol. 12, pp. 209–220, 2018.
- [25] Q. Dao, M. A. El-Yacoubi, and A.-S. Rigaud, "Detection of alzheimer disease on online handwriting using 1d convolutional neural network," *IEEE Access*, vol. 11, pp. 2148–2155, 2022.
- [26] L. Kang, X. Zhang, J. Guan, K. Huang, and R. Wu, "Early alzheimer's disease diagnosis via handwriting with self-attention mechanisms," *Journal of Alzheimer's Disease*, vol. 102, no. 1, pp. 173–180, 2024.
- [27] J. Sweidan, M. A. El-Yacoubi, and A.-S. Rigaud, "Explainability of cnn-based alzheimer's disease detection from online handwriting," *Scientific Reports*, vol. 14, no. 1, p. 22108, 2024.
- [28] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.
- [29] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *Advances in neural information processing systems*, vol. 35, pp. 22982–22994, 2022.
- [30] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in neural information processing systems*, vol. 35, pp. 35971–35983, 2022.
- [31] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [32] S. Müller, O. Preische, P. Heymann, U. Elbing, and C. Laske, "Increased diagnostic accuracy of digital vs. conventional clock drawing test for discrimination of patients in the early course of alzheimer's disease from cognitively healthy individuals," *Frontiers in aging neuroscience*, vol. 9, p. 101, 2017.
- [33] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [34] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [35] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [36] N. Mwamsojo, F. Lehmann, M. A. El-Yacoubi, K. Merghem, Y. Frignac, B.-E. Benkelfat, and A.-S. Rigaud, "Reservoir computing for early stage alzheimer's disease detection," *IEEE Access*, vol. 10, pp. 59821–59831, 2022.
- [37] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
- [38] J. Cohen, *Statistical power analysis for the behavioral sciences*. routledge, 2013.