Spatio-Temporal Representation of Eye-Tracking Data for Cognitive Failure Detection in VR

Tanaz Ghahremani[1,*], Sahar Niknam[1], Berin Venedik[1] and Jean Botev[1]

[1]VR/AR Lab, University of Luxembourg, 6 Avenue de la Fonte, L-4364 Esch-sur-Alzette, Luxembourg

Abstract

Human error remains a pervasive risk across safety-critical domains, motivating research on predictive approaches that enable selective automation and adaptive training. This paper presents a longitudinal Virtual Reality (VR) study for cognitive failure prediction, based on 50 hours of eye-tracking data collected over 18 months from a participant solving mental arithmetic problems. We propose a semantic spatio-temporal representation that transforms quantitative eye-tracking features into compact visualizations, and evaluates its effectiveness using machine learning models, including convolutional, recurrent, and fusion neural architectures. These models leverage both spatial and temporal dynamics to detect performance outcomes, achieving up to 87% accuracy in identifying missed responses. However, distinguishing wrong from correct answers proved more difficult, likely due to similar ocular behavior during confident but incorrect responses and the impact of class imbalance. Overall, the results highlight the feasibility of personalized AI systems trained on rich within-subject data, and position VR combined with deep learning as a practical platform for real-time monitoring and adaptive training to reduce human error.

Keywords

Personalized neural networks, cognitive failure, eye-tracking, spatio-temporal representation, data conversion, virtual reality, deep learning

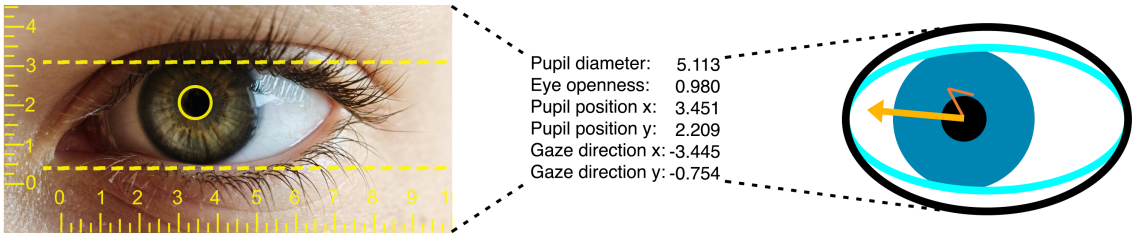


Figure 1: Conceptual illustration of the study pipeline: from ocular behavior to spatio-temporal representation. Visual elements and values are illustrative.

1. Introduction

Human error is a major contributor to adverse incidents across domains, from medical malpractice and manufacturing accidents to transportation fatalities and nuclear power plant catastrophes [1, 2, 3, 4]. Given the inevitability of mistakes in human performance, *human error* constitutes a systemic problem that must be addressed within the design of work systems. Reason classifies human error into two categories of *slips and lapses* and *mistakes* [5]. While mistakes happen during decision-making processes and is usually the result of cognitive shortcomings in addressing a new issue, slips and lapses of attention and memory happens during skill-based task, when the individual fails a task they already mastered. A

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus

*Corresponding author.
tanaz.ghahremani.001@student.uni.lu (T. Ghahremani); sahar.niknam@uni.lu (S. Niknam); berin.venedik.001@student.uni.lu (B. Venedik); jean.botev@uni.lu (J. Botev)
0009-0008-0038-272X (T. Ghahremani); 0000-0002-8021-1021 (S. Niknam); 0009-0006-2550-7176 (B. Venedik); 0000-0001-8492-7708 (J. Botev)

substantial portion of these errors stems from cognitive failures, such as slips of attention or lapses of memory [5], while performing routine or precise tasks.

Advances in automation and artificial intelligence (AI) are transforming all these domains by delegating repetitive and precision-demanding tasks to intelligent machines [6, 7, 8]. Such technologies have the potential to reduce reliance on manual performance and decrease the occurrence of human error. However, successful automation that ensures societal acceptance requires conditional and minimal delegation. Predicting human error supports such automation by intervening selectively and assigning tasks to the machine only when human failure is likely.

Growing research evidence shows that human error is not only explainable with post-error neural activity [9, 10, 11], but also predictable through computational models trained on physiological and behavioral data [12, 13, 14]. Achieving such predictive capacity, however, demands large-scale datasets that capture diverse physiological and behavioral patterns. Virtual reality (VR) offers a unique opportunity to address this challenge. Within virtual environments, not only can integrated biosensors provide extensive physiological and behavioral data, but such data can also be gathered in scenarios that would be difficult—if not impossible—to reproduce in real-world settings. Moreover, VR is already widely adopted as a training environment, from surgical simulation and medical education to flight training and industrial safety[15, 16, 17, 18], serving as an unprecedented platform for studying human performance and error through continuous and objective monitoring of cognitive markers, such as attentional focus and engagement level.

Eye-tracking is a well-established and reliable method for assessing cognitive load and relevant cognitive states [19, 20, 21]. It is particularly popular in VR-based research, where the headset technology enables effortless integration of eye-tracking and provides pixel-level access to gaze location within the virtual environment. With growing computational power and increasing complexity of machine learning models, many works have trained these models—especially neural networks—on eye-tracking data as an effective means of modeling cognitive processes and markers [22, 23, 24, 25, 26]. Eye-tracking data are typically represented as quantitative time series or aggregated measures (e.g., mean pupil size or eye movement speed) when used to train these models.

Building on this growing trend, our work contributes by re-expanding quantitative eye-tracking data into spatio-temporal visualizations of eye behaviors (Figure 1). By transforming numerical eye-tracking features into a minimal visualization of the eyes, we reintroduce the missing spatial context—for example, how pupil dilation relates to gaze direction—helping models to capture subtle signatures such as individual gaze habits. Feeding these visualizations as time series, we create spatio-temporal data to train Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and fusion architectures that combine both spatial and temporal representations. This approach allows us to capture complementary aspects of the data—spatial patterns, temporal dynamics, and their interactions, thereby maximizing the information available for predicting cognitive failure.

2. Related Work

CNNs have recently gained popularity for time-series and sequential data, supplementing traditional recurrent models. CNN architectures enable parallelization, efficient parameter sharing, and convolution and pooling operations, reducing the tensor dimensionality and highlighting relevant patterns. This lowers the complexity of the model while preserving crucial temporal-spatial data structures, resulting in an effective predictive model. Particularly, recent Human-Computer Interaction (HCI) studies have utilized image-based representations of time-series data to use CNNs for sequence modeling [27, 28, 29]. This approach combines temporal dynamics with spatially structured representations, capturing sequential dependencies faster and more efficiently than traditional recurrent neural networks (RNNs) [30, 31].

Bai et al. discovered that a simple Temporal Convolutional Network (TCN) consistently outperformed LSTM and Gated Recurrent Unit (GRU) baselines on varied sequence modeling benchmarks [30], indicating CNNs can successfully capture long-range dependencies. Similarly, Zhang et al. demonstrated

that purely character-level CNNs can achieve state-of-the-art text classification accuracy [32]. In human-activity recognition, Yang et al. applied a deep CNN to raw multichannel sensor streams and showed it automatically learns high-level features that outperform traditional hand-crafted methods [33]. These studies show that CNNs are a viable alternative to RNNs for sequence data, and many recent works adopt convolutional architectures for time-series tasks.

In addition to one-dimensional (1D) convolutions, a common strategy is to convert time series into images and apply 2D CNNs. Wang and Oates introduced Gramian angular fields (GAF) and Markov Transition Fields (MTF), using tiled CNNs on the resulting images and achieving highly competitive classification accuracy [31]. This idea has been used across many domains. For example, Qiu et al. transformed multimodal signals into GAF images and applied CNNs for emotion recognition, improving on prior methods [29]. Camara et al. converted electrocardiogram waveforms into GAF images and used a VGG19 CNN for user identification [28]. More recently, Qin et al. fused raw ECG and GAF-encoded images in a split-attention CNN (GAF-FusionNet) to achieve 94–99% accuracy on standard ECG datasets [34]. Chansri and Srinonchat similarly applied GAF encodings and CNNs to flex-sensor glove data for hand-gesture recognition [27]. Vortmann et al. applied an Imaging Time Series Approach (ITSA) to eye-tracking data, creating 2D Gramian/Markov representations of gaze features and training a CNN. Their model outperformed classical summary-feature classifiers in attentional-state tasks [35]. These examples demonstrate that imaging and convolutional techniques for time series are now common practice across applications.

This study proposes and evaluates a practical alternative to complex time-series encodings by transforming quantitative eye-tracking sequences into temporal eye frames and utilizing 2D CNNs for predicting cognitive failure. Our motivation is twofold; firstly, image encodings offer an instantly comprehensible visual summary of temporal and spatial eye dynamics, thereby enhancing exploratory data analysis and supporting human-in-the-loop validation. Second, we investigate whether a conceptually direct rendered temporal eye representation, rather than engineered transformations like GAFs or MTFs, can achieve similar predictive efficacy with prioritizing interpretability and the preservation of spatial intuition within the feature space.

3. Methodology

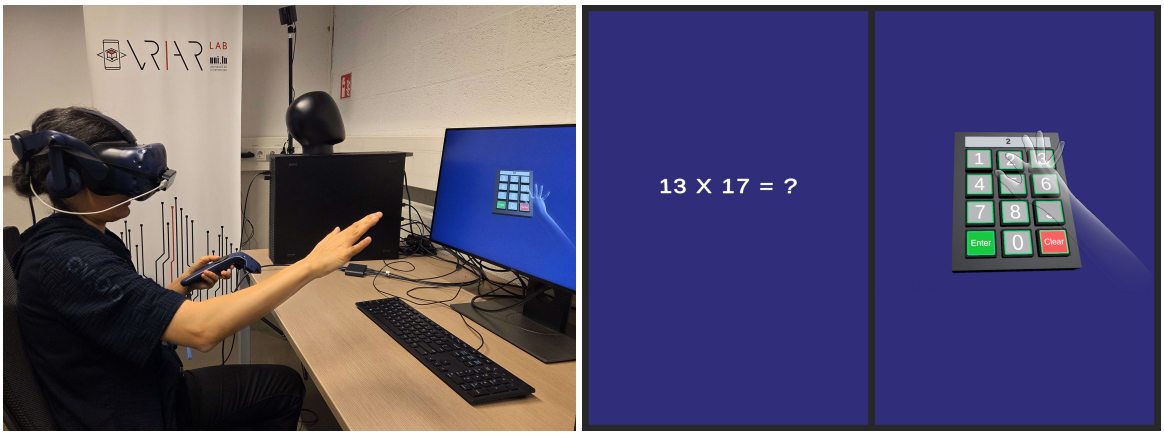
In this longitudinal study, we collected eye-tracking data from a participant, who completed 300 VR sessions solving multiplication questions over 18 months. The participant was employed as research assistants in the VR/AR Lab at the University of Luxembourg. The study received ethical approval from the Ethics Review Panel of the University of Luxembourg. The dataset associated with this work is publicly available on Zenodo (DOI: 10.5281/zenodo.18376588).

3.1. Study Design

The study was conducted in VR as a series of 10-minute sessions. During each session, we collected eye-tracking data at 120 Hz using the integrated Tobii tracker in the HTC VIVE Pro Eye headset[1] (Figure 2a). Sessions began with eye-tracking calibration, followed by a test 1×1 question to verify system functionality. The environment was designed to be minimal, with questions displayed in white on a solid blue background to control for confounds such as pupil light reflex and attention-related cognitive load (Figure 2b).

Multiplication problems were displayed for a maximum of 15 seconds. If the participant gave no indication of finding an answer during the interval, the problem was immediately replaced by the next one. When ready to respond, she pressed a controller button to activate a virtual keypad, which was operated through hand gestures tracked by a Leap Motion Controller[2] (Figure 2a).

[1] https://www.vive.com/sea/product/vive-pro-eye/overview/
[2] https://www.ultraleap.com/products/



(a) Study setup: the participant is wearing a VR headset while interacting with a virtual keypad with hand gestures, and holding a controller to trigger responses.

(b) Virtual environment: multiplication problems were presented in white on a blue background (left), and answers were entered using a virtual keypad operated via hand tracking (right).

The question set evolved across sessions. In the first session, the participant was presented with a random set of about 1,200 two-digit multiplication problems. Thereafter, a personalized subset of the questions was selected, consisting of problems the participant judged manageable yet challenging—demanding genuine cognitive effort without being so difficult as to cause disengagement or so simple as to require only superficial processing. Whenever the participant made fewer than five errors in five consecutive sessions, the personal subset was updated with more difficult questions. With this approach, we kept the cognitive load at a manageable level, mainly limiting the mistakes to two categories of *slips* and *lapses*—the categories that account for the majority of accidents caused by human error [5].

Upon completion of 300 sessions, we collected 17,343 eye-tracking samples, which included *eye openness*, *pupil diameter*, *pupil relative location* (2D vector), and *gaze direction* (3D vector). Samples were labeled as *Correct* (2) for accurate responses (15,217 samples), *Missed* (1) when no response was given within 15 seconds (1,100 samples), and *Wrong* (0) for inaccurate responses (1,026 samples) (Figure 3).

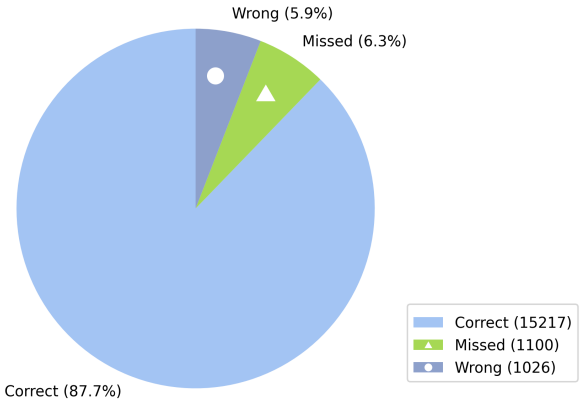


Figure 3: Distribution of eye-tracking samples across classes: *Correct*, *Missed*, and *Wrong*.

3.2. Preprocessing

Since response times varied across questions, the data samples had different shapes. Therefore, we first standardized the samples by truncating them to their last 30 frames. The 30-frame window corresponds to the last 250 ms, an interval reported by Urai et al. as critical for high-level cognitive processes, which inspired our choice [36].

Next, to take advantage of convolutional architectures for the prediction of cognitive failure, we transformed these sequential measurements into spatio-temporal eye frames. To generate these frames, each image captures a specific timestamp of eye behavior using standardized anatomy, as shown in Figure 4. The eyes are positioned at fixed horizontal offsets, with constant elliptical outlines. Eyelid movement is represented by openness signals, which have continuous representations for closed and open states. Blinks were explicitly visualized rather than treated as missing data, allowing the network to recognize them as relevant cognitive events. Pupil diameter, illustrated with green circles, is discretized into four levels, allowing for a stable representation without noise amplification. The pupil's position is transferred from sensor data to a 2D eye coordinate space, and the iris is rendered for size. Gaze direction is shown by arrows originating from the pupil. Furthermore, a fading movement trace retains temporal structure, which helps convolutional neural networks recognize temporal patterns. This visualization reflects actual anatomical scales, including proportional eye form, distance between eyes, and natural blinking mechanisms, to ensure that the visuals generated are interpretable.

Given the highly imbalanced nature of the dataset, we applied a two-step balancing technique. First, downsampling was performed on the majority class to reduce skew. Second, augmentation techniques were applied to the minority class by injecting controlled noise into the feature values prior to image transformation, thereby generating additional synthetic samples. Following these procedures, all images were normalized to standardize feature distributions and then resized to a fixed resolution (168 px ×96 px) suited for deep learning architectures.

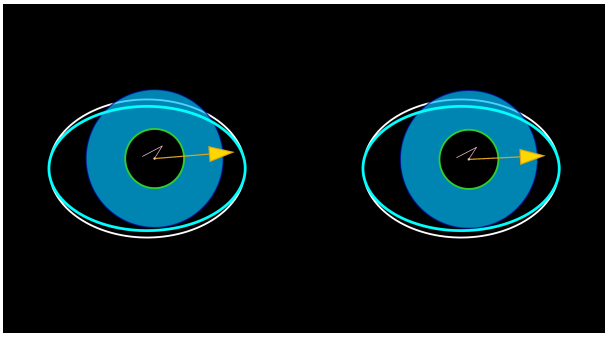


Figure 4: Example of an eye representation frame.

This preprocessing pipeline ensures that the temporal representations of the eye encode both instantaneous and temporal aspects of the behavior of the eyes while addressing class imbalance, thus preparing the data for robust CNN-based analysis.

3.3. Model Architectures and Training

This section outlines the learning strategies used in our study. To thoroughly assess the efficiency of various representations of eye-tracking data, we employed three complementary training procedures. First, a CNN-based model was trained on the temporal eye frames generated from the quantitative time series, allowing us to exploit the advantages of convolutional architectures in spatial feature extraction. Second, an LSTM-based model was trained directly on the raw time-series data, serving as a sequential baseline that captures temporal dependencies without image conversion. Finally, we designed a fusion model that combined convolutional and recurrent representations, integrating the strengths of both spatial and temporal learning. In the following subsections, we detail the architecture design.

All models were trained using a batch size of 32 for a maximum of 250 epochs, with an initial learning rate of 10^{-3}, weight decay of 10^{-4}, and an early stopping criterion with a patience of 50 epochs. These hyperparameters were selected based on empirical tuning to balance convergence speed and generalization. The dataset was split into training, validation, and test sets using an 80:10:10 ratio. Training and evaluation were conducted on an NVIDIA Tesla V100 GPU, with 16 GB RAM.

3.3.1. CNN-based Model

The first method employs a CNN trained on temporal eye images to utilize spatial information inside each frame and short-term temporal dependencies across the sequence. All the frames in the time steps were concatenated before passing to the network, resulting in a 30-channel input corresponding to the last 250 ms of data at 120 Hz as previously stated. Figure 5 shows the architecture, which consists of convolutional stages for spatial feature extraction, followed by a temporal attention module that emphasizes informative dynamics across frames. Dropout is used throughout the network to prevent overfitting. The final classifier consists of a stack of fully connected layers with normalization and non-linear activations, which generate the final class predictions. This CNN architecture uses local connectivity, parameter sharing, and hierarchical feature abstraction while focusing on short-term temporal cues. It served as a baseline image-based model for comparing sequential and fusion approaches.

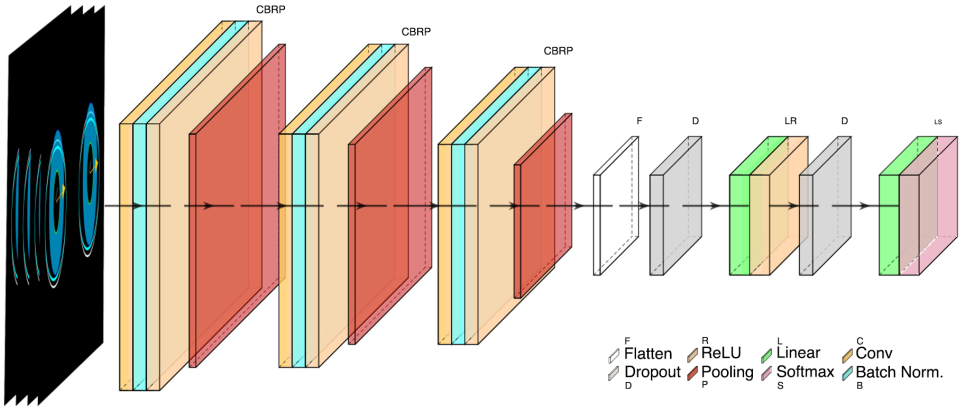


Figure 5: CNN model architecture.

3.3.2. LSTM-based Model

The second approach uses a recurrent neural network with LSTM units to model raw quantitative eye-tracking time series. This bidirectional LSTM architecture incorporates both forward and backward temporal contexts, improving its capacity to recognize patterns before cognitive failures. The model output was condensed into a classifier head with Softmax activation and dropout regularization, providing a complementary perspective to the CNN-only approach.

3.3.3. Fusion Model

The Fusion model integrates per-frame visual embeddings from temporal eye frames with parallel quantitative feature sequences to learn joint spatio-temporal representations. The architecture is composed of three modules as shown in Figure 6: a CNN-based image encoder, a recurrent feature encoder, and a fusion LSTM followed by a classifier. The image encoder extracts spatial representations from the eye-image sequence, while the feature encoder uses a bidirectional recurrent layer to capture temporal patterns in the numerical attributes. Their outputs were combined and processed by a fusion LSTM to model cross-modal temporal dependencies. The resulting representation is then passed to a fully connected classifier with nonlinear activation and dropout for the final prediction. The Fusion model was designed to capture complementary patterns, such as instantaneous visual cues and long-term sequential trends, in order to outperform single-modality baselines.

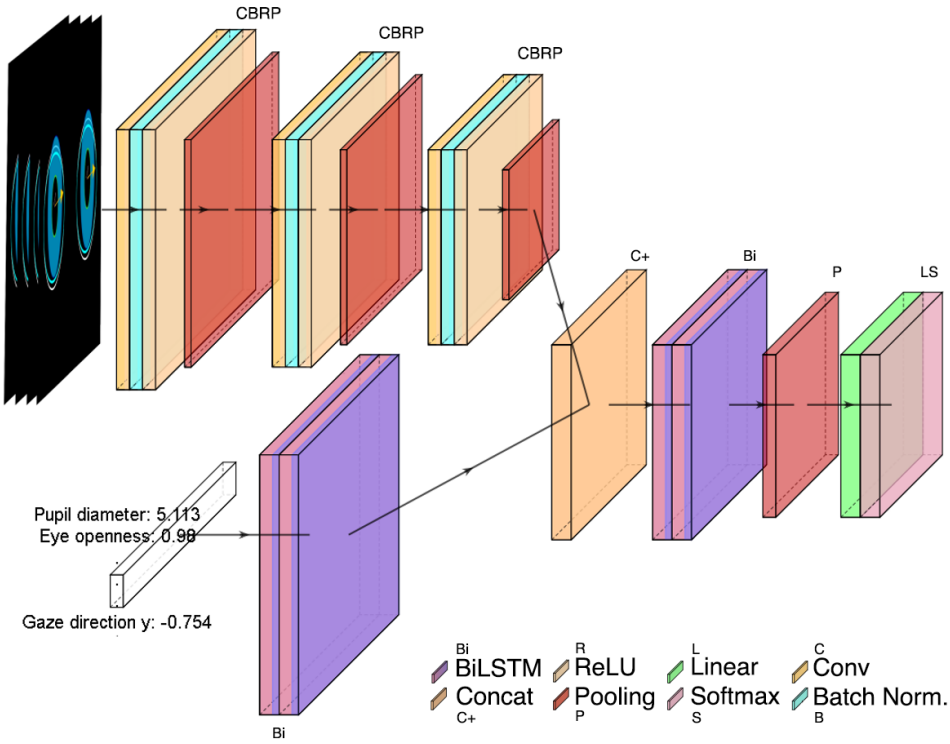


Figure 6: CNN and LSTM Fusion model architecture.

4. Results

We evaluated our models on the dataset with three classes: *Correct*, *Missed*, and *Wrong* answers. The *Missed* class showed a clear and consistent ocular pattern, allowing models across all architectures to identify these samples with relatively high accuracy. However, performance was lower when discriminating between *Correct* and *Wrong* responses, as their eye-tracking signatures appeared highly similar.

Table 1 presents the performance of all models on the original three-class classification task. As shown, none of the models achieved satisfactory classification performance on the *Wrong* class, although the *Missed* class was detected with reasonable recall. The LSTM-only model performed best overall, particularly on the *Missed* class. The CNN-only model maintained its competitiveness, while the Fusion model produced balanced results, but neither addressed the poor separability of the *Wrong* class.

In the three-class setup, models struggled to identify the *Wrong* class, although *Missed* responses were detected more reliably, and *Correct* responses performed relatively well. The difficulties appeared to originate not from training capacity, but from the behavioral similarities between *Correct* and *Wrong* trials, as the participant frequently provided incorrect answers rapidly and confidently. These characteristics prompted a change to a binary task contrasting *Correct*. Therefore, due to challenges in learning the *Wrong* class and its limited presence in the dataset, we excluded it from subsequent experiments. We reformulated the task as a binary classification between *Correct* and *Missed* classes.

All models were retrained in this setting, resulting in substantially improved performance as shown in Table 2. Overall, the LSTM-only model outperformed the CNN-only model in both three-class and binary classification settings, largely due to its ability to model pupil diameter and its limited time window of 250 ms (30 frames). The loss curves in Figure 7 and the confusion matrix in Figure 8 illustrate the stability of training and the class separation achieved by the model. However, the CNN-only model showed highly competitive results, suggesting that convolutional architectures are a promising alternative for longer time sequences or complex feature sets. The Fusion model, which combined CNN-derived spatial encodings with LSTM-based sequential representations, showed comparable

Table 1
Model performance on the original three-class test set. Precision (P), Recall (R), and F1-score are reported for each class, along with overall accuracy (Acc.).

	Correct			Missed			Wrong			Acc.
Model	P	R	F1	P	R	F1	P	R	F1	
CNN-only	0.572	0.553	0.563	0.647	0.647	0.643	0.348	0.364	0.356	0.522
LSTM-only	0.607	0.773	0.680	0.626	0.902	0.739	0.458	0.100	0.164	0.605
Fusion	0.624	0.707	0.662	0.638	0.814	0.716	0.387	0.218	0.279	0.588

performance to single-modality baselines, indicating its potential for more complex applications with richer spatial-temporal feature interactions, although it comes with a higher computational cost.

Table 2
Model performance on the two-class test set.

	Correct			Missed			Acc.
Model	P	R	F1	P	R	F1	
CNN-only	0.865	0.813	0.838	0.750	0.816	0.781	0.814
LSTM-only	0.932	0.827	0.876	0.783	0.913	0.843	0.862
Fusion	0.920	0.847	0.882	0.800	0.893	0.844	0.866

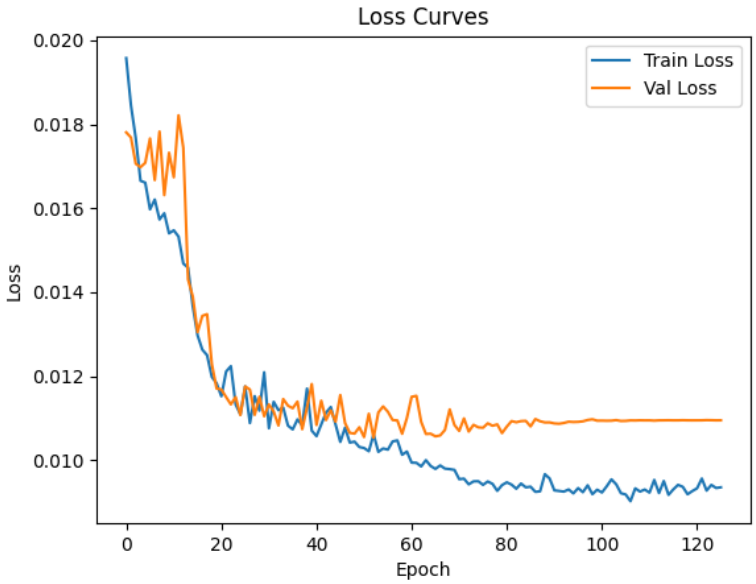


Figure 7: Loss curves of the LSTM-only model's predictions on the test dataset.

5. Discussion

This work investigates the potential of semantic visualizations of eye-tracking time series for predicting cognitive failure in VR. We trained a CNN-only, an LSTM-only, and a Fusion model on spatio-temporal data representations to evaluate how well different architectures capture patterns indicative of impending failure. While all models reliably distinguished *Missed* from *Correct* responses (Table 2), separating *Wrong* from *Correct* samples remained unresolved (Table 1).

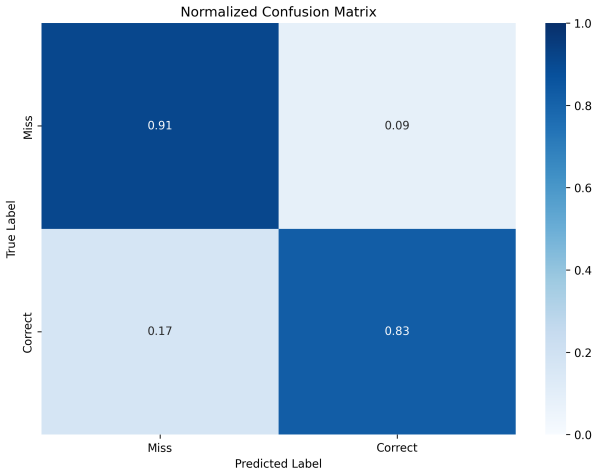


Figure 8: Normalized confusion matrix of the LSTM-only model's predictions on the test dataset.

As highlighted in Table 1, in the three-class task, all models achieved their highest scores on the *Missed* class, suggesting that this class exhibits pronounced and learnable ocular patterns. Performance on *Correct* responses was moderate, while *Wrong* responses remained poorly detected across architectures. The slightly better performance on *Correct* samples can be attributed to the class imbalance (Figure 3). Despite downsampling the majority class and augmenting the minorities, *Correct* samples still offered greater variation and information content, biasing the training process toward classifying ambiguous cases as *Correct*. Overall accuracy in the three-class task remained modest (0.52–0.60), reflecting limited robustness, with the LSTM-only model performing best overall, particularly on the *Missed* class, due to its ability to describe short sequences dominated by pupil diameter dynamics.

Given the quick convergence of the models (Figure 7), this limitation is unlikely to stem from insufficient training capacity, but rather from subtle, difficult-to-capture nuances in ocular behavior. Reviewing the wrong answers revealed frequent examples of attentional slips (*e.g.*, 468 instead of 668, or 281 instead of 261). Moreover, responses in the *Wrong* class were often submitted well before the 15 s time limit, signaling confidence in the accuracy of the answer, even though it was incorrect. Together, these observations and the rule-based structure and relative simplicity of the multiplication task suggest that the participant was not struggling with the arithmetic, and produced incorrect results with confidence. This false confidence likely explains the similarity in cognitive and ocular behavior between the *Correct* and *Wrong* classes.

Building on this intuition, we next focused on the two-class task, contrasting *Correct* against *Missed* responses, where clearer distinctions were expected. Table 2 confirms this expectation: all models exceeded 81% accuracy, with the LSTM-only and Fusion architectures outperforming CNN-only. The LSTM-only model achieved the highest precision on *Correct* and the highest recall on *Missed* responses. The Fusion model achieved marginally higher overall accuracy (0.866 vs. 0.862 for LSTM) and a very similar F1 score on the Missed class. The Fusion model provided the most balanced performance across the two classes, although it is more complex and computationally more costly than the single-modality models. This suggests that while combining spatial and temporal features can yield comparable or slightly improved performance, the simpler LSTM-only model, leveraging the strong temporal signal, remains highly effective and efficient for this specific task.

A central contribution of this work lies in the spatio-temporal representation of eye-tracking data. Eye-tracking inherently distills ocular behavior into numerical signals, which we re-expressed as compact visualizations enriched with key descriptors. By retaining their temporal structure, these visualizations function as time series, enabling models to capture temporal dynamics alongside spatial features in a informative manner. This representation reduces data complexity while preserving meaningful patterns of cognitive processes, which are temporal in nature ([37, 38, 39]). Thereby, we offer a practical pathway for extending eye-tracking analysis to resource-constrained settings such

as mobile VR, where lightweight headsets must operate under limited computational budgets. Our approach enables scalable, real-time monitoring and study of cognitive performance and failures in immersive environments.

5.1. Limitations and Future Work

Imbalanced dataset A key limitation of this study is the class imbalance inherent in the dataset. Although data downsampling and augmentation strategies were applied, the uneven distribution of responses meant that the models were exposed to a greater variation within the *Correct* class than either *Missed* or *Wrong*. This imbalance undermined the generalizability of the models, especially on the three-class task, and contributed to their tendency to classify ambiguous cases as *Correct*.

Task design Another challenge lies in the overall simplicity of the arithmetic task. The question set was personalized to maintain an optimal level of difficulty—challenging without being discouraging. While this design ensured that the participant invested genuine cognitive effort rather than disengaging, it also meant that most questions remained relatively easy. As a result, samples of the classes *Missed* and *Wrong* reflected mainly slips and lapses of attention and memory, rather than substantive cognitive mistakes. Additionally, this led the participant to produce few incorrect responses, amplifying the class imbalance and limiting the variability of error patterns available to the models.

Sample size Finally, the limited sample size restricts the generalizability of the findings. With a single participant, the results should be interpreted as exploratory rather than conclusive. Nonetheless, this design is consistent with the broader vision of personalized modeling: rather than diluting individual patterns in pooled datasets, training neural networks on single-participant data emphasizes the unique cognitive signatures that underlie error and performance. Future work should extend this approach across larger and more diverse cohorts to determine how individualized models can be generalized or adapted in practice.

6. Conclusion

Our findings show that deep learning models can predict cognitive failure in VR from eye-tracking data, though performance varied across error types. In a longitudinal study spanning 18 months and comprising 50 hours of eye-tracking data, this work demonstrated that VR can serve as a practical platform for monitoring cognitive performance and predicting impending failures in real time. As VR headsets become lighter and biosensors more integrated, we can collect more physiological data—traditionally difficult to obtain—at scale and in naturalistic scenarios that real-life training cannot replicate. Combined with the growing computational power and complexity of machine learning models and AI systems, this enables the extraction of subtle cognitive markers from biosignals, allowing the design of personalized models of cognitive failure to enhance safety and performance in both routine and high-risk environments.

Declaration on Generative AI

During the preparation of this work, the author(s) used Openai Chat-GPT-4.2 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. La Fata, L. Adelfio, R. Micale, G. La Scalia, Human error contribution to accidents in the manufacturing sector: A structured approach to evaluate the interdependence among performance shaping factors, Safety science 161 (2023) 106067. doi:10.1016/j.ssci.2023.106067.

[2] M. A. Makary, M. Daniel, Medical error—the third leading cause of death in the us, BMJ 353 (2016). doi:10.1136/bmj.i2139.

[3] Z. Meihui, D. Licao, C. Wenming, P. Ensheng, Analysis of human errors in nuclear power plant event reports, Nuclear Engineering and Technology 57 (2025) 103687. doi:10.1016/j.net.2025.103687.

[4] P. M. Salmon, M. Regan, I. Johnston, Human error and road transport: phase one: a framework for an error tolerant road transport system, 2005.

[5] J. Reason, Human Error, Cambridge university press, 1990. doi:10.1017/cbo9781139062367.

[6] J. N. K. Wah, Revolutionizing surgery: AI and robotics for precision, risk reduction, and innovation, Journal of Robotic Surgery 19 (2025) 47. doi:0.1007/s11701-024-02205-0.

[7] N. J. Ward, Automation of task processes: An example of intelligent transportation systems, Human Factors and Ergonomics in Manufacturing & Service Industries 10 (2000) 395–408.

[8] T. Wendler, F. W. B. van Leeuwen, N. Navab, M. N. van Oosterom, How molecular imaging will enable robotic precision surgery: the role of artificial intelligence, augmented reality, and navigation, European Journal of Nuclear Medicine and Molecular Imaging 48 (2021) 4201–4224. doi:10.1007/s00259-021-05445-6.

[9] R. J. Compton, D. Gearinger, H. Wild, D. Rette, E. C. Heaton, S. Histon, P. Thiel, M. Jaskir, Simultaneous eeg and pupillary evidence for post-error arousal during a speeded performance task, European Journal of Neuroscience 53 (2021) 543–555. doi:10.1111/EJN.14947.

[10] C. Danielmeier, M. Ullsperger, Post-error adjustments, Frontiers in psychology 2 (2011) 233. doi:10.3389/fpsyg.2011.00233.

[11] C. Wirth, P. Dockree, S. Harty, E. Lacey, M. Arvaneh, Towards error categorisation in bci: single-trial eeg classification between different errors, Journal of neural engineering 17 (2019) 016008. doi:10.1088/1741-2552/ab53fe.

[12] Y. J. Oh, Y. H. Lee, J. H. Yun, A study on the operator's erroneous responses to the new human interface of a digital device to be introduced to nuclear power plants, in: International Conference on Human-Computer Interaction, Springer, 2011, pp. 337–341. doi:10.1007/978-3-642-22098-2_68.

[13] C.-J. Lin, C. Wu, W. A. Chaovalitwongse, Integrating human behavior modeling and data mining techniques to predict human errors in numerical typing, IEEE Transactions on Human-Machine Systems 45 (2014) 39–50. doi:10.1109/THMS.2014.2357178.

[14] N. Takada, T. Laohakangvalvit, M. Sugaya, Human error prediction using heart rate variability and electroencephalography, Sensors 22 (2022) 9194. doi:10.3390/s22239194.

[15] P. Dymora, B. Kowal, M. Mazurek, S. Romana, The effects of virtual reality technology application in the aircraft pilot training process, in: IOP conference series: materials science and engineering, volume 1024, IOP Publishing, 2021, p. 012099. doi:10.1088/1757-899X/1024/1/012099.

[16] J. E. Naranjo, D. G. Sanchez, A. Robalino-Lopez, P. Robalino-Lopez, A. Alarcon-Ortiz, M. V. Garcia, A scoping review on virtual reality-based industrial training, Applied Sciences 10 (2020) 8224. doi:10.3390/app10228224.

[17] G. S. Ruthenbeck, K. J. Reynolds, Virtual reality for medical training: the state-of-the-art, Journal of Simulation 9 (2015) 16–26. doi:10.1057/jos.2014.14.

[18] R. Seil, C. Hoeltgen, H. Thomazeau, H. Anetzberger, R. Becker, Surgical simulation training should become a mandatory part of orthopaedic education, Journal of Experimental Orthopaedics 9 (2022) 22. doi:10.1186/s40634-022-00455-1.

[19] S. P. Marshall, Identifying cognitive state from eye metrics, Aviation, space, and environmental medicine 78 (2007) B165–B175.

[20] D. S. Rudmann, G. W. McConkie, X. S. Zheng, Eyetracking in cognitive state detection for hci, in: Proceedings of the 5th international conference on Multimodal interfaces, 2003, pp. 159–163.

doi:10.1145/958432.958464.

[21] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. S. Tachos, E. Tripoliti, K. Marias, D. I. Fotiadis, M. Tsiknakis, Review of eye tracking metrics involved in emotional and cognitive processes, IEEE reviews in biomedical engineering 16 (2021) 260–277. doi:10.1109/RBME.2021.3066072.

[22] B. Ibragimov, C. Mello-Thoms, The use of machine learning in eye tracking studies in medical imaging: a review, IEEE journal of biomedical and health informatics 28 (2024) 3597–3612. doi:10.1109/JBHI.2024.3371893.

[23] M. Kaczorowska, M. Plechawska-Wójcik, M. Tokovarov, Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features, Brain sciences 11 (2021) 210. doi:10.3390/brainsci11020210.

[24] M. Krol, M. Krol, A novel approach to studying strategic decisions with eye-tracking and machine learning, Judgment and Decision Making 12 (2017) 596–609. doi:10.1017/S1930297500006720.

[25] J. Z. Lim, J. Mountstephens, J. Teo, Eye-tracking feature extraction for biometric machine learning, Frontiers in neurorobotics 15 (2022) 796895. doi:10.3389/fnbot.2021.796895.

[26] A. Rizzo, S. Ermini, D. Zanca, D. Bernabini, A. Rossi, A machine learning approach for detecting cognitive interference based on eye-tracking data, Frontiers in Human Neuroscience 16 (2022) 806330. doi:10.3389/fnhum.2022.806330.

[27] C. Chansri, J. Srinonchat, Utilizing gramian angular fields and convolution neural networks in flex sensors glove for human–computer interaction, IEEE Transactions on Human–Machine Systems 54 (2024) 475–483. doi:10.1109/thms.2024.3404101.

[28] C. Camara, P. Peris-Lopez, M. Safkhani, N. Bagheri, Ecg identification based on the gramian angular field and tested with individuals in resting and activity states, Sensors 23 (2023) 937. doi:10.3390/s23020937.

[29] J.-L. Qiu, X.-Y. Qiu, K. Hu, Emotion recognition based on gramian encoding visualization, in: Brain Informatics (Lecture Notes in Computer Science, vol. 11309), volume 11309 of *Lecture Notes in Computer Science*, Springer, Cham, 2018, pp. 3–12. doi:10.1007/978-3-030-05587-5_1.

[30] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018). doi:10.48550/arXiv.1803.01271, v2.

[31] Z. Wang, T. Oates, Imaging time-series to improve classification and imputation, in: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, AAAI Press, 2015, p. 3939–3945. doi:10.5555/2832747.2832798.

[32] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in neural information processing systems 28 (2015). doi:10.5555/2969239.2969312.

[33] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, pp. 3995–4001. doi:10.5555/2832747.2832806.

[34] J. Qin, F. Liu, et al., GAF-FusionNet: Multimodal ECG Analysis via Gramian Angular Fields and Split Attention, arXiv preprint arXiv:2501.01960v1 (2025). doi:10.32388/I182MY, preprint.

[35] L.-M. Vortmann, J. Knychalla, S. Annerer-Walcher, M. Benedek, F. Putze, Imaging time series of eye tracking data to classify attentional states, Frontiers in Neuroscience 15 (2021). doi:10.3389/fnins.2021.664490.

[36] A. E. Urai, A. Braun, T. H. Donner, Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias, Nature communications 8 (2017) 14637. doi:10.1038/ncomms14637.

[37] C. Freksa, Spatial and temporal structures in cognitive processes, in: Foundations of computer science: potential—theory—cognition, Springer, 2006, pp. 379–387.

[38] M. Maniadakis, P. Trahanias, Temporal cognition: a key ingredient of intelligent systems, Frontiers in neurorobotics 5 (2011) 2.

[39] W. J. Matthews, W. H. Meck, Temporal cognition: Connecting subjective time to perception, attention, and memory., Psychological bulletin 142 (2016) 865.