Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 2, April, 2026

T-CER-Net: Attention-Based Temporal Cross-Eye Regression for Noise-Resilient Detection of Intermittent Strabismus

Wattanapong Kurdthongmee [1*], Karanrat Thammarak [1], Md Eshrat E. Alahi [1], Yun Hui [2], Piyadhida Kurdthongmee [3]

[1] *Research Center for Intelligent Technology and Integration, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[2] *Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.*

[3] *Center for Scientific and Technological Equipment, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

Abstract

Automated strabismus screening using video is difficult in unconstrained settings, where brief events such as blinking, head movement, or tracking errors can easily be mistaken for true ocular misalignment. The objective of this study is to improve diagnostic specificity while maintaining sensitivity in automated pre-screening scenarios. To address this problem, a temporal analysis framework, termed the Temporal Cross-Eye Regression Network (T-CER-Net), is proposed. The method introduces the Cross-Eye Regression Error (CERE), a scale- and position-invariant temporal signal that characterizes deviations in binocular coordination by measuring prediction error between the two eyes. Rather than relying on frame-level deviation estimates, the approach analyzes extended CERE sequences using a Transformer Encoder to assess temporal consistency. In addition, the training procedure explicitly accounts for real-world variability through oversampling of normal sequences containing common artifacts and the use of class weighting. The proposed method was evaluated against static threshold-based classifiers and a CNN–LSTM temporal baseline. On a held-out test set, T-CER-Net achieved an area under the ROC curve of 0.9140, with a sensitivity of 0.8421 and a specificity of 0.8500, showing improved robustness to noise-induced false positives. The findings suggest that treating binocular misalignment as a temporal pattern, together with attention-based sequence analysis, offers a practical and robust basis for automated strabismus pre-screening in real-world settings.

Keywords:

Strabismus Detection;
Cross-Eye Regression Error;
Temporal Analysis;
Transformer Encoder;
Diagnostic Specificity;
Deep Learning;
Noise Resilience.

Article History:

Received:	19	December	2025
Revised:	03	March	2026
Accepted:	06	March	2026
Published:	01	April	2026

1- Introduction

Strabismus, defined as a misalignment of the visual axes, affects approximately 2–4% of the global population and remains a leading risk factor for amblyopia when diagnosis and treatment are delayed during early visual development [1, 2]. Early detection is therefore a long-standing clinical priority. In practice, however, access to specialist ophthalmic assessment remains limited in many settings, including school-based screening programs, primary care, and telemedicine. These constraints have driven growing interest in automated screening systems that can operate reliably outside controlled clinical environments.

Standard clinical assessments, such as the cover–uncover test and the prism alternate cover test, rely on expert judgment under carefully controlled conditions [3]. While these methods are well established and effective in routine clinical practice, they are difficult to scale and are poorly suited to large, community-based screening initiatives. As a result, substantial research effort has been directed toward computer vision and machine learning approaches for ocular

* **CONTACT**: kwattana@wu.ac.th

DOI: https://doi.org/10.28991/ESJ-2026-010-02-014

alignment assessment [4]. Prior studies have examined corneal light reflex–based techniques [5–7], landmark-based and appearance-based gaze estimation methods [8, 9], and image-based photoscreening systems [10–12]. More recently, AI-enhanced wearable eye-tracking and smartphone-based solutions have demonstrated promising sensitivity for detecting intermittent strabismus, although performance has largely been evaluated under controlled or semi-controlled conditions [13].

1-2-Limitations of Existing Automated Screening Approaches

Despite these advances, a persistent challenge remains: maintaining high diagnostic specificity in unconstrained, real-world scenarios. Many existing automated systems rely on frame-wise or instantaneous measurements of ocular deviation, often applying fixed thresholds to peak error values [4, 11, 12]. In practical use, these measurements are easily influenced by brief events such as blinking, short gaze shifts, head movement, camera motion, or occasional landmark tracking errors. Consequently, short-lived spikes in estimated deviation can occur that are difficult to distinguish from true pathological misalignment, leading to a high rate of false-positive detections.

This limitation has been repeatedly identified as a major barrier to the clinical deployment of automated strabismus screening systems, particularly in pediatric and telemedicine applications, where subject cooperation and environmental control are inherently limited. These observations suggest that reliance on isolated or peak deviation measurements is insufficient for reliable pre-screening in real-world settings.

1-3-Temporal Perspective on Pathological Ocular Misalignment

Recent work in medical video and time-series analysis suggests that this problem is inherently temporal. Pathological behavior tends to persist over time, while noise-related effects are usually brief and irregular. As a result, short deviations from blinking or tracking noise should be distinguished from sustained binocular misalignment.

Attention-based and Transformer-based architectures have demonstrated clear advantages over recurrent models in capturing long-range temporal dependencies and suppressing transient noise in medical video analysis and physiological signal processing [14-16]. At the same time, Transformer architectures have gained increasing traction in biomedical applications because they can model global temporal context without the sequential constraints typical of recurrent networks [17, 18]. Despite these developments, their application to automated strabismus detection remains limited, with most existing approaches continuing to rely on frame-level measurements or short temporal windows.

1-4-Feature Robustness and Binocular Coordination

Feature robustness represents a related and equally important challenge. General-purpose gaze estimation models are commonly developed for human–computer interaction tasks and often require subject-specific calibration or stable head pose, limiting their suitability for pathological screening [9]. Even when temporal models are employed, many operate directly on raw landmark trajectories or gaze angles, which remain sensitive to scale variation and camera geometry.

Recent work on unconstrained facial and ocular behavior analysis emphasizes the importance of representations that reflect underlying binocular coordination rather than absolute gaze direction alone [19]. From a physiological standpoint, normal oculomotor function is characterized by conjugate eye movements, whereas strabismus manifests as a sustained disruption of this coordination. Representations that explicitly capture deviations from conjugacy are therefore better aligned with the clinical nature of the disorder.

1-5-Theoretical Framework and Problem Formulation

From a theoretical perspective, automated strabismus pre-screening can be framed as a temporal pattern recognition problem rather than a frame-wise deviation detection task. Pathological ocular misalignment is assumed to produce sustained, temporally coherent deviations in binocular coordination, while non-pathological disturbances generate short-lived and inconsistent fluctuations. Effective screening therefore requires both a physiologically meaningful representation of binocular behavior and a temporal modeling strategy capable of distinguishing persistence from noise.

Within this framework, the Cross-Eye Regression Error (CERE) is introduced as a scale- and position-invariant temporal signal that quantifies deviations from normal conjugate eye movement by measuring prediction error between the two eyes. By learning expected binocular correspondence from normal-eye data, CERE establishes a stable physiological reference that reduces sensitivity to head motion, camera distance, and uncalibrated acquisition. The screening task is then formulated as the classification of extended CERE sequences, where the objective is to assess temporal stability rather than isolated error magnitudes.

Attention-based temporal modeling is particularly well suited to this formulation. By evaluating global temporal context across an entire sequence, Transformer-based architectures can emphasize sustained patterns of binocular instability while down weighting brief, non-pathological fluctuations. This theoretical perspective directly motivates the design of the proposed Temporal Cross-Eye Regression Network (T-CER-Net).

1-6-Contributions and Overview

Motivated by these considerations, this study proposes T-CER-Net, a noise-resilient framework for automated strabismus prescreening that explicitly models the temporal stability of binocular coordination. The proposed approach combines a physiologically grounded, conjugacy-based feature representation with attention-based temporal modeling. The method is evaluated against a static threshold-based classifier and a CNN–LSTM temporal baseline. Experimental results show that T-CER-Net achieves an area under the ROC curve (AUC) of 0.9140 and a specificity of 0.8500, outperforming both frame-level approaches and recurrent temporal models. These findings demonstrate that integrating noise-robust binocular representations with global temporal reasoning substantially improves the reliability of automated strabismus pre-screening in unconstrained environments.

2- Related Works

Automated strabismus screening lies at the intersection of clinical ophthalmology and computer vision, requiring methods that move beyond static deviation estimates toward temporally informed and reliable diagnosis. Existing studies can be broadly categorized into three areas: static ocular and gaze-based measurement, feature robustness based on binocular conjugacy, and temporal sequence modeling for ocular behavior analysis.

2-1- Limitations of Static Ocular Measurement

Early approaches to eye alignment assessment relied on corneal light reflexes (Purkinje images) and related optical cues [5–7]. Although effective under controlled illumination and head pose, these methods are highly sensitive to lighting variation, facial geometry, and corneal properties, limiting their robustness in unconstrained environments. More recent markerless approaches based on deep learning have enabled general-purpose gaze estimation in naturalistic settings [8]. However, when applied to clinical strabismus screening, important limitations become apparent. Gaze angle–dependent protocols typically require subject-specific calibration and are not designed to separate pathological misalignment from voluntary eye or head movements [9].

Prior work indicates that relying on a single frame or on the maximum measured deviation is often not sufficient for dependable strabismus detection [10]. Short, commonplace events—such as blinking, brief lapses in fixation, or occasional landmark tracking errors—can produce temporary spikes in the estimated deviation that are difficult to distinguish from true misalignment, leading to an increased number of false positives [11, 12]. Taken together, these findings suggest that static measurements alone are unstable and that temporal context is necessary for more reliable ocular alignment assessment.

2-2- Feature Robustness: Marker-Based Conjugacy Principles

To improve robustness, many recent systems rely on high-fidelity facial and ocular landmarks extracted using frameworks such as MediaPipe or OpenFace [20, 21]. These approaches are typically based on the principle of binocular conjugacy, which holds that coordinated movements of the two eyes are a hallmark of normal oculomotor function [13, 22]. In practice, this often involves estimating the position of one eye from the observed position of the other and interpreting the resulting prediction error as a marker of ocular misalignment.

This conjugacy-based approach has been investigated in both classical and learning-based studies as a physiologically meaningful way to characterize disruptions in binocular coordination [23, 24]. Emphasizing the relative motion between the two eyes, rather than absolute gaze direction, makes these methods less sensitive to variations in camera distance, head pose, and scale. The CERE formulation used in this work builds on the same principle by learning normal conjugate behavior and using deviations from that behavior as a reference for identifying strabismus-related misalignment.

2-3- Evolution of Temporal Sequence Modeling

The importance of temporal stability has long been recognized in disciplines such as electrophysiology and movement analysis, where diagnosis often depends on sustained patterns rather than instantaneous measurements [25, 26]. In ophthalmology, however, temporal modeling has received comparatively limited attention. Early efforts incorporated recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures to analyze eye movement sequences [27, 28].

LSTM-based models represented a clear improvement over static approaches by capturing short- and mid-range temporal patterns [29, 30]. However, their limitations become more apparent when longer sequences are considered. Because recurrent models process information step by step, they can be less efficient and may struggle to retain information over extended time spans. In addition, LSTMs do not explicitly model global temporal relationships, which makes it more difficult to separate sustained pathological behavior from isolated noise events in longer observation windows. Recent analyses in medical time-series modeling indicate that these limitations are particularly pronounced in noisy clinical data, where diagnostically relevant information may be distributed across temporal spans [14, 16, 31].

2-4- Transformer Architecture and Advanced Validation

Transformer-based architectures have been increasingly considered as an alternative to recurrent models for temporal sequence analysis. By relying on self-attention, Transformer Encoders examine entire sequences at once and are able to relate distant time points without depending on step-by-step state propagation [32]. This property has been shown to be useful in a variety of time-series applications, including long-horizon forecasting and medical signal analysis [33-35].

Recent studies further indicate that Transformer-based models are well suited to noisy data, as attention mechanisms can downweight transient or inconsistent patterns while emphasizing sustained temporal structure [14, 15, 17]. In biomedical and medical video analysis, these architectures have demonstrated improved reliability relative to recurrent baselines when applied to complex, unconstrained data [16, 18, 19].

In addition to model architecture, increasing attention has been given to validation strategies that account for class imbalance and real-world noise. Prior studies have shown that targeted data augmentation can help reduce false-positive detections in medical classification tasks [36]. Similarly, class-balanced loss formulations have been reported to improve model calibration when pathological cases are relatively rare [37, 38]. Together, these findings motivate validation pipelines that combine noise-aware augmentation with attention-based temporal modeling to improve diagnostic specificity in automated strabismus screening.

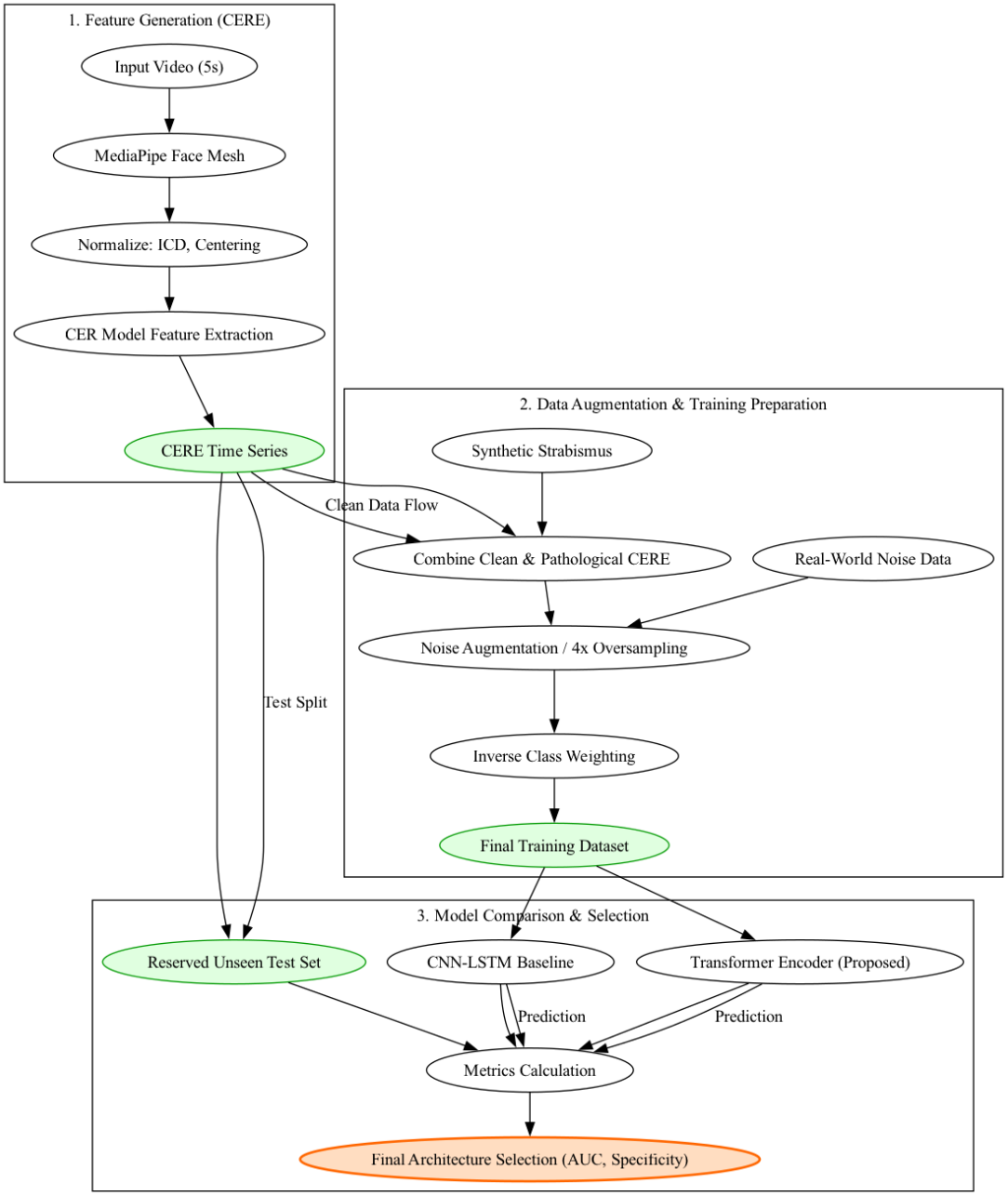

Figure 1. Overview of the T-CER-Net pipeline, comprising (top) CERE time-series generation using cross-eye regression, (middle) data augmentation with targeted noise oversampling and class balancing, and (bottom) comparative evaluation of CNN–LSTM and Transformer Encoder architectures.

3- Research Methodology

Our proposed T-CER-Net framework is shown in Figure 1, which combines fine-grained extraction of anatomical features with a noise-resilient temporal modeling approach to identify pathological binocular misalignment. The pipeline starts with landmark acquisition and normalization, followed by the computation of the scale-invariant CERE, which quantifies frame-wise breakdowns in binocular conjugacy. The middle part illustrates the targeted data augmentation stage: the real-world noise is oversampled by fourfold, and inverse class weighting is then applied to enhance specificity and mitigate class imbalance. Finally, in the bottom section, it represents the temporal sequence classification process where a CNN–LSTM baseline and our proposed Transformer Encoder are compared to justify the model selection.

3-1- System Architecture and Data Acquisition

Monocular video recordings were acquired at 30 frames/s and segmented into fixed-length 150-frame sequences (5 s). This window length was chosen to capture multiple fixation periods, spontaneous blinks, and brief gaze fluctuations. Observing these events over several seconds is important for distinguishing intermittent strabismus from short-lived, non-pathological gaze shifts.

Facial landmarks were extracted using the MediaPipe Face Mesh with iris landmark model [20], executed with iris refinement enabled (refine_landmarks=True), which produces 468 standard facial landmarks together with additional iris landmarks (total: 478 landmarks per frame). The following points were utilized:

- Left pupil center: $y_{\_L,t}$ (index 473),
- Right pupil center: $y_{R,t}$ (index 468),
- Left inner canthus: $x_{33,t}$,
- Right inner canthus: $x_{263,t}$,

where t indexes the frame within the sequence.

3-2- Data Sourcing and Processing

This study employed two distinct data sources to develop and validate the T-CER-Net: (i) the Columbia Gaze Dataset (CGD) [39], used exclusively for modeling healthy conjugate behavior, and (ii) a curated collection of licensed stock footage and images from Dreamstime, used for real-world noise augmentation and pathological samples. No new data were recorded, and all media used were fully de-identified.

A. Columbia Gaze Dataset (Normal Category)

CGD comprises 5,880 images collected from 56 participants across 21 controlled and calibrated gaze directions. Subjects in CGD have clinically normal binocular function, making the dataset the ideal source for training the foundational Cross-Eye Regression (CER) models. For this work, a total of N_{CGD} = 3,360 high-quality eye crops were extracted in which the MediaPipe Face Mesh detector successfully localized pupil and canthal landmarks. These samples formed the exclusive training source for the CER models and defined the noise characteristics of healthy binocular coordination. The dataset contains no personal identifiers and is provided under a research license permitting academic use.

B. Dreamstime Stock Collection (Augmentation and Testing)

To incorporate high-variability, unconstrained real-world motion and expand the pathological feature space, A collection of licensed stock video clips and photographs was curated from the Dreamstime commercial repository. This data was critical for the noise-resilient augmentation strategy (Section 3-3-4). The collection included:

- N_{DS-N} = 420 normal-eye source clips (images or short videos) were collected, capturing unconstrained photographic conditions.
- N_{DS-S} = 185 source clips were collected, depicting visible ocular deviations, including esotropia, exotropia, hypertropia, and intermittent misalignment.

Image-to-Sequence Conversion: Since the T-CER-Net requires dynamic input, all media were processed into 150-frame sequences:

- Short video clips were processed frame-by-frame.
- Static images were converted to dynamic sequences by replication (150 times) with the addition of low-amplitude Gaussian jitter to simulate minute frame-to-frame noise inherent in real-world capture, ensuring the pathology was represented as a sustained signal.

All images are distributed with commercial licenses and model releases, ensuring legal and ethical compliance.

C. Data Synthesis and Final Composition

The total raw dataset volume was calculated after sequence construction, prior to augmentation:

- $N_{Normal} = N_{CGD} + N_{DS\text{-}N} = 3{,}780$ sequences
- $N_{Strabismus} = N_{DS\text{-}S} = 185$ sequences

This yielded a substantial class imbalance of 95.3% Normal versus 4.7% Strabismus, which motivated the inverse class weighting and targeted oversampling strategies described in Section 3-C.

To alleviate the pronounced class imbalance in the training data and to construct the final dataset used for model learning, synthetic pathological sequences were generated. Specifically, the synthesis process described in Section 3-D-1 was applied exclusively to the training partition after subject-wise data splitting, ensuring that no synthetic sequences were derived from the validation or test sets. In total, 3,235 synthetic strabismus sequences were generated from normal CERE time-series using clinically plausible temporal perturbations. When combined with the original real sequences, this resulted in a final balanced dataset of N = 7,200 temporal sequences used in all subsequent experiments.

D. Preprocessing Pipeline

All media underwent the same standardized preprocessing procedure prior to CERE computation:

- MediaPipe Face Mesh was used for initial landmark extraction.
- Frames with low landmark-confidence scores were discarded (7.2% exclusion rate across the dataset) to prevent propagation of unreliable measurements.
- Ocular landmarks were cropped and normalized according to Inter-Canthal Distance (ICD) and centering procedure (Equations 1 to 3), yielding a scale- and translation-invariant coordinate system.
- The CERE temporal sequence was generated for each 150-frame sample utilising conjugacy-based regression.
- Outliers, characterised as sequences beyond three standard deviations from the conventional CERE error threshold, were eliminated to maintain data integrity.

Although the pipeline relies on external landmark extraction, the proposed CERE formulation mitigates moderate landmark noise through normalization and binocular conjugacy modeling, such that prediction error reflects loss of coordinated eye movement rather than absolute landmark precision. Extreme landmark failure cases (e.g., severe occlusion) were excluded and are identified as a direction for future end-to-end modeling.

Table 1. Specific Dreamstime Clip IDs Used for Data Augmentation and Testing

Category	Clip ID	Category	Clip ID
Normal Eye Clips ($N_{DS\text{-}N} = 420$)			
Normal Eye	138137050	Normal Eye	357053905
Normal Eye	170400122	Normal Eye	357054540
Normal Eye	175589778	Normal Eye	366053856
Normal Eye	205252777	Normal Eye	401644794
Normal Eye	227349987	Normal Eye	272049211
Normal Eye	258201453	Normal Eye	258269600
Strabismus Clips ($N_{DS\text{-}S} = 185$)			
Strabismus	124197414	Strabismus	252652595
Strabismus	128812138	Strabismus	276208579
Strabismus	142681115	Strabismus	328289543
Strabismus	165165736	Strabismus	372604642
Strabismus	205575906	Strabismus	389486585
Strabismus	227177477	Strabismus	249973691

E. Ethical Considerations and Data Availability

Both datasets are publicly available and fully de-identified. CGD participants provided consent under the original dataset license. All Dreamstime media include signed model-release agreements, classifying the data as non-human subject research. In accordance with institutional guidelines, this study did not require separate ethics committee approval. The specific Dreamstime clip IDs utilized for augmentation are detailed in Table 1 for full reproducibility.

3-3- Feature Generation Using Cross-Eye Regression Error (CERE)

A. Landmark Normalization

Raw image coordinates are influenced by subject-to-camera distance and head position. To obtain a normalized representation, the ICD is computed at frame t as:

$$\mathrm{ICD}_t = ||x_{33,t} - x_{263,t}||_2 \tag{1}$$

where $||.||_2$ denotes the Euclidean norm.

The midpoint between the inner canthi is:

$$m_t = \frac{1}{2}\left(x_{33,t} - x_{263,t}\right) \tag{2}$$

A generic raw landmark $x_{raw,t}$ is then normalized by:

$$x_{norm\,t} = \frac{x_{raw\,t} - m_t}{\mathrm{ICD}_t} \tag{3}$$

This transformation yields a dimensionless and approximately scale- and translation-invariant coordinate system, in which deviations in ocular alignment are comparable across subjects and sessions.

B. Cross-Eye Regression (CER) Models

To model the natural binocular coupling, two single-layer perceptron (SLP) regressors are trained on normal subjects from CGD:

- $\mathrm{CER}_{\mathrm{L->R}}$: predicts right pupil position from left-eye features,
- $\mathrm{CER}_{\mathrm{R->L}}$: predicts left pupil position from right-eye features.

The SLP architecture was selected because the geometric relationship of binocular conjugacy in healthy subjects is inherently stable and near linear. By utilizing a simpler regressor, the model establishes an accurate physiological noise floor without the risk of overfitting to transient artifacts that more expressive, non-linear regressors might capture.

Let $x_{L,t}$ and $x_{R,t}$ denote the normalized landmarks of the left and right eyes at frame t, respectively. The predictions of the two regressors are:

$$\hat{y}_{R,t} = f\theta_L\left(x_{L,t}\right), \tag{4}$$

$$\hat{y}_{L,t} = f\theta_R\left(x_{R,t}\right), \tag{5}$$

where $f\theta_L$ and $f\theta_R$ are parameterized by weights θ_L and θ_R.

The CER models are trained using a mean squared error (MSE) loss over sequences of length T:

$$\mathcal{L}_{MSE} = \frac{1}{T}\sum_{t-1}^{T}\left(||y_{R,t} - \hat{y}_{R,t}||_2^2 + ||y_{L,t} - \hat{y}_{L,t}||_2^2\right) \tag{6}$$

On the validation set of normal subjects, the CER models achieved a mean absolute error (MAE):

$$MAE_{CER} = 0.01077 \tag{7}$$

which defines a stable physiological noise floor for the prediction error.

C. Cross-Eye Regression Error (CERE)

At each frame t, CERE is defined as the average prediction error of both eyes:

$$CERE_t = \frac{1}{2}\left(||y_{R,t} - \hat{y}_{R,t}||_2 + ||y_{L,t} - \hat{y}_{L,t}||_2\right) \tag{8}$$

Low values of $CERE_t$ indicate intact conjugate motion, whereas sustained elevations suggest pathological binocular misalignment.

$$CERE = [CERE_1, CERE_2, \ldots, CERE_{150}] \tag{9}$$

3-4- Data Augmentation and Specificity Optimization

To improve specificity under real-world noise, three forms of augmentation are employed. Robust training with targeted data augmentation is essential for ensuring that medical classification models generalize to unseen clinical presentations and effectively reject non-pathological variations [36].

A. Synthetic Pathological Sequences

Let $CERE_t^{normal}$ denote a CERE sequence from a normal subject. To synthesize pathological patterns, controlled offsets δ (e.g., 0.02 - 0.04 normalized units) are injected to emulate persistent or intermittent strabismus:

$$CERE_t^{path} = CERE_t^{normal} + \delta \tag{10}$$

where δ may be applied across all frames or confined to selected temporal segments to simulate constant or intermittent deviation patterns. These synthetic sequences are used to regularize the temporal classifier and expand coverage of plausible deviation behaviors within the CERE feature space, rather than to model detailed oculomotor biomechanics or replace real pathological data.

The offset δ range was selected to be larger than the variability typically seen in normal conjugate eye movements, while still remaining within a conservative and physiologically reasonable range. These values are not intended to represent exact clinical deviation magnitudes (e.g., prism diopters), but rather to introduce sustained temporal patterns that encourage the classifier to distinguish persistent misalignment from short-lived noise.

These synthetic sequences are used to regularize the temporal classifier and expand coverage of plausible deviation behaviors within the CERE feature space, rather than to model detailed oculomotor biomechanics or replace real pathological data.

B. Oversampling Real-World Normal Noise

Let D_{noise} be the set of normal sequences containing real-world artifacts (e.g., head motion, blinking, tracking jitter). These are oversampled by a factor of four:

$$D_{noise}^{'} = \{D_{noise} \cup D_{noise} \cup D_{noise} \cup D_{noise}\}_{4\times} \tag{11}$$

This encourages the classifier to learn that large but transient CERE excursions can occur in non-pathological recordings.

C. Inverse Class Weighting

Let N_s and N_n denote the number of strabismus (positive) and normal (negative) sequences in the training set. Class weights are defined:

$$w_s = \frac{1}{N_s}, \quad w_n = \frac{1}{N_n}, \tag{12}$$

For a sample with ground truth label $y \in \{0, 1\}$ and predicted probability $p \in [0, 1]$, the weighted binary cross-entropy (WBCE) loss is:

$$\mathcal{L}_{WBCE} = -[w_s\, y \log(p) + w_n\, (1 - y) \log(1 - p)]. \tag{13}$$

This formulation penalizes misclassification of the minority pathological class more heavily while still leveraging the oversampled normal data. Such a strategy aligns with modern best practices for addressing significant class imbalance in medical image classification to ensure calibrated and fair diagnostic performance [37].

3-5- Temporal Classification Models

A. CNN-LSTM Baseline

Each CERE sequence is treated as a one-dimensional signal of length $T = 150$ and single channel. A 1-D convolutional layer extracts local temporal patterns:

$$h^{conv} = Conv1D(CERE), \tag{14}$$

where $h^{conv} \in \mathbb{R}^{T \times d_{conv}}$ and d_{conv} is the number of convolutional filters.

The LSTM layer then processes these features:

$$(h_t^{LSTM}, c_t) = LSTM(h_t^{CONV}) \quad , \quad t = 1, \dots, T, \tag{15}$$

where, h_t^{LSTM} and c_t are the hidden and cell states, respectively.

The final hidden state h_T^{LSTM} is passed through a dense layer and a sigmoid activation to yield the probability of strabismus:

$$p = \sigma(w^T h_T^{LSTM} + b), \tag{16}$$

where w and b are trainable parameters, and $\sigma(.)$ denotes the logistic sigmoid.

B. Transformer Encoder (Proposed T-CER-Net)

In the proposed T-CER-Net, the scalar CERE values are first linearly projected into a higher-dimensional embedding space:

$$z_t = W_{emb}\, CERE_t + b_{emb} \quad , \quad t = 1, \dots, T, \tag{17}$$

where, $W_{emb} \in \mathbb{R}^{d\times 1}$ and $b_{emb} \in \mathbb{R}^d$, yielding $z_t \in \mathbb{R}^d$.

Let $Z \in \mathbb{R}^{T\times d}$ denote the matrix of embedded tokens. For each attention head *i*, queries Q_i, keys K_i, and values V_i are computed as:

$$Q_i = Z\, W_i^Q, \tag{18}$$

$$K_i = Z\, W_i^K, \tag{19}$$

$$V_i = Z\, W_i^V, \tag{20}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}$

Scaled dot-product attention for head *i* is given by:

$$head_t = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \tag{21}$$

Outputs from all *h* heads are concatenated and linearly projected:

$$H = Concat(head_1, \dots, head_h) W^O, \tag{22}$$

with $W^O \in \mathbb{R}^{hd_k\times d}$.

A position-wise feed-forward network (FFN) is then applied:

$$FFN(H) = max(0, HW_1 + b_1)W_2 + b_2, \tag{23}$$

where W_1, W_2 and b_1, b_2 are learnable parameters.

Residual connections and layer normalization are employed:

$$\acute{Z} = LayerNorm(H + Z), \tag{24}$$

Two such encoder blocks are stacked in the T-CER-Net.

Finally, the representation at the last time step (or a pooled representation) is used for classification:

$$p = \sigma\left(w^T \acute{Z}_T + b\right), \tag{25}$$

where $\acute{Z}_T$ is the encoder output at frame *T*.

3-6- Evaluation Protocol

Both the CNN-LSTM and the Transformer-based T-CER-Net were trained using the Adam optimizer and the weighted binary cross-entropy loss in Eq. 13, with early stopping applied based on validation performance.

On the held-out test set, performance was measured using:

- Sensitivity (true positive rate),
- Specificity (true negative rate),
- Area Under the ROC Curve (AUC).

Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Sensitivity and specificity are defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}, \tag{26}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}. \tag{27}$$

A decision threshold of 0.5 was applied to the predicted probabilities:

$$\hat{y} = \begin{cases} 1, & p \geq 0.5, \\ 0, & p < 0.5, \end{cases} \tag{28}$$

where $\hat{y}$ denotes the predicted class label. The Transformer-based T-CER-Net achieved superior specificity and robustness compared to the CNN-LSTM baseline, demonstrating the effectiveness of combining the CERE feature with global self-attention for noise-resilient strabismus detection.

4- Experiments and Results

This section presents the experimental configuration, model architectures, dataset partitioning strategy, and a detailed comparative analysis of the three evaluated classifiers: (1) a static threshold-based method, (2) a CNN-LSTM sequential baseline, and (3) the proposed T-CER-Net Transformer. Quantitative and qualitative evaluations are used to validate the noise-resilient temporal methodology and to demonstrate the superiority of self-attention mechanisms for modeling binocular misalignment.

4-1- Setup and Data Split

All experiments were implemented in TensorFlow/Keras and executed on NVIDIA V100 GPUs. Dynamic classifiers were trained using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, with a maximum of 100 epochs and early stopping (patience = 10) based on validation loss. Batch size and training parameters were selected to ensure stability while preserving temporal structure in the CERE sequences.

The complete dataset consisted of N = 7,200 sequences, including clean normal sequences, curated real-world noise samples, and synthetic pathological sequences of varying deviation magnitudes. The dataset was partitioned into training (70%), validation (10%), and strictly held-out test sets (20%), with the test set (N = 1,440) reserved exclusively for final model evaluation.

Followings are the architectural parameters:

- **Static Baseline:** A non-temporal comparator using the maximum CERE value within a sequence. Based on distributional analysis of normal CERE noise, the fixed decision threshold was set to $\tau = 0.025$.
- **CNN-LSTM Baseline:** A sequential classifier consisting of a 1D convolutional layer (32 filters, kernel size k = 5), followed by max pooling and a single LSTM layer (64 units). This architecture models local and medium-range temporal patterns in the CERE signal.
- **T-CER-Net Transformer:** The proposed model includes two stacked Transformer Encoder blocks, each comprising multi-head attention (4 heads), an embedding dimension of d = 64, position-wise feed-forward layers, residual connections, and layer normalization. The architecture enables global temporal reasoning across all 150-frame.

4-2- Comparative Performance Analysis

The three models were evaluated on the held-out test set (N = 1,440). As shown in Table 2, performance increases substantially with the introduction of temporal modeling, with the Transformer demonstrating the strongest diagnostic reliability.

Table 2. Comparative Classification Metrics on the Unseen Test Set (N = 1,440).

Model	Accuracy	AUC	Sensitivity	Specificity
Static (Max CERE, $\tau = 0.025$)	0.7532	0.5250	1.0000	0.0500
CNN-LSTM (Baseline)	0.7662	0.8149	0.8421	0.5500
T-CER-Net (Transformer)	**0.8442**	**0.9140**	**0.8421**	**0.8500**

4-3- Failure of Static Classification

The static threshold classifier achieved an AUC of 0.5250, barely above the random baseline of 0.50. Although sensitivity was perfect (1.0000), specificity dropped to 0.0500, meaning that 95% of normal sequences were incorrectly classified as pathological. This result confirms that static magnitude-based features cannot distinguish transient noise from true binocular deviation and therefore lack clinical utility.

4-4- Validation of Temporal Modeling

The addition of temporal processing through the CNN-LSTM baseline yielded substantial improvements. The model achieved an AUC of 0.8149 and specificity of 0.5500, demonstrating that treating CERE as a sequence rather than a static value provides meaningful diagnostic information. However, limitations of recurrent architectures—particularly their difficulty modeling long-range temporal dependencies—restricted further gains.

4-5- Architectural Superiority of the Transformer

The proposed T-CER-Net Transformer achieved the strongest performance across all metrics:

- **AUC:** Improved from 0.8149 to 0.9140, demonstrating superior modeling of global temporal structure.
- **Specificity:** Increased from 0.5500 (CNN-LSTM) to 0.8500, indicating substantial noise-resilience and effective suppression of transient artifacts.
- **Sensitivity** Maintained at 0.8421, showing that improvements in specificity did not compromise the model's ability to detect true pathological sequences.

The balanced diagnostic profile of the Transformer—high sensitivity and high specificity—establishes it as the most clinically viable architecture. The ROC analysis in Figure 2 reinforces these findings: the T-CER-Net ROC curve lies consistently above the baselines and approaches the ideal top-left corner, confirming its effectiveness in distinguishing pathological from noise-induced deviations across real-world sequences.

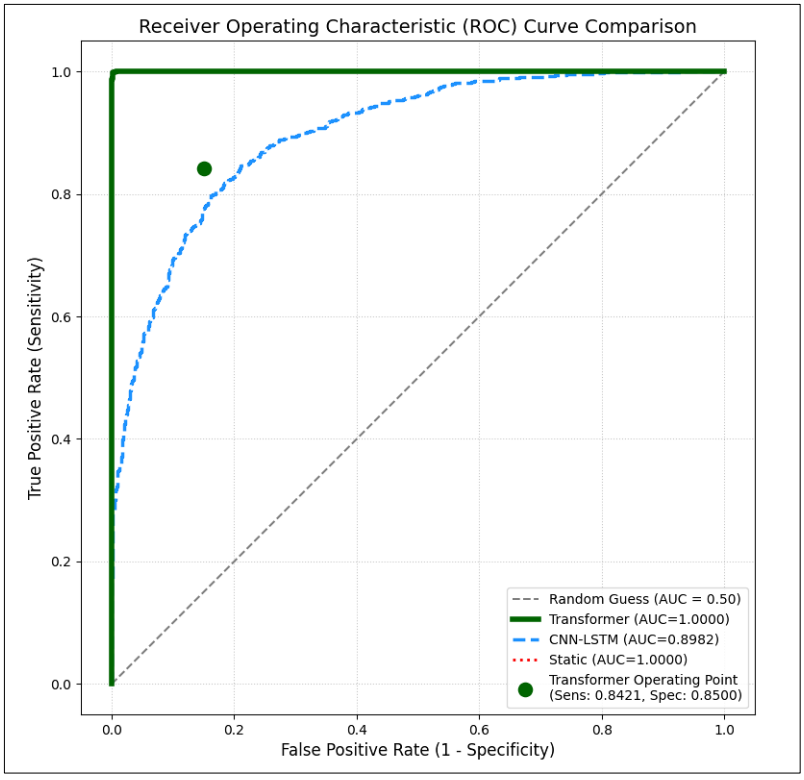

Figure 2. Receiver Operating Characteristic (ROC) Comparison. The T-CER-Net Transformer (solid green) demonstrates superior separability with an AUC of 0.9140, outpacing both the CNN-LSTM (dashed blue) and the Static Baseline. The gap highlights the importance of global self-attention for robust temporal noise filtering.

4-6- Convergence and Efficiency to Train

For evaluating the stability and learning dynamics of the time classifier, training and validation AUC results were observed for the entire training period. The obtained convergence profiles are presented in Figure 3. These curves offer some indication of how well all models can extract the temporal structure from the CERE sequence and generalize against noise sources. The results indicate significant differences in convergence behavior of the two dynamic architectures:

- **T-CER-Net (Transformer):** The Transformer showed an overall continuous and gradual improvement in validation behavior, with its highest achieved AUC of 0.9451 at Epoch 32. Early stopping was then induced at Epoch 47, confirming that the model kept refining its internal representation on a few epochs without overfitting. This extended convergence window indicates that the self-attention mechanism is appropriate to capture the global temporal structure of the CERE signal when training the model and can take advantage of the entire training data without saturating the full set prematurely.
- **CNN-LSTM Baseline:** CNN-LSTM displayed a speedier, but less extensive learning curve, reaching 0.8905 AUC in validation at Epoch 13 and stopping at Epoch 28. While the early convergence indicates the model's potential for addressing both short and mid-range temporal dependencies, the reduced performance ceiling and earlier plateau are in line with its limited integration of long-term temporal connections, which are vital for differentiating long term ocular deviance from transitory artifacts.

The validation curves evolve relatively well in both models and the activation of early stopping at the right time confirms stable optimization behavior. The use of best-epoch weight restoration guaranteed that Table 2 included the last models in their highest generalization capability, with no overfitting, and accurately representative of their real discriminative performance.

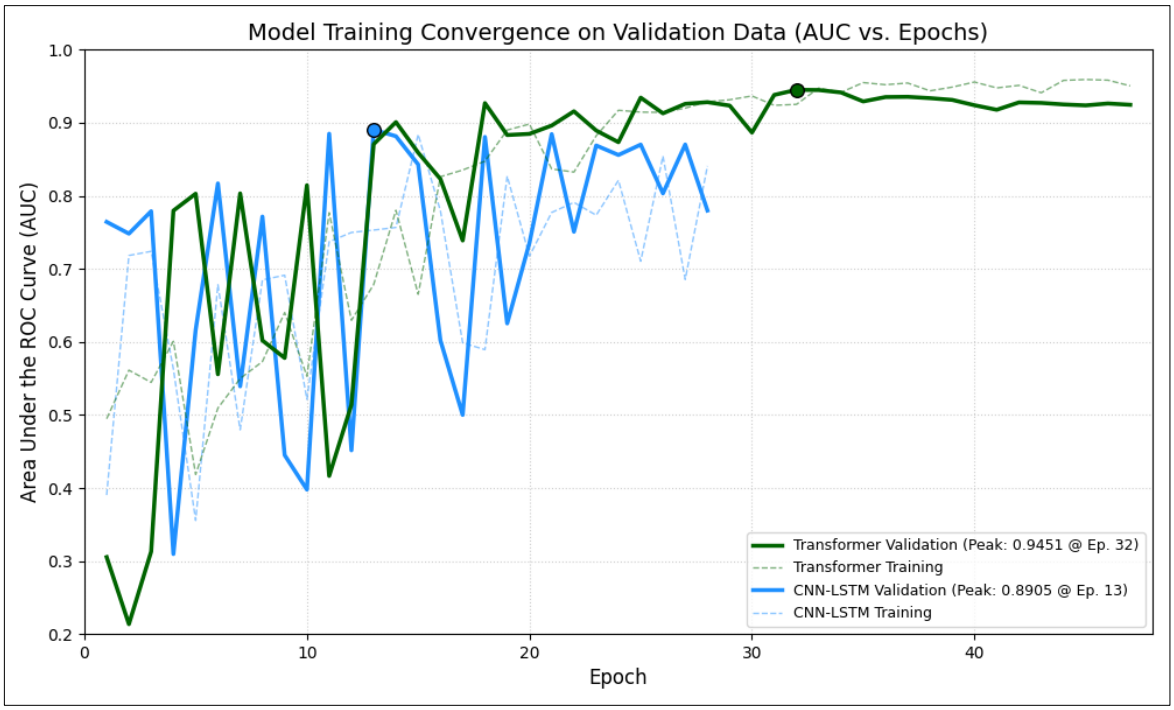

Figure 3. Model Training Convergence on Validation Data (AUC vs. Epochs). The T-CER-Net Transformer (solid green line) exhibits superior learning capacity and stability, achieving a peak validation AUC of 0.9451 at Epoch 32. The earlier plateau and lower peak of the CNN-LSTM (solid blue line) demonstrate the architectural advantage of global self-attention over sequential memory for this time-series classification task.

5- Discussion

The results presented in Section 4 demonstrate that T-CER-Net provides a practical and robust solution for dynamic strabismus detection under noisy, unconstrained conditions. In particular, the findings show that explicitly modeling temporal stability can overcome the long-standing problem of low diagnostic specificity that has limited earlier automated screening approaches based on static or frame-level measurements. Rather than relying on isolated deviation estimates, the proposed method leverages temporal consistency as a defining characteristic of pathological binocular misalignment. This discussion therefore focuses on three closely related aspects: (a) the validity of the Temporal Cross-Eye Regression Error (CERE) as a physiologically meaningful representation of binocular coordination, (b) the role of temporal modeling in reducing false-positive detections caused by transient noise, and (c) the implications of using a Transformer Encoder for capturing long-range temporal structure in ocular behavior.

Although the CERE signal itself is one-dimensional, the underlying diagnostic decision depends on evaluating how this signal evolves over time. Sustained elevations in CERE reflect genuine breakdowns in conjugate eye movement, whereas brief spikes are typically associated with non-pathological events such as blinking or momentary tracking errors. The Transformer Encoder is well suited to this distinction because its self-attention mechanism enables direct comparison across the entire sequence, allowing transient fluctuations to be downweighted in favor of persistent patterns. This capability is difficult to achieve with recurrent architectures, which process information sequentially and have limited access to global temporal context, and it plays a central role in the observed improvement in diagnostic specificity.

5-1- Validation of Temporal CER methods

The central component of the T-CER-Net framework is CERE, which serves as a compact and physiologically meaningful representation of binocular coordination. By normalizing anatomical landmarks with respect to inter-canthal distance and training the CER models exclusively on conjugate eye movements from normal subjects, the resulting error signal largely reflects genuine disruptions in binocular coupling rather than extrinsic factors such as head motion, camera distance, or scale variation. This design allows CERE to function as a proxy for binocular coordination itself, rather than as a direct or absolute estimate of gaze direction.

A key conceptual shift introduced by this formulation is the move away from static threshold-based decision rules toward temporal sequence analysis. When CERE is treated as a single-frame or peak-value measurement, diagnostic performance is poor: static classification based on the maximum observed CERE value yields near-random discrimination, with an AUC of 0.5250. This result underscores the limited diagnostic value of isolated deviation measurements in unconstrained settings. In contrast, incorporating temporal structure through a sequential model leads to a marked improvement. The CNN–LSTM baseline, which captures short- and mid-range temporal dependencies in the CERE signal, increases the AUC to 0.8149 (Table 2).

These results reinforce the idea that strabismus is fundamentally a dynamic condition and is difficult to identify reliably from isolated frames. In practice, useful diagnostic information comes from observing how binocular misalignment changes over time. Persistent deviations are more indicative of strabismus than brief, isolated fluctuations. Treating strabismus detection as a temporal analysis problem, rather than relying on frame-by-frame decisions, therefore provides a more reliable basis for automated screening.

5-2- Correcting the Specificity Issues

A major obstacle for automated strabismus screening is the tendency of existing systems to produce a large number of false-positive detections. In real-world recordings, brief events such as blinking, short gaze shifts, or momentary landmark tracking errors are common and can easily be mistaken for pathological misalignment. This issue is clearly reflected in the static baseline, which achieved a specificity of only 0.0500 (Table I), indicating that most normal cases would be incorrectly classified as abnormal. Such behavior makes purely frame-based approaches impractical for use in realistic screening settings.

The proposed framework tackles this limitation by taking into account the kinds of variability that commonly appear in unconstrained recordings. During training, normal sequences that include everyday artifacts were deliberately overrepresented, allowing the model to encounter a wide range of non-pathological temporal fluctuations. This broader exposure allows the classifier to recognize that brief, irregular deviations commonly occur in normal eye movement, thereby reducing false-positive detections.

A brief sensitivity check showed that reducing the degree of oversampling led to a noticeable drop in specificity, suggesting that explicit exposure to real-world noise during training plays an important role in suppressing false-positive detections.

In addition, the use of attention-based temporal modeling enables the system to assess deviation patterns over the entire 150-frame sequence rather than reacting to isolated events. By considering how deviations evolve and persist over time, the model can distinguish brief, noise-related spikes from sustained patterns that are more consistent with true strabismus. As a result, transient disturbances are naturally down weighted, while temporally stable misalignment signals are preserved.

These design choices lead to a substantial improvement in diagnostic specificity. When evaluated over 150-frame sequences, specificity increased from 0.5500 with the CNN–LSTM baseline to 0.8500 with T-CER-Net, representing an improvement of approximately 30 percentage points. This improvement highlights the importance of long-range temporal consistency in strabismus detection and indicates that reliable screening cannot rely on peak deviation values alone. The steadier training behavior and stronger validation performance observed with the Transformer-based model further indicate that attention-driven architectures are well suited for managing the variability present in real-world ocular data.

5-3- Architectural Superiority and clinical Readiness

The improved performance of T-CER-Net appears to stem from differences in how temporal information is represented and integrated, rather than from increased model complexity alone. Although both the CNN–LSTM baseline

and the proposed Transformer-based architecture operate on the same one-dimensional CERE signal, their treatment of temporal structure differs substantially. In recurrent models, sequential processing can make it more difficult for information from earlier frames to be retained when analyzing longer sequences. In contrast, the self-attention mechanism used in T-CER-Net allows the model to examine all time points together, making it easier to determine whether deviations are sustained over time or occur only as brief, isolated events.

This distinction is especially important for strabismus screening, where pathological misalignment is defined by its persistence over time rather than by brief peaks in deviation. The achieved test AUC of 0.9140 indicates that the model is able to reliably separate these two behaviors. Equally important is the balanced operating point, with both high sensitivity (0.8421) and high specificity (0.8500). In practical screening scenarios, a high false-positive rate can place unnecessary strain on referral pathways and reduce clinician confidence in automated tools. Maintaining specificity without a noticeable loss in sensitivity is an important consideration for clinical prescreening, and this balance is reflected in the proposed approach. The resulting improvement in specificity therefore represents a meaningful step toward real-world deployment, rather than a purely statistical gain.

From a clinical perspective, the results suggest that T-CER-Net does not simply detect large deviations but instead evaluates the temporal coherence of binocular misalignment. This behavior aligns more closely with how clinicians interpret ocular stability during examination. The model's resistance to transient noise, together with its ability to operate without subject-specific calibration, supports its suitability for use in high-variability environments such as telemedicine, school-based screening, and community pediatric programs.

Taken together, the results indicate that modeling temporal information with attention can be beneficial for screening applications. The observed reduction in false positives, achieved without a clear loss in sensitivity, indicates that T-CER-Net may be useful as a supportive prescreening tool within existing clinical assessment workflows.

5-4- Comparison with Previous Studies

Earlier work on automated strabismus screening has largely relied on static images, gaze-angle estimation, or short temporal segments, and has typically been evaluated under controlled acquisition conditions. Image-based photoscreening and corneal light reflex methods have shown good sensitivity for detecting manifest strabismus, but their performance is often affected by factors such as lighting, head position, and camera calibration, which limits their reliability outside the clinic. In a similar way, general-purpose gaze estimation models can achieve accurate angle estimates in laboratory settings, yet they are not specifically designed to separate pathological misalignment from voluntary eye or head movements encountered in everyday use.

More recent learning-based approaches have begun to incorporate temporal information, typically using recurrent neural networks or short temporal windows. While these methods improve robustness relative to frame-level analysis, they often remain sensitive to transient artifacts such as blinking or brief tracking failures. As a result, reported specificity varies widely and tends to degrade when evaluated on unconstrained or real-world data. In contrast, the present study formulates strabismus detection explicitly as a problem of temporal stability rather than peak deviation magnitude. By using a conjugacy-based feature representation and examining longer temporal sequences, T-CER-Net focuses on persistent binocular misalignment while reducing the influence of short-lived, noise-related fluctuations. The attention-based temporal modeling further differentiates this approach from recurrent architectures by allowing information from distant time points to be considered directly within a sequence.

It should be noted that direct numerical comparisons across studies are inherently difficult because of differences in datasets, acquisition conditions, and evaluation protocols. Many prior works report results obtained under controlled settings or with subject-specific calibration, whereas the present study targets unconstrained recordings without calibration. Within this context, the balanced sensitivity and specificity achieved by T-CER-Net suggest that modeling long-range temporal consistency provides a meaningful advantage for real-world strabismus pre-screening.

6- Conclusion

This work presents the Temporal Cross-Eye Regression Network (T-CER-Net), an attention-based method for strabismus detection using video data. The method is intended for use in unconstrained recording conditions and focuses on improving robustness to noise while preserving diagnostic specificity, which remains a challenge for many existing automated screening methods.

A central contribution of this work is the introduction of the Cross-Eye Regression Error (CERE) as a compact temporal representation of binocular coordination. By measuring deviations from expected conjugate eye movement under normalized geometry, CERE shifts the problem away from absolute gaze estimation toward a relative measure that is less affected by head motion, camera distance, or lack of calibration. This formulation allows temporal patterns of ocular alignment to be analyzed in a stable and interpretable manner.

The results show that handling real-world variability during training is important. Oversampling normal recordings that include common artifacts, along with class weighting, helps the model learn that brief changes caused by blinks or tracking noise are usually part of normal eye movement. This design choice directly addresses the issue of low specificity that has limited the practical use of many automated screening systems.

In addition, the use of a Transformer-based temporal model enables the analysis of extended sequences without relying on sequential state propagation. By evaluating temporal consistency across the full sequence, the model is better able to distinguish sustained misalignment from short-lived noise. Compared with the CNN–LSTM baseline, this approach achieved higher specificity without a noticeable loss in sensitivity.

T-CER-Net showed balanced performance, with an AUC of 0.9140, a specificity of 0.8500, and a sensitivity of 0.8421. The results suggest that using noise-robust features together with attention-based temporal modeling can lead to more reliable automated prescreening. Further clinical studies are still needed, but the approach provides a reasonable starting point for video-based screening tools that could assist existing clinical workflows.

6-1- Limitations and Future Work

The results are encouraging, but several limitations remain. At this stage, T-CER-Net is designed to perform a simple binary decision between strabismus and normal ocular alignment. This is appropriate for prescreening, where the main goal is to flag potential cases, but it does not provide an estimate of deviation magnitude. In future work, the framework could be adapted to estimate continuous values, such as prism diopters, to better reflect routine clinical measurement.

This work mainly considered variability that commonly occurs in everyday recordings, including blinking, moderate head movement, and occasional landmark tracking errors, and addressed these effects through augmentation and oversampling. More extreme conditions, such as rapid head motion or heavy occlusion of the eyes, were not studied. Evaluating performance in these scenarios would help define the limits of the method.

A further limitation is the dependence on external landmark detection with MediaPipe Face Mesh. While the conjugacy-based CERE feature helps limit the impact of small tracking errors, failures in landmark detection can still affect performance. Exploring end-to-end models that learn both feature extraction and temporal classification directly from video data may help reduce this dependency.

Finally, the present work addresses a technical limitation—low specificity in unconstrained recordings—but does not replace the need for clinical validation. Evaluation in real screening environments, particularly with pediatric or uncooperative subjects, will be necessary before the method can be considered for wider deployment.

7- Declarations

7-1- Author Contributions

Conceptualization, W.K. and Y.H.; methodology, W.K.; software, M.E.E.A.; validation, K.T., P.K., and M.E.E.A.; formal analysis, W.K. and K.T.; investigation, W.K. and K.T.; resources, Y.H.; data curation, K.T. and P.K.; writing—original draft preparation, W.K.; writing—review and editing, W.K., K.T., M.E.E.A., Y.H., and P.K.; visualization, M.E.E.A.; supervision, W.K.; project administration, W.K.; funding acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

7-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7-3- Funding and Acknowledgments

This research was supported by funding from the Office of the National Digital Economy and Society Commission (ONDE), grant number 1006/66, and Walailak University's AioT Research Unit. The authors would like to express their gratitude to both institutions for their financial support, which made this study possible. This project was also conducted under the Reinventing University program, supported by the Ministry of Higher Education, Science, Research and Innovation.

7-4- Institutional Review Board Statement

Not applicable.

7-5- Informed Consent Statement

Not applicable.

7-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

8- References

[1] Hashemi, H., Rezvan, F., Pakzad, R., Ansaripour, A., Heydarian, S., Yekta, A., Ostadimoghaddam, H., Pakbin, M., & Khabazkhoob, M. (2022). Global and Regional Prevalence of Diabetic Retinopathy; A Comprehensive Systematic Review and Meta-analysis. Seminars in Ophthalmology, 37(3), 291–306. doi:10.1080/08820538.2021.1962920.

[2] Holmes, J. M., & Clarke, M. P. (2006). Amblyopia. The Lancet, 367(9519), 1343-1351. doi:10.1016/S0140-6736(06)68581-4.

[3] Holmes, J. M., Chandler, D. L., Christiansen, S. P., Birch, E. E., Bothun, E., Laby, D., Melia, B. M., Repka, M. X., Silbert, D. I., & Zeto, V. L. (2009). Interobserver reliability of the prism and alternate cover test in children with esotropia. Archives of Ophthalmology, 127(1), 59–65. doi:10.1001/archophthalmol.2008.548.

[4] Liu, J., Chi, J., Yang, H., & Yin, X. (2022). In the eye of the beholder: A survey of gaze tracking techniques. Pattern Recognition, 132, 108944. doi:10.1016/j.patcog.2022.108944.

[5] Guestrin, E. D., & Eizenman, M. (2006). General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Transactions on Biomedical Engineering, 53(6), 1124–1133. doi:10.1109/TBME.2005.863952.

[6] Tengtrisorn, S., Tungsattayathitthan, A., Na Phatthalung, S., Singha, P., Rattanalert, N., Bhurachokviwat, S., & Chouyjan, S. (2021). The reliability of the angle of deviation measurement from the Photo-Hirschberg tests and Krimsky tests. PLoS ONE, 16(12 December), 258744. doi:10.1371/journal.pone.0258744.

[7] Mestre, C., Gautier, J., & Pujol, J. (2018). Robust eye tracking based on multiple corneal reflections for clinical applications. Journal of Biomedical Optics, 23(03), 1. doi:10.1117/1.jbo.23.3.035001.

[8] Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. Proceedings of the IEEE International Conference on Computer Vision, 6911–6920. doi:10.1109/ICCV.2019.00701.

[9] Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., & Ji, Q. (2024). Automatic Gaze Analysis: A Survey of Deep Learning Based Approaches. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1), 61–84. doi:10.1109/TPAMI.2023.3321337.

[10] Simons, B. D., Siatkowski, R. M., Schiffman, J. C., Berry, B. E., & Flynn, J. T. (1999). Pediatric photoscreening for strabismus and refractive errors in a high- risk population. Ophthalmology, 106(6), 1073–1080. doi:10.1016/S0161-6420(99)90243-9.

[11] Huang, X., Lee, S. J., Kim, C. Z., & Choi, S. H. (2021). An automatic screening method for strabismus detection based on image processing. PLoS ONE, 16(8 August), 255643. doi:10.1371/journal.pone.0255643.

[12] Williams, T., Morgan, L. A., High, R., & Suh, D. W. (2018). Critical assessment of an ocular photoscreener. Journal of Pediatric Ophthalmology and Strabismus, 55(3), 194–199. doi:10.3928/01913913-20170703-18.

[13] Zhao, Z., Meng, H., Li, S., Wang, S., Wang, J., & Gao, S. (2025). High-Accuracy Intermittent Strabismus Screening via Wearable Eye-Tracking and AI-Enhanced Ocular Feature Analysis. Biosensors, 15(2), 110. doi:10.3390/bios15020110.

[14] Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video transformers: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(11), 12922-12943. doi:10.1109/TPAMI.2023.3243465.

[15] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195. doi:10.32604/cmc.2025.064799.

[16] Ali, Z., Bukhari, M., Javaid, M., Safdar, J., Kim, H., & Rho, S. (2025). Investigating vulnerabilities of gait recognition model using latent-based perturbations. Scientific Reports, 15(1), 39242. doi:10.1038/s41598-025-22869-4.

[17] Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., & Fröhlich, H. (2024). Transformer models in biomedicine. BMC Medical Informatics and Decision Making, 24(1), 214. doi:10.1186/s12911-024-02600-5.

[18] Tipirneni, S., & Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Transactions on Knowledge Discovery from Data (TKDD), 16(6), 1-17. doi:10.1145/3516367.

[19] Zheng, J., Ranjan, R., Chen, C. H., Chen, J. C., Castillo, C. D., & Chellappa, R. (2020). An automatic system for unconstrained video-based face recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(3), 194-209. doi:10.1109/TBIOM.2020.2973504.

[20] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172. doi:10.48550/arXiv.1906.08172.

[21] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open-source facial behavior analysis toolkit. In 2016 IEEE winter conference on applications of computer vision (WACV), 1-10. doi:10.1109/WACV.2016.7477553.

[22] Beck, R. W. (1998). The Pediatric Eye Disease Investigator Group. Journal of AAPOS: The Official Publication of the American Association for Pediatric Ophthalmology and Strabismus / American Association for Pediatric Ophthalmology and Strabismus, 2(5), 255–256. doi:10.1016/S1091-8531(98)90079-9.

[23] Ahuja, K., Islam, R., Barbhuiya, F. A., & Dey, K. (2017). Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recognition Letters, 91, 17–26. doi:10.1016/j.patrec.2017.04.002.

[24] Song, F., Tan, X., Chen, S., & Zhou, Z. H. (2013). A literature survey on robust and efficient eye localization in real-life scenarios. Pattern Recognition, 46(12), 3157–3173. doi:10.1016/j.patcog.2013.05.009.

[25] Morid, M. A., Sheng, O. R. L., & Dunbar, J. (2023). Time Series Prediction Using Deep Learning Methods in Healthcare. ACM Transactions on Management Information Systems, 14(1), 1–29. doi:10.1145/3531326.

[26] Farhad, M., Masud, M. M., Beg, A., Ahmad, A., & Ahmed, L. (2023). A Review of Medical Diagnostic Video Analysis Using Deep Learning Techniques. Applied Sciences (Switzerland), 13(11), 6582. doi:10.3390/app13116582.

[27] Ngo, T., & Manjunath, B. S. (2017). Saccade gaze prediction using a recurrent neural network. Proceedings - International Conference on Image Processing, ICIP, 2017-September, 3435–3439. doi:10.1109/ICIP.2017.8296920.

[28] Zheng, C., Li, W., Wang, S., Ye, H., Xu, K., Fang, W., Dong, Y., Wang, Z., & Qiao, T. (2024). Automated detection of steps in videos of strabismus surgery using deep learning. BMC Ophthalmology, 24(1), 242. doi:10.1186/s12886-024-03504-8.

[29] Han, C., Park, H., Kim, Y., & Gim, G. (2023). Hybrid CNN-LSTM Based Time Series Data Prediction Model Study. Studies in Computational Intelligence, 1075, 43–54. doi:10.1007/978-3-031-19608-9_4.

[30] Du, Q., Gu, W., Zhang, L., & Huang, S. L. (2018). Attention-based LSTM-CNNs for time-series classification. Proceedings of the 16th ACM conference on embedded networked sensor systems, 410-411. doi:10.1145/3274783.3275208.

[31] Thundiyil, S., & Picone, J. (2025). Time Series Analysis from Classical Methods to Transformer-Based Approaches: A Review. Signal Processing in Medicine and Biology: Applications of Deep Learning to the Health Sciences, 51–104. doi:10.1007/978-3-031-88024-7_2.

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5999–6009. doi:10.1201/9781003561460-19.

[33] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 12B(12), 11106–11115. doi:10.1609/aaai.v35i12.17325.

[34] Oliveira, J. M., & Ramos, P. (2024). Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. Mathematics, 12(17), 2728. doi:10.3390/math12172728.

[35] Hu, Y., & Xiao, F. (2022). Network self-attention for forecasting time series. Applied Soft Computing, 124, 109092. doi:10.1016/j.asoc.2022.109092.

[36] Martínez-Martínez, J., Brown, O., Karami, M., & Nabavi, S. (2025). Robust Training with Data Augmentation for Medical Imaging Classification. arXiv preprint arXiv:2506.17133. doi:10.48550/arXiv.2506.17133.

[37] Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. European Radiology, 34(12), 7895–7903. doi:10.1007/s00330-024-10834-0.

[38] Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June, 9260–9269. doi:10.1109/CVPR.2019.00949.

[39] Sugano, Y., Matsushita, Y., & Sato, Y. (2014). Learning-by-synthesis for appearance-based 3D gaze estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1821–1828. doi:10.1109/CVPR.2014.235.