This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



# Predicting Longitudinal Visual Field Progression with Class Imbalanced Data

Ling Chen, *Member, IEEE,* Chun-Hung Chen, Wei Wang, Da-Wen Lu, and Vincent S. Tseng*, *Fellow, IEEE*

***Abstract*—Glaucoma is the leading cause of irreversible blindness worldwide. The clinical standard for glaucoma diagnosis and progression tracking remains visual field (VF) testing via standard automated perimetry. One outstanding challenge of many ophthalmic prediction tasks is the issue of class imbalance, where the majority class outnumbers the minority class(es). Although this issue has been reported in several prior studies on the prediction of VF progression or glaucoma, it has not been addressed in the context of longitudinal VF data. In this work, we proposed, VF-Transformer, a transformer-based framework for VF progression prediction based on longitudinal VF examination results. In particular, we addressed the class imbalance issue by incorporating our proposed *inverted class-dependent temperature* (ICDT) loss and weight normalization.**

**The proposed framework was developed and evaluated on a public VF dataset and further validated on an external hospital dataset, using accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) as evaluation metrics. Extensive experiments and comparisons with existing state-of-the-art methods and class imbalance handling strategies confirmed the effectiveness of the proposed framework in predicting VF progression in the presence of class imbalance.**

***Index Terms*—visual field progression prediction, imbalanced data, transformer-based network**

## I. Introduction

GLAUCOMA is a heterogeneous group of progressive optic neuropathies that causes damage to retinal ganglion cells, their axons and associated glial cells. It is the leading cause of irreversible blindness worldwide [1]. Visual field (VF) testing via standard automated perimetry remains a clinical standard for glaucoma diagnosis and progression tracking [2]. An example of Humphrey field analyzer (HFA) 24-2 VF sensitivity values for a right eye is shown in Fig. 1, where each VF result contains 52 test points (excluding 2 blind spots), and each point is expressed on a dB scale of light attenuation.

Detecting VF worsening is important for tracking and managing glaucoma progression. However, manually classifying glaucomatous VF loss over time can be challenging due to subjectivity in VF reviewing and variability of VF measures across tests. Clinical indices have been designed to summarize VF results, including global measures such as mean deviation (MD) and visual field index (VFI), and local measures such as pointwise threshold values [3]. Trend-based analyses are commonly applied to determine the rates of change in these measures using linear regression, rendering the MD slope, VFI slope, pointwise linear regression (PLR), and permutation of pointwise linear regression (PoPLR). Event-based algorithms, such as the advanced glaucoma intervention study (AGIS) and collaborative initial glaucoma treatment study (CIGTS) scoring systems, are also used to determine the progression. AGIS scores each VF according to the defects in the nasal, superior, and inferior hemifields, while CIGTS scores VF points based on the density and depth of defects [4].

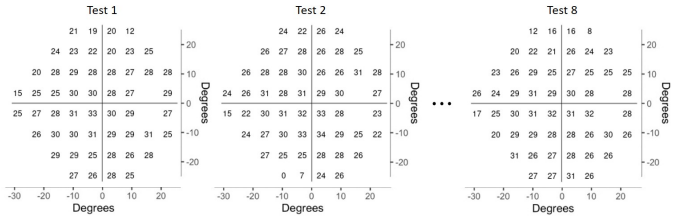


Fig. 1: Example of the sensitivity values of longitudinal visual field tests of a patient's right eye, expressed on a dB scale of light attenuation.

Machine learning (ML)-based or deep learning (DL)-based algorithms have been developed to detect VF progression within given longitudinal VFs [5]–[7] or to forecast future VF results [8]–[11]. Although several studies [6], [12], [13] using VF data have reported observing class-imbalanced data, the issue has not been specifically addressed for longitudinal VFs. However, longitudinal relationships amongst consecutive VF test results are crucial for tracking glaucoma progression [14], which warrants careful study. The class imbalance issue is characterized by the majority class outnumbering the minority class(es), which generally causes an undesirable bias towards the majority class and limits the model performance [15]. It is commonly observed in medical data, where the number of cases is much smaller than the controls [12], [16].

In this study, we introduce VF-Transformer, a DL framework for predicting VF progression based on a transformer architecture [17] with a novel loss function to handle class imbalance. The main contributions of this work include:

This research was supported by the National Science and Technology Council, Taiwan, R.O.C. under Grant number NSTC 113-2221-E-A49-028-.

Ling Chen and Wei Wang are with the Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan (e-mails: {ling.chen, ww.md11}@nycu.edu.tw).

Da-Wen Lu is with the Department of Ophthalmology, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan (e-mail: ludawen@yahoo.com).

Chun-Hung Chen is with the Department of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan (e-mail: jimchen1551.y@nycu.edu.tw).

Vincent S. Tseng is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan (e-mail: vtseng@cs.nycu.edu.tw).

* corresponding author

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



- We repropose VF-Transformer, a fully automated transformer-based framework for VF progression prediction in the presence of class imbalance. To our knowledge, this is the first study to tackle the issue of class imbalance in longitudinal VF data.
- We introduce a loss function called *inverted class-dependent temperature* (ICDT) loss in combination with $\tau$-norm weight normalization to handle the class imbalance issue commonly observed in ophthalmic data.
- The proposed framework effectively detects and predicts VF progression despite class imbalance, achieving significantly improved sensitivity with minimal impact on AUC and specificity across internal and external test sets. This is highly important for practical clinical applications, especially in the presence of class imbalance, which often results in severely diminished sensitivity.

Section II surveys the related work in the literature. Section III formally defines the learning problem. The proposed VF-Transformer framework for VF progression prediction is introduced in Section IV. Section V reports the experimental settings and results, followed by the discussions in Section VI. Finally, Section VII concludes the paper.

## II. Related Work

### A. Learning with Longitudinal Visual Field Data

Prior ML-based or DL-based studies on longitudinal VF data focused mainly on automatic VF results estimation [8]–[11], VF progression detection [5]–[7], or defect pattern categorization [13], [18]. For VF results estimation, regression-based models [9], recurrent neural network (RNN) [10], and Gaussian process-based models [8] were introduced to capture temporal relationships amongst VFs and estimate their values. For defect pattern categorization, VF results were converted into an image shape [18] or Voronoi images [13] for CNN models to classify VF defects. However, these studies did not consider the longitudinal aspect of the VF data.

For progression analysis, standard ML methods [5] and elastic-net Cox regression models [7] were used in prior studies. Dixit et al [6] developed a convolutional long short-term memory (LSTM) network to detect glaucoma progression. Although the class imbalance issue was reported in [6], it was not addressed. Herbert et al [19] trained 3D CNNs to forecast VF worsening using 1-3 VFs with OCT features; however, the longitudinality of VF data was not handled.

### B. Learning with Class Imbalance Data

Class imbalance issue occurs when some classes have many more instances than others. It has been referred to as the long-tailed learning problem in recent years by the DL community [15], when the distribution of class sizes shows a long-tailed shape. Mainstream class rebalancing approaches include re-sampling and class-sensitive learning:

- **Re-sampling**. Class imbalance issue was traditionally handled at the sample level, using random over-sampling or random under-sampling methods. A more recent classic over-sampling method, SMOTE [20], generates tail-class samples by mixing intra-class neighboring samples, which was used to handle class imbalance for early detection of glaucoma [12]. Other studies designed special over-sampling methods, for example, based on clustering and graph representation [21]. However, these methods do not consider the longitudinal aspect of the data.
- **Class-sensitive learning** addresses the problem of uneven gradients for different classes caused by class imbalance at the objective level [22]–[25]. Focal loss [22] assigns higher weights to the harder tail classes and lower weights to the easier head classes. Balanced softmax (BS) loss [23] reweights the exponential of predicted logits by label frequencies to alleviate the bias caused by class imbalance. It was adopted to improve Vision Transformer performance with imbalanced image data [25]. Class-dependent temperature (CDT) loss [24] rebalances per-class logits by a factor known as "temperature" [26].

Other strategies were also introduced. For example, class-center triplet loss was designed to enable a more compact distribution for each class [27].

In this study, we focus on class-sensitive learning approaches. Our proposed ICDT loss is close to CDT loss [24], as both approaches address class imbalance by adjusting the class temperature. However, CDT loss fails to handle imbalanced classes specific to the VF data, as will be shown in the later sections.

## III. Problem Definition and Preliminaries

### A. Problem Definition

Given a dataset $\mathcal{I} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$ of $N$ eyes, the consecutive VF test results of an eye $i$ is represented as a matrix $\mathbf{X}_i \in \mathbb{R}^{M\times T}$, where $M$ is the number of VF tests and $T$ is the number of test points in a test. $y_i \in \{0, 1\}$ is the ground truth progression label for eye $i$.

Let the output logit $z_c$ of class $c$ of a deep classifier be:

$$z_c = \mathbf{w}_c^\top f_\theta(\mathbf{X}) \tag{1}$$

where $\mathbf{X}$ is the input, $f_\theta(\cdot)$ is the feature extractor of a deep network, parameterized by $\theta$, and $\mathbf{w}_c$ is the linear classifier of class $c$. The VF progression prediction problem can be defined as a binary classification problem, learning a deep classifier:

$$\hat{y} = argmax_{c\in\{0,1\}} z_c \tag{2}$$

where $z_c$ is defined as Eq. (1).

For the detection task, progression is defined based on changes within the input $M$ consecutive VFs. For the forecasting task, progression is defined based on changes across $K$ consecutive VFs, given prior $M$ VFs, where $K > M$.

### B. Preliminaries

The predicted probability $p$ of the positive class ($c = 1$) for binary classification is:

$$p = \sigma(z) = \frac{1}{1 + e^{-z_{c=1}}} \tag{3}$$

where $\sigma(\cdot)$ is a sigmoid function and $z_c$ is defined as Eq. (1). Then the standard binary cross entropy loss can be defined as:

$$Loss_{BCE} = -y \cdot log(p) - (1 - y) \cdot log(1 - p) \tag{4}$$

where $y$ is the ground truth label and $p$ is defined as Eq. (3).

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



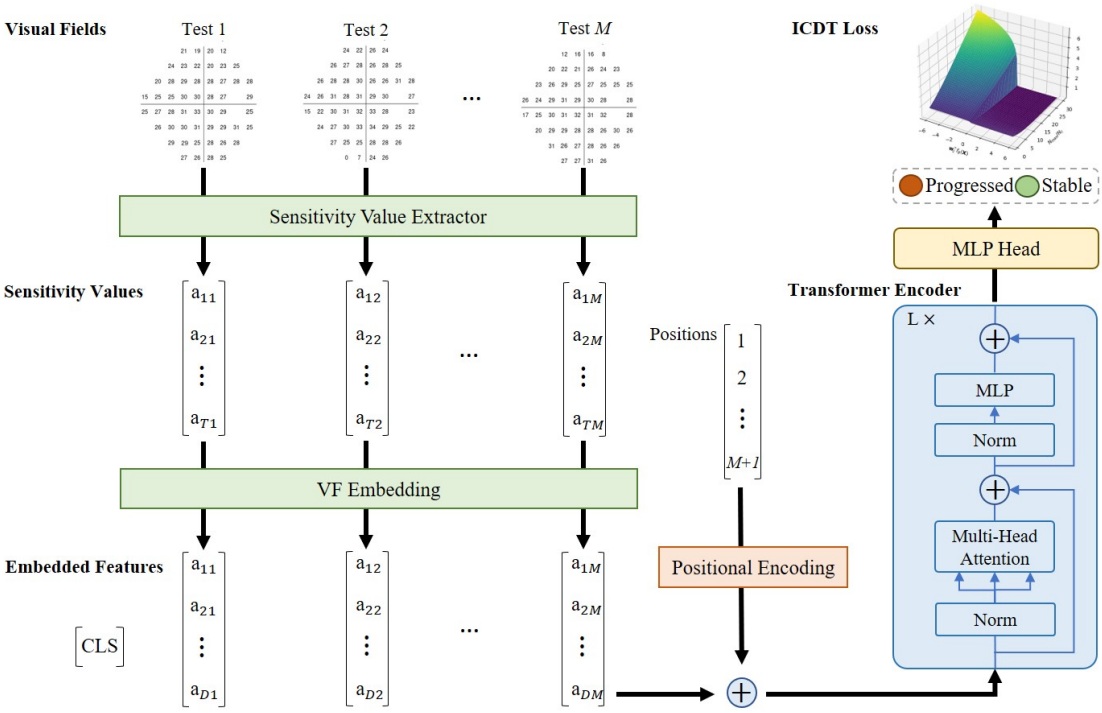


Fig. 2: An overview of the proposed VF-Transformer framework

## IV. The Proposed Framework

The proposed VF-Transformer framework contains three modules, including a VF embedding extraction module, a transformer encoder block, and a class imbalance handling module. An overview of the framework is shown in Fig. 2

### A. VF Embedding Extraction

To extract features from input $\mathbf{X}_i$ of VF sensitivity values, a VF embedding layer was constructed as a trainable linear projection $\mathbf{E} \in \mathbb{R}^{T \times D}$, where $T$ is the number of test points and $D$ is the embedded dimension.

### B. Transformer Block

To prepare input for the transformer block, we first encode the positions of $M$ individual VF tests following [17] and add the encoded positions into the embedded features with an additional class position as in [28]. The output is then passed to a standard transformer encoder with $D$ embedding dimensions, $h$ heads for multi-head self-attention, and $L$ encoder layers. Finally, a multilayer perceptron (MLP) head is applied to the transformer encoder for binary classification.

### C. ICDT Loss Function

Our proposed ICDT loss, similar to the design of the CDT [24] loss, reweights per-class logits by adding class temperature $a_c$ in Eq. (1):

$$z'_c = \mathbf{w}_c^\top f_\theta(\mathbf{X})/a_c \tag{5}$$

where $\mathbf{X}$, $f_\theta(\cdot)$ and $\mathbf{w}_c$ are defined as in Eq. (1).

In [24], the decision values (logits) of the test features of a minority class were observed to move away from the training features to low decision values. Therefore, $a_c$ was designed as a factor of imbalance ratio $IR$ to artificially simulate the effect of feature deviation in training:

$$IR = N_{max}/N_c \tag{6}$$

where $N_{max}$ and $N_c$ are sizes of the majority class and a given class $c$, respectively. Then $a_c$ is defined as:

$$a_c = (IR)^\gamma \tag{7}$$

where $\gamma$ controls the strength of the temperature.

A later study [29] reported based on the CIFAR-10 image dataset that CDT loss can harm the performance of the minority class in binary classification in the initial phase of training, by guiding the classifier in the wrong direction. However, empirically we found that the performance of the minority class did not recover well when training with longitudinal VF data, as will be demonstrated in Section V-E1.

Inspired by the observation of [30] that training with scaled logits produces smaller gradients for correctly classified samples and larger gradients for incorrectly classified samples, we defined $a'_c$ to be the inverse of the $IR$ ratio to scale up the logits of the minority class:

$$a'_c = (IR)^{-\gamma} = (N_c/N_{max})^\gamma \tag{8}$$

where $N_{max}$, $N_c$, and $\gamma$ are defined as in Eqs. (6) and (7).

Finally, our proposed ICDT loss for binary classification can be derived using Eqs. (3), (4), (5), and (8):

$$\begin{aligned} Loss_{ICDT} = &-y \cdot log(\frac{1}{1+e^{-\mathbf{w}_{c=1}^\top f_\theta(\mathbf{X})(N_{max}/N_{c=1})^\gamma}}) \\ &-(1-y) \cdot log(1-\frac{1}{1+e^{-\mathbf{w}_{c=1}^\top f_\theta(\mathbf{X})(N_{max}/N_{c=1})^\gamma}}) \end{aligned} \tag{9}$$

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



TABLE I: UWHVF dataset: sizes and imbalance ratios (IR) for various lengths (M) of VFs.

| | | | MD Slope | | VFI Slope | | PLR | | PoPLR | | AGIS | | CIGTS | | M6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | Patients | Eyes | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR |
| 5 | 1103 | 2065 | 323 | 5.39 | 199 | 9.38 | 465 | 3.44 | 222 | 8.30 | 63 | 31.78 | 78 | 25.47 | 111 | 16.71 |
| 6 | 784 | 1436 | 289 | 3.97 | 228 | 5.30 | 334 | 3.30 | 201 | 6.14 | 59 | 23.34 | 98 | 13.65 | 124 | 9.81 |
| 7 | 538 | 975 | 267 | 2.65 | 220 | 3.43 | 247 | 2.95 | 194 | 4.03 | 62 | 14.73 | 86 | 10.34 | 121 | 6.48 |

TABLE II: TSGH external test set: sizes and imbalance ratios (IR) for various lengths (M) of VFs.

| | | | MD Slope | | VFI Slope | | PLR | | PoPLR | | AGIS | | CIGTS | | M6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | Patients | Eyes | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR | No. Pos. | IR |
| 5 | 625 | 1129 | 231 | 3.89 | 139 | 7.12 | 272 | 3.15 | 143 | 6.90 | 42 | 25.88 | 47 | 23.02 | 69 | 14.41 |
| 6 | 589 | 630 | 176 | 2.58 | 126 | 4.00 | 147 | 3.29 | 117 | 4.38 | 25 | 24.20 | 38 | 15.58 | 64 | 8.08 |
| 7 | 195 | 350 | 116 | 2.02 | 82 | 3.27 | 76 | 3.61 | 87 | 3.02 | 17 | 19.59 | 29 | 11.07 | 43 | 6.33 |

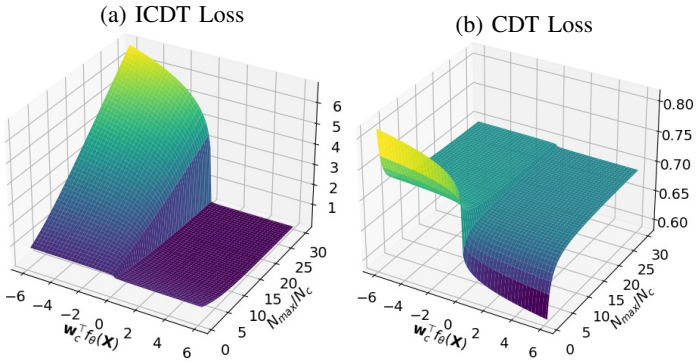


Fig. 3: Comparing loss value distributions of ICDT loss and CDT loss with $\gamma$ =1 and $\tau$ =0.8 at $c$ =1 and $y$ =1.

where higher $N_{max}/N_{c=1}$ renders more scaled up per-class logits for the minority class ($c$ =1), resulting in larger gradients for the misclassified minority class samples.

To further understand the behavior of the proposed ICDT loss in comparison with CDT loss, we plotted Fig. 3 using $\mathbf{w}_{c=1}^\top f(\theta)$ values in [-6, 6] inclusive and IR ($N_{max}/Nc$) values in [1, 30] inclusive, at $c$ =1 and $y$ =1 (minority class as positive) with $\gamma$ =1 and $\tau$ =0.8. ICDT loss demonstrated superior efficacy in adjusting logits dynamically based on IR, which results in higher gradients for the minority class when misclassified (i.e., logits <0) as IR increases (Fig. 3a). This leads to faster convergence rates and encourages corrective feedback during training. By contrast, CDT loss only adjusts the logits more effectively when IR is smaller (Fig. 3a), which can limit its performance when the imbalance issue is severe.

### D. Weight Normalization

An additional $\tau$-normalization [31] was applied to the classifier weights, namely the weights of the last fully connected layers of the MLP head, to enhance the training stability:

$$\mathbf{w}_c' == \frac{\mathbf{w}_c}{||\mathbf{w}_c||^\tau} \tag{10}$$

where $\mathbf{w}_c \in \mathbb{R}^d$ denotes the classifier weights corresponding to class $c$, $||\cdot||$ denotes $L_2$-normalization. Empirically, $\tau$ is often chosen as $\tau \in (0, 1)$. When $\tau$ =0, no scaling is imposed, while when $\tau$ =1, it reduces to the standard $L_2$-normalization.

## V. Evaluation Methods and Results

### A. Datasets

In this study, we used a public dataset for model development and testing, and an additional hospital dataset for external testing. The public UWHVF [32] from the Department of Ophthalmology at the University of Washington contains 28,943 VFs from 7,248 for eyes of 3,871 patients who underwent VF tests with or without glaucoma diagnosis. It was collected between 1998 and 2018 using a Humphrey field analyzer (HFA) II. The hospital dataset was collected from glaucoma patients and suspects visiting the Tri-Service General Hospital (TSGH), Taiwan, between 2013 and 2022 using HFA III. Patients with 5 or more VFs were included, while VFs with false positive rate $\geq$0.15 and false negative rate $\geq$0.25 were excluded to ensure data reliability. The current study was reviewed and approved by the International Review Board (IRB) of TSGH and the need for informed consent was waived.

The VF data were stored in a right-eye layout, with the left-eye VFs flipped to the same layout as the right-eye. Sensitivity values of 52 VF test points from the 24-2 pattern were used, excluding two blind spots. UWHVF dataset was randomly divided at the patient level into a development set (80%) for training and validation, and a held-out test set (20%) for testing. The entire TSGH set was used for external validation.

The ground truth labels for VF progression were defined by commonly used trend-based algorithms (MD slope, VFI slope, PLR, and PoPLR) and event-based algorithms (AGIS and CIGTS). As in [5], we included a combined progression algorithm called M6, where the "progressed" or "stable" label of an eye is determined by the majority of the six algorithms. If there is a tie, the eye is labeled as "unclear", and will not be used in model training and testing. The Python wrapper of the R *vfprogression* and *visualField* packages were used to compute these reference standards [33].

We divided the data into subsets of a range of input VF lengths $M$. TABLE I and TABLE II show the numbers of patients and eyes in these subsets for the UWHVF and TSGH datasets respectively, with corresponding positive case size (No. Pos.) and IR defined in Eq. (6) for seven progression algorithms.

### B. Comparative Methods

To evaluate our proposed approach, we compared the following methods relevant to our approach for handling longitu-

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



dinal input, with different combinations of losses and weight normalization settings described shortly below.

- **LSTM** [34] is the standard long short-term memory (LSTM) network. For fair comparison with our approach, the extracted VF embedded features described in Section IV-A was used as inputs.
- **ConvLSTM** [6] is a DL network to predict VF progression based on two layers of convolutional LSTM that treats each input VF as a 2D image-like matrix.

To examine the effectiveness of our proposed ICDT loss against other high-performing losses for handling class imbalance, the following losses were compared:

- **BCE Loss**. Binary cross entropy (BCE) loss is a standard loss for binary classification, as defined in Eq. (4). It is used as a baseline without considering class imbalance.
- **BS Loss** [35]. Balanced softmax (BS) loss handles class imbalance by reweighting the exponential terms in the original softmax function by label frequencies of classes.
- **FL** [22]. Focal loss (FL) addresses the class imbalance by adding a modulating factor to the cross entropy loss to prevent gradients from being dominated by easily classified negatives. Different from the BS loss, FL loss does not use label frequencies.
- **CDT Loss** [24]. Class-dependent temperature (CDT) loss imposes class-dependent weights on model output logits via per-class temperature using the class imbalance ratio.

In addition, we compared the use of the $\tau$ normalization [31] with no weight normalization.

### C. Evaluation Metrics

We employed accuracy (ACC), sensitivity (SEN), specificity (SPE), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC) to evaluate the effectiveness of our proposed framework. We use TP, FP, FN, and TN to denote true positive, false positive, false negative, and true negative. Accuracy, sensitivity, and specificity are defined as follows.

$$ACC = (TP + TN)/(TP + FP + FN + TN)$$

$$SEN = TP/(TP + FN)$$

$$SPE = TN/(TN + FP)$$

### D. Experimental Settings

The experiments were conducted using Python and the PyTorch library running on an RTX Titan GPU. All models underwent training for 200 epochs using a batch size of 32. We employed the Adam optimizer with a weight decay of 0.00001. The initial learning rate was set at 0.0001 and was adjusted using a cosine annealing scheduler.

For VF-Transformer, the number of test points $T$ =52, the size of the embedding dimension $D$ =768, the number of heads $h$ for multi-head self-attention was 8, and the number of encoder layers $L$ =6. For fair comparison, the LSTM shared the same input VF embedded features as VF-Transformer. The original VFs were in the shape of 8×9, which were padded to 9×9 as input for ConvLSTM. For BCE loss, we used PyTorch BCE loss with logits implementation that utilizes the log-sum-exp trick to enhance numerical stability. The meta-parameter $\gamma$ for CDT loss and our ICDT loss were set to 1. $\tau$ was experimentally set to 0.8.

### E. Results

*1) Comparing Class Imbalance Strategies:* We compared the performance of VF-Transformer with the models listed in Section V-B, using different combinations of losses and weight normalization settings. TABLE III shows the results based on 5 VFs in detecting M6 progression on the internal test set (UWHVF), with the mean ± standard deviation of predicted logits for positive and negative samples of the training and test sets, while TABLE IV presents the results on the entire external test set (TSGH).

Our proposed VF-Transformer with ICDT loss and $\tau$-normalization outperformed all the other models, losses, and normalization settings compared, achieving an AUC of 0.921, sensitivity of 0.857, and specificity of 0.937 on the internal test set (TABLE III), and an AUC of 0.913, sensitivity of 0.899, and specificity of 0.915 on the external test set (TABLE IV). While other approaches excelled in individual metrics such as accuracy, AUC, or specificity, they did so at the cost of significantly compromised sensitivity. For instance, models trained with BS loss exhibited a bias toward predicting either exclusively positive cases (TABLE III) or exclusively negative cases (TABLE IV). Meanwhile, models trained with other loss functions generally prioritized specificity due to class imbalance.

The proposed ICDT loss played a crucial role in addressing class imbalance, significantly improving sensitivity across all models (by 0.095–0.238 on the internal test set, TABLE III, and 0.203–0.261 on the external test set, TABLE IV) — a critical factor in clinical practice, where higher sensitivity is prioritized. Additionally, incorporating $\tau$-normalization further enhanced the VF-Transformer's performance.

The transformer model, on the other hand, seems to be more robust in the presence of imbalanced data, since its combination with individual losses generally outperformed their combinations with LSTM or ConvLSTM. This could be attributed to the attention mechanism of transformer to better extract representative features for the minority class.

*2) Analysis of the predicted logits:* In [24], the predicted logits of the test features of a minority class were observed to move away from the training features to low decision values, a tendency observed, but to a lesser degree, in our predicted logits. As TABLE III shows, our proposed ICDT loss demonstrated a superior ability to push up logits of positive samples regardless of the models used. Furthermore, transformer seems to be able to push up these logits regardless of the loss used.

*3) VF Progression Detection:* TABLE V shows the performance of VF-Transformer for progression detection given 5 VFs on the internal (UWHVF) and external (TSGH) test sets. Overall, the proposed framework demonstrated AUCs of 0.835-0.975 on UWHVF and 0.813-0.932 on TSGH, with 0.721-0.889 sensitivity and 0.795-0.961 specificity on

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



TABLE III: Comparing model performance and output logits for detecting M6 progression given 5 VFs on UWHVF test set.

| Model | Loss and Norm | Metrics | | | | Train Logits | | Test Logits | |
|---|---|---|---|---|---|---|---|---|---|
| | | ACC | AUC | SEN | SPE | Positive | Negative | Positive | Negative |
| LSTM | BCE | 0.9557 | 0.9515 | 0.6190 | 0.9752 | 0.66 ± 0.29 | 0.02 ± 0.09 | 0.59 ± 0.36 | 0.03 ± 0.13 |
| | BCE + $\tau$-norm | 0.9531 | 0.9488 | 0.5238 | 0.978 | 0.63 ± 0.30 | 0.02 ± 0.09 | 0.57 ± 0.37 | 0.03 ± 0.12 |
| | BS | 0.0544 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.0544 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9609 | 0.9477 | 0.4286 | 0.9917 | 0.49 ± 0.21 | 0.07 ± 0.08 | 0.44 ± 0.22 | 0.08 ± 0.09 |
| | FL + $\tau$-norm | 0.9662 | 0.9477 | 0.4286 | 0.9972 | 0.48 ± 0.21 | 0.09 ± 0.08 | 0.44 ± 0.22 | 0.10 ± 0.09 |
| | CDT | 0.9559 | 0.9498 | 0.2857 | 0.9945 | 0.31 ± 0.42 | 0.00 ± 0.01 | 0.33 ± 0.39 | 0.01 ± 0.07 |
| | CDT + $\tau$-norm | 0.9559 | 0.9412 | 0.2381 | 0.9972 | 0.25 ± 0.36 | 0.00 ± 0.01 | 0.24 ± 0.31 | 0.00 ± 0.04 |
| | ICDT | 0.9507 | 0.9461 | 0.8095 | 0.9587 | 0.76 ± 0.18 | 0.08 ± 0.16 | 0.70 ± 0.30 | 0.09 ± 0.17 |
| | ICDT + $\tau$-norm | 0.9507 | 0.9528 | 0.8095 | 0.9587 | 0.76 ± 0.18 | 0.08 ± 0.16 | 0.72 ± 0.29 | 0.09 ± 0.17 |
| ConvLSTM | BCE | 0.9633 | 0.9751 | 0.5238 | 0.989 | 0.51 ± 0.26 | 0.04 ± 0.09 | 0.51 ± 0.22 | 0.05 ± 0.10 |
| | BCE + $\tau$-norm | 0.9611 | **0.9781** | 0.4286 | 0.9917 | 0.48 ± 0.27 | 0.04 ± 0.08 | 0.48 ± 0.23 | 0.04 ± 0.09 |
| | BS | 0.0546 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.0546 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9532 | 0.9782 | 0.1429 | **1.00** | 0.38 ± 0.13 | 0.12 ± 0.07 | 0.39 ± 0.11 | 0.13 ± 0.07 |
| | FL + $\tau$-norm | 0.9532 | 0.9764 | 0.1429 | **1.00** | 0.39 ± 0.14 | 0.13 ± 0.07 | 0.38 ± 0.11 | 0.14 ± 0.07 |
| | CDT | 0.9454 | 0.9704 | 0.00 | **1.00** | 0.05 ± 0.06 | 0.00 ± 0.01 | 0.04 ± 0.03 | 0.00 ± 0.01 |
| | CDT + $\tau$-norm | 0.9454 | 0.9212 | 0.00 | **1.00** | 0.06 ± 0.07 | 0.01 ± 0.01 | 0.05 ± 0.05 | 0.01 ± 0.01 |
| | ICDT | 0.9378 | 0.9706 | 0.7619 | 0.9477 | 0.61 ± 0.14 | 0.18 ± 0.14 | 0.62 ± 0.13 | 0.20 ± 0.15 |
| | ICDT + $\tau$-norm | 0.9506 | 0.9716 | 0.7619 | 0.9614 | 0.60 ± 0.15 | 0.17 ± 0.13 | 0.61 ± 0.15 | 0.19 ± 0.14 |
| VF-Transformer | BCE | 0.9635 | 0.8936 | 0.7619 | 0.9752 | 0.99 ± 0.02 | 0.00 ± 0.00 | 0.75 ± 0.41 | 0.03 ± 0.15 |
| | BCE + $\tau$-norm | 0.9613 | 0.9333 | 0.619 | 0.9807 | 1.00 ± 0.01 | 0.00 ± 0.00 | 0.61 ± 0.47 | 0.02 ± 0.13 |
| | BS | 0.0544 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.0544 | 0.50 | **1.00** | 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9665 | 0.9329 | 0.619 | 0.9862 | 0.94 ± 0.10 | 0.01 ± 0.02 | 0.63 ± 0.39 | 0.03 ± 0.11 |
| | FL + $\tau$-norm | 0.9715 | 0.9233 | 0.6667 | 0.989 | 0.96 ± 0.10 | 0.01 ± 0.03 | 0.65 ± 0.39 | 0.03 ± 0.10 |
| | CDT | **0.9716** | 0.9708 | 0.6667 | 0.989 | 0.71 ± 0.44 | 0.00 ± 0.01 | 0.66 ± 0.46 | 0.01 ± 0.11 |
| | CDT + $\tau$-norm | 0.969 | 0.9689 | 0.5714 | 0.9917 | 0.63 ± 0.46 | 0.00 ± 0.01 | 0.57 ± 0.48 | 0.01 ± 0.09 |
| | ICDT | 0.7154 | 0.8544 | 0.8571 | 0.708 | 0.53 ± 0.01 | 0.19 ± 0.24 | 0.48 ± 0.15 | 0.21 ± 0.25 |
| | ICDT + $\tau$-norm | 0.9323 | 0.9213 | 0.8571 | 0.9366 | 0.65 ± 0.11 | 0.08 ± 0.17 | 0.59 ± 0.23 | 0.10 ± 0.18 |

*Average logits for positive and negative samples from the train and test sets are presented as mean ± standard deviation.

TABLE IV: Comparing model performance and output logits for detecting M6 progression given 5 VFs on TSGH test set.

| Model | Loss and Norm | Metrics | | | | Test Logits | |
|---|---|---|---|---|---|---|---|
| | | ACC | AUC | SEN | SPE | Positive | Negative |
| LSTM | BCE | 0.9518 | 0.9533 | 0.4928 | 0.9838 | 0.52 ± 0.38 | 0.03 ± 0.11 |
| | BCE + $\tau$-norm | 0.9500 | 0.9500 | 0.4348 | 0.9858 | 0.51 ± 0.38 | 0.03 ± 0.10 |
| | BS | 0.9347 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.9347 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9527 | 0.9486 | 0.3333 | 0.9959 | 0.45 ± 0.26 | 0.07 ± 0.09 |
| | FL + $\tau$-norm | 0.9527 | 0.9484 | 0.3333 | 0.9959 | 0.44 ± 0.26 | 0.08 ± 0.09 |
| | CDT | 0.9517 | 0.9315 | 0.2754 | 0.999 | 0.29 ± 0.41 | 0.00 ± 0.03 |
| | CDT + $\tau$-norm | 0.9497 | 0.9455 | 0.2319 | **1.00** | 0.24 ± 0.37 | 0.0 ± 0.02 |
| | ICDT | 0.9491 | 0.9560 | 0.7536 | 0.9625 | 0.69 ± 0.28 | 0.08 ± 0.16 |
| | ICDT + $\tau$-norm | 0.9472 | 0.9554 | 0.7246 | 0.9625 | 0.69 ± 0.29 | 0.08 ± 0.16 |
| ConvLSTM | BCE | 0.9525 | 0.9476 | 0.4203 | 0.9899 | 0.44 ± 0.30 | 0.05 ± 0.09 |
| | BCE + $\tau$-norm | 0.9515 | 0.9407 | 0.3913 | 0.9909 | 0.42 ± 0.31 | 0.05 ± 0.09 |
| | BS | 0.9344 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.9344 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9478 | 0.9404 | 0.2174 | 0.999 | 0.37 ± 0.15 | 0.13 ± 0.07 |
| | FL + $\tau$-norm | 0.9478 | 0.9172 | 0.2174 | 0.999 | 0.37 ± 0.17 | 0.14 ± 0.07 |
| | CDT | 0.9344 | 0.9485 | 0.00 | **1.00** | 0.05 ± 0.07 | 0.00 ± 0.01 |
| | CDT + $\tau$-norm | 0.9344 | 0.9239 | 0.00 | **1.00** | 0.07 ± 0.08 | 0.01 ± 0.01 |
| | ICDT | 0.9367 | 0.9532 | 0.6232 | 0.9585 | 0.59 ± 0.18 | 0.18 ± 0.15 |
| | ICDT + $\tau$-norm | 0.9394 | 0.9496 | 0.5362 | 0.9676 | 0.56 ± 0.20 | 0.16 ± 0.13 |
| VF-Transformer | BCE | 0.9565 | 0.9531 | 0.6522 | 0.9777 | 0.63 ± 0.46 | 0.02 ± 0.15 |
| | BCE + $\tau$-norm | **0.9633** | 0.9684 | 0.6812 | 0.9828 | 0.68 ± 0.45 | 0.02 ± 0.12 |
| | BS | 0.9347 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | BS + $\tau$-norm | 0.9347 | 0.50 | 0.00 | **1.00** | 0.50 ± 0.00 | 0.50 ± 0.00 |
| | FL | 0.9565 | 0.9534 | 0.6087 | 0.9807 | 0.62 ± 0.41 | 0.02 ± 0.13 |
| | FL + $\tau$-norm | 0.9593 | 0.9515 | 0.5652 | 0.9868 | 0.61 ± 0.43 | 0.02 ± 0.11 |
| | CDT | 0.9613 | 0.9649 | 0.5652 | 0.9889 | 0.57 ± 0.48 | 0.01 ± 0.10 |
| | CDT + $\tau$-norm | 0.9612 | **0.9676** | 0.5072 | 0.9929 | 0.52 ± 0.46 | 0.01 ± 0.08 |
| | ICDT | 0.9233 | 0.9618 | 0.8261 | 0.9301 | 0.56 ± 0.21 | 0.05 ± 0.15 |
| | ICDT + $\tau$-norm | 0.9137 | 0.9131 | **0.8986** | 0.9149 | 0.54 ± 0.18 | 0.07 ± 0.17 |

*Average logits for positive and negative samples from the train and test sets are presented as mean ± standard deviation.

UWHVF, and 0.659-0.899 sensitivity and 0.764-0.938 specificity on TSGH.

In M6 progression detection, the framework achieved 0.921 AUC, 0.857 sensitivity and 0.937 specificity on UWHVF, with

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



TABLE V: Performance for progression detection on the internal and external test sets given 5 VFs ($M$=5) as input.

| | Internal (UWHVF) Test Set | | | | External (TSGH) Test Set | | | |
|---|---|---|---|---|---|---|---|---|
| | ACC | AUC | SEN | SPE | ACC | AUC | SEN | SPE |
| M6 | 0.9323 | 0.9213 | 0.8571 | 0.9366 | 0.9137 | 0.9131 | 0.8986 | 0.9149 |
| AGIS | 0.9324 | 0.938 | 0.8889 | 0.9333 | 0.767 | 0.8473 | 0.8571 | 0.7635 |
| CIGTS | 0.9327 | 0.9751 | 0.8889 | 0.9344 | 0.9277 | 0.9294 | 0.7021 | 0.9376 |
| MD Slope | 0.9447 | 0.9506 | 0.8491 | 0.9607 | 0.8954 | 0.9324 | 0.8391 | 0.9101 |
| PLR | 0.8081 | 0.8908 | 0.8462 | 0.7952 | 0.794 | 0.8716 | 0.8303 | 0.7821 |
| PoPLR | 0.9151 | 0.9518 | 0.8696 | 0.9201 | 0.9027 | 0.9028 | 0.7692 | 0.9212 |
| VFI | 0.8174 | 0.8349 | 0.7209 | 0.8284 | 0.7725 | 0.8130 | 0.6594 | 0.7882 |

TABLE VI: Performance of forecasting progression on the internal and external test sets across different lengths of VFs ($K$) given 5 VFs ($M$=5) as input.

| | | Internal (UWHVF) Test Set | | | | External (TSGH) Test Set | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | K | ACC | AUC | SEN | SPE | ACC | AUC | SEN | SPE |
| M6 | 6 | 0.8525 | 0.9286 | 0.9200 | 0.8442 | 0.8939 | 0.9189 | 0.8095 | 0.9045 |
| | 7 | 0.8812 | 0.9362 | 0.9474 | 0.8723 | 0.8295 | 0.8782 | 0.7750 | 0.8387 |
| AGIS | 6 | 0.9263 | 0.9507 | 0.6667 | 0.9385 | 0.8961 | 0.9369 | 0.7917 | 0.9007 |
| | 7 | 0.8545 | 0.9326 | 0.9167 | 0.8500 | 0.8851 | 0.9046 | 0.7500 | 0.8914 |
| CIGTS | 6 | 0.8963 | 0.9571 | 0.8636 | 0.8985 | 0.8934 | 0.8965 | 0.7027 | 0.9054 |
| | 7 | 0.8080 | 0.9291 | 0.9412 | 0.7943 | 0.8159 | 0.7869 | 0.4615 | 0.8469 |
| MD Slope | 6 | 0.8519 | 0.9322 | 0.8800 | 0.8447 | 0.7677 | 0.8429 | 0.8155 | 0.7500 |
| | 7 | 0.6938 | 0.8562 | 0.9216 | 0.6099 | 0.6875 | 0.8024 | 0.8972 | 0.5822 |
| PLR | 6 | 0.8070 | 0.8651 | 0.7586 | 0.8232 | 0.7985 | 0.8621 | 0.7500 | 0.8120 |
| | 7 | 0.7768 | 0.8023 | 0.6458 | 0.8194 | 0.7914 | 0.8361 | 0.7042 | 0.8153 |
| PoPLR | 6 | 0.7853 | 0.8791 | 0.8889 | 0.7682 | 0.7647 | 0.8451 | 0.8739 | 0.7404 |
| | 7 | 0.7675 | 0.8704 | 0.7647 | 0.7658 | 0.7220 | 0.8319 | 0.8500 | 0.6792 |
| VFI Slope | 6 | 0.7885 | 0.8307 | 0.7333 | 0.7984 | 0.7529 | 0.8030 | 0.7083 | 0.7623 |
| | 7 | 0.7414 | 0.7662 | 0.6341 | 0.7682 | 0.6670 | 0.7538 | 0.6364 | 0.6790 |

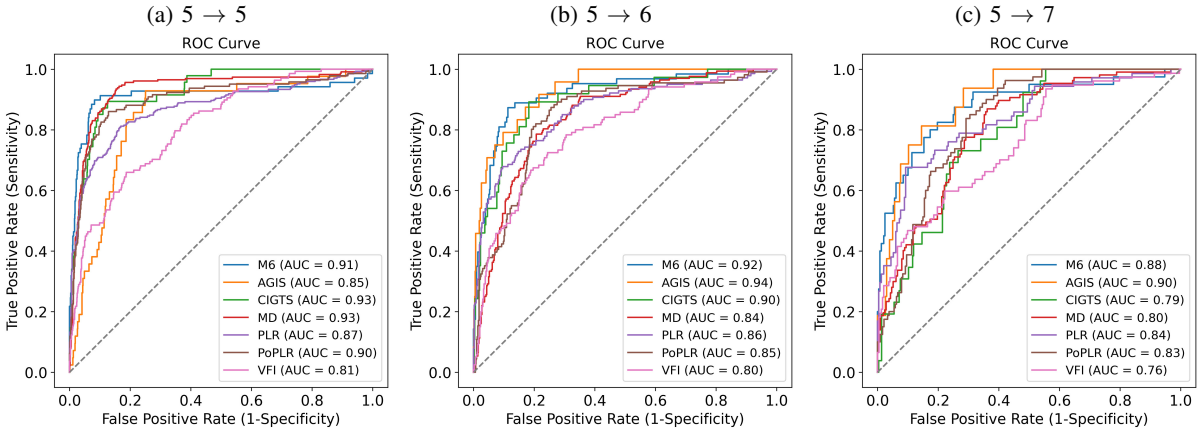


Fig. 4: ROC curves for progression prediction at $K$ (5→ $K$) on the TSGH test set given 5 VFs.

0.914 AUC, 0.899 sensitivity, and 0.955 specificity on TSGH. Amongst individual progression algorithms, VFI slope had the worst AUC and sensitivity on both test sets. Performance on the internal test set (UWHVF) was generally superior over the external test set (TSGH), as expected.

*4) VF Progression Forecasting:* We further evaluated VF-Transformer for forecasting progression across a range of VF lengths ($K = 6, 7$) given 5 VFs on both test sets. As shown in TABLE VI, the framework achieved performance of 0.929-0.936 AUCs, 0.920-0.947 sensitivity and 0.844-0.872 specificity on UWHVF, and 0.878-0.919 AUCs, 0.775-0.810 sensitivity and 0.839-0.905 specificity on TSGH, in forecasting M6 progression. Prediction biases towards sensitivity or specificity were observed in forecasting CIGTS and MD slope, especially at $K$=7. ROC curves in Fig. 4 demonstrated AUC performance on TSGH at $K$=5,6,7 given 5 VFs. There is a general trend of decreasing AUC with an increasing $K$, which confirms that forecasting is a harder prediction problem than detection.

*5) Performance analysis on the severity groups:* To analyze how the proposed framework performs in eyes with different levels of severity, we further compared the performance in severity groups defined using baseline MD, including the mild (MD >-6 dB), moderate (-12 dB < MD ≤-6 dB), and severe (MD ≤-12 dB) groups. TABLE VII shows the results on the external test set for M6 progression detection given 5 VFs. The framework demonstrated stable performance across the groups with 0.875-0.968 AUCs and 0.842-1.00

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



sensitivity. In particular, it was able to identify severe cases with 1.0 sensitivity and a reasonable specificity of 0.760, while maintaining above 0.842 sensitivity and 0.858 specificity for the other groups.

## VI. Discussion

The class imbalance issue, often observed in medical data, challenges the performance of the ML and DL models. This research tackled the issue in the context of predicting VF progression based on longitudinal VF data. With the proposed ICDT loss to handle imbalanced data along with $\tau$-normalization, our VF-Transformer framework demonstrated superior performance (0.913-0.921 AUCs, 0.857-0.899 sensitivity and 0.915-0.937 specificity on test sets) over all the models and class imbalance strategies compared in our extensive experiments (TABLES III and IV).

Comparing the losses, BS loss was unable to handle the class imbalance issue by simply reweighting the exponential term, while baseline BCE loss, FL and CDT loss had the tendency to bias toward specificity on our VF data (TABLES III and IV). The superior performance of the proposed ICDT loss over CDT loss can be explained by Fig. 3, where ICDT loss demonstrated greater efficacy in adjusting logits dynamically and creating higher gradients for the minority class when misclassified, compared to CDT loss, especially when IR is large as in our datasets.

For progression detection, Saeedi et al [5] reported 0.72-0.89 sensitivity and 0.71-0.96 specificity for majority of six (i.e., M6) given 5 VFs based on standard ML models using aggregated features. Our VF-Transformer achieved comparable 0.857-0.899 sensitivity, and 0.915-0.937 specificity for M6 progression detection given 5 VFs on the test sets (TABLE V), with higher IRs (14.41-16.71) than theirs (57/40 =1.43), indicating a much more severe class imbalance issue in our problem setting. Dixit et al [6] reported AUCs of 0.815 for MD slope, 0.794 for VFI slope, and 0.794 for PLR using ConvLSTM based on 4 VFs. When comparing using our VF data, our framework outperformed ConvLSTM with all losses and normalization settings (TABLES III and IV). Notably, Dixit et al reported the sizes of positive and negative cases for progression defined by three trend-based algorithms, which were equal to IRs of 4.63, 3.66, and 2.58 for MD slope, VFI slope, and PLR, respectively. These IRs derived from their dataset are generally lower than ours (TABLES I and II), which explains why their model did not perform well on our test sets.

Existing studies on VF progression prediction focused mainly on progression detection, rather than forecasting. Although forecasting is harder than detection, our proposed framework given 5 VFs demonstrated AUCs of 0.919 and 0.878 for forecasting M6 progression across 6 and 7 VFs, respectively, on the external test set (Fig. 4).

Further analysis on baseline MD severity groups using the external test set demonstrated stable performance of our framework across the groups (0.875-0.968 AUCs and 0.842-1.00 sensitivity), with excellent ability to capture the severe cases (1.0 sensitivity and 0.760 specificity), followed by the mild cases and moderate cases.

TABLE VII: Comparing severity group performance for M6 progression detection on TSGH test set.

| Severity | ACC | AUC | SEN | SPE | Support |
|---|---|---|---|---|---|
| Mild | 0.9468 | 0.8749 | 0.8421 | 0.9517 | 763 |
| Moderate | 0.8668 | 0.9356 | 0.9412 | 0.8582 | 158 |
| Severe | 0.7835 | 0.9677 | 1.0000 | 0.7603 | 135 |

## VII. Conclusion

In this work, we presented VF-Transformer for predicting VF progression in the presence of imbalanced classes commonly observed in medical data. To our knowledge, this is the first study to directly tackle class imbalance on longitudinal VF data. The proposed framework addressed the issue by introducing ICDT loss and integrating $\tau$-normalization. Our extensive experiments and comparisons with the state-of-the-art methods and class-imbalance handling strategies demonstrated the effectiveness and stability of VF-Transformer over all models and strategies compared. In the future, the pre-trained framework has the potential to assist clinicians in managing, tracking, and forecasting VF progression with timely predictions, following sufficient external validation.

## References

[1] K. Allison, D. Patel, and O. Alabi, "Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future," *Cureus*, vol. 12, no. 11, 2020.

[2] A. Heijl, M. C. Leske, B. Bengtsson, B. Bengtsson, M. Hussein, K. Wettrell, P. Åsman, M. Wennberg, G. Ranelycke, M. Wollmer, G. Lundskog, K. Magnusson, L. Hyman, Q. He, E. Komaroff, L. Armstrong, K. Sjöström, L. Brenner, G. Svensson, I. Abrahamson, N. E. Ahlgren, U. Andersson, A. Engkvist, L. Hagert, A. Bergstr̈om, C. Holmin, A. Glöck, C. Dahling, and I. Karlsson, "Measuring visual field progression in the early manifest glaucoma trial," *Acta Ophthalmologica Scandinavica*, vol. 81, no. 3, pp. 286–293, 2003.

[3] C. G. De Moraes, J. M. Liebmann, and L. A. Levin, "Detection and measurement of clinically meaningful visual field progression in clinical trials for glaucoma," *Progress in Retinal and Eye Research*, vol. 56, pp. 107–147, 2017.

[4] R. Hu, L. Racette, K. S. Chen, and C. A. Johnson, "Functional assessment of glaucoma: Uncovering progression," *Survey of Ophthalmology*, vol. 65, no. 6, pp. 639–661, 2020.

[5] O. Saeedi, M. V. Boland, L. D'acunto, R. Swamy, V. Hegde, S. Gupta, A. Venjara, J. Tsai, J. S. Myers, S. R. Wellik, G. Demoraes, L. R. Pasquale, L. Q. Shen, Y. Li, and T. Elze, "Development and comparison of machine learning algorithms to determine visual field progression," *Translational Vision Science and Technology*, vol. 10, no. 7, pp. 1–9, 2021.

[6] A. Dixit, J. Yohannan, and M. V. Boland, "Assessing Glaucoma Progression Using Machine Learning Trained on Longitudinal Visual Field and Clinical Data," *Ophthalmology*, vol. 128, no. 7, pp. 1016–1026, 2021.

[7] A. Chen, G. Montesano, R. Lu, C. S. Lee, D. P. Crabb, and A. Y. Lee, "Visual Field Endpoints for Neuroprotective Trials: A Case for AI-Driven Patient Enrichment," *American Journal of Ophthalmology*, vol. 243, pp. 118–124, 2022.

[8] B. Chesley and D. L. Barbour, "Visual Field Estimation by Probabilistic Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3499–3506, 2020.

[9] T. Omoto, R. Asaoka, T. Akagi, A. Oishi, M. Miyata, H. Murata, Y. Fujino, K. Hirasawa, T. Inoue, M. Tanito, and N. Shoji, "The number of examinations required for the accurate prediction of the progression of the central 10-degree visual field test in glaucoma," *Scientific Reports*, vol. 12, no. 1, pp. 1–9, 2022.

[10] K. Park, J. Kim, and J. Lee, "Visual Field Prediction using Recurrent Neural Network," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[11] Y. Taketani, H. Murata, Y. Fujino, C. Mayama, and R. Asaoka, "How many visual fields are required to precisely predict future test results in glaucoma patients when using different trend analyses?" *Investigative Ophthalmology and Visual Science*, vol. 56, no. 6, pp. 4076–4082, 2015.

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3547346



[12] S. C. K. Tékouabou, E. A. A. Alaoui, I. Chabbar, H. Toulni, W. Cherif, and H. Silkan, "Optimizing the early glaucoma detection from visual fields by combining preprocessing techniques and ensemble classifier with selection strategies," *Expert Systems with Applications*, vol. 189, no. April 2020, p. 115975, 2022.

[13] X. Huang, K. Jin, J. Zhu, Y. Xue, K. Si, C. Zhang, S. Meng, W. Gong, and J. Ye, "A Structure-Related Fine-Grained Deep Learning System With Diversity Data for Universal Glaucoma Visual Field Grading," *Frontiers in Medicine*, vol. 9, no. March, 2022.

[14] P. H. Artes and B. C. Chauhan, "Longitudinal changes in the visual field and optic disc in glaucoma," *Progress in Retinal and Eye Research*, vol. 24, no. 3, pp. 333–354, 2005.

[15] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep Long-Tailed Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023.

[16] X. Luo, J. Li, M. Chen, X. Yang, and X. Li, "Ophthalmic Disease Detection via Deep Learning with a Novel Mixture Loss Function," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 9, pp. 3332–3339, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5999–6009.

[18] M. Abu, N. A. H. Zahri, A. Amir, M. I. Ismail, A. Yaakub, S. A. Anwar, and M. I. Ahmad, "A Comprehensive Performance Analysis of Transfer Learning Optimization in Visual Field Defect Classification," *Diagnostics*, vol. 12, no. 5, 2022.

[19] P. Herbert, K. Hou, C. Bradley, G. Hager, M. V. Boland, P. Ramulu, M. Unberath, and J. Yohannan, "Forecasting Risk of Future Rapid Glaucoma Worsening Using Early Visual Field, OCT, and Clinical Data," *Ophthalmology Glaucoma*, vol. 6, no. 5, pp. 466–473, 2023.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority over-Sampling Technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[21] A. Guzmán-Ponce, J. S. Sánchez, R. M. Valdovinos, and J. R. Marcial-Romero, "DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem," *Expert Systems with Applications*, vol. 168, no. January 2020, 2021.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[23] J. Ren, C. Yu, S. Sheng, and X. Ma, "Balanced Meta-Softmax for Long-Tailed Visual Recognition," *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–12, 2020.

[24] H.-J. Ye, H.-Y. Chen, D.-C. Zhan, and W.-L. Chao, "Identifying and Compensating for Feature Deviation in Imbalanced Deep Learning," *arXiv preprint arXiv:2001.01385*, 2020.

[25] Z. Xu, R. Liu, S. Yang, Z. Chai, and C. Yuan, "Learning Imbalanced Data with Vision Transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 793–15 803.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] K. Chen, W. Lei, S. Zhao, W. S. Zheng, and R. Wang, "PCCT: Progressive Class-Center Triplet Loss for Imbalanced Medical Image Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 2026–2036, 2023.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021, pp. 1–22.

[29] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis, "Label-Imbalanced and Group-Sensitive Classification under Overparameterization," in *Advances in Neural Information Processing Systems*, vol. 23, 2021, pp. 18 970–18 983.

[30] R. Q. Wang, F. Zhu, X. Y. Zhang, and C. L. Liu, "Training with scaled logits to alleviate class-level over-fitting in few-shot learning," *Neurocomputing*, vol. 522, pp. 142–151, 2023.

[31] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling Representation and Classifier for Long-Tailed Recognition," *8th International Conference on Learning Representations, ICLR 2020*, pp. 1–16, 2020.

[32] G. Montesano, A. Chen, R. Lu, C. S. Lee, and A. Y. Lee, "UWHVF: A Real-World, Open Source Dataset of Perimetry Tests From the Humphrey Field Analyzer at the University of Washington," *Translational Vision Science and Technology*, vol. 11, no. 1, pp. 1–8, 2022.

[33] I. Mańin-Franch and W. H. Swanson, "The visualFields package: A tool for analysis and visualization of visual fields," *Journal of Vision*, vol. 13, no. 4, pp. 1–12, 2013.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] J. Ren, C. Yu, s. sheng, X. Ma, H. Zhao, S. Yi, and h. Li, "Balanced meta-softmax for long-tailed visual recognition," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 4175–4186.

Authorized licensed use limited to: National Yang Ming Chiao Tung University. Downloaded on March 04,2025 at 01:33:20 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.