

Confidence-aware multi-modality learning for eye disease screening

Ke Zou^{a,b}, Tian Lin^{c,d}, Zongbo Han^e, Meng Wang^f, Xuedong Yuan^{a,b,*}, Haoyu Chen^{c,d,*},
Changqing Zhang^e, Xiaojing Shen^{a,g}, Huazhu Fu^{f,*}

^a*National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China*

^b*College of Computer Science, Sichuan University, Chengdu, 610065, China*

^c*Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, Shantou 515041, China.*

^d*Medical College, Shantou University, Shantou 515041, China.*

^e*College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.*

^f*Institute of High Performance Computing, Agency for Science, Technology and Research, 138632, Singapore*

^g*College of Mathematics, Sichuan University, Chengdu, 610065, China*

Abstract

Multi-modal ophthalmic image classification plays a key role in diagnosing eye diseases, as it integrates information from different sources to complement their respective performances. However, recent improvements have mainly focused on accuracy, often neglecting the importance of confidence and robustness in predictions for diverse modalities. In this study, we propose a novel multi-modality evidential fusion pipeline for eye disease screening. It provides a measure of confidence for each modality and elegantly integrates the multi-modality information using a multi-distribution fusion perspective. Specifically, our method first utilizes normal inverse gamma prior distributions over pre-trained models to learn both aleatoric and epistemic uncertainty for uni-modality. Then, the normal inverse gamma distribution is analyzed as the Student's t distribution. Furthermore, within a confidence-aware fusion framework, we propose a mixture of Student's t distributions to effectively integrate different modalities, imparting the model with heavy-tailed properties and enhancing its robustness and reliability. More importantly, the confidence-aware multi-modality ranking regularization term induces the model to more reasonably rank the noisy single-modal and fused-modal confidence, leading to improved reliability and accuracy. Experimental results on both public and internal datasets demonstrate that our model excels in robustness, particularly in challenging scenarios involving Gaussian noise and modality missing conditions. Moreover, our model exhibits strong generalization capabilities to out-of-distribution data, underscoring its potential as a promising solution for multimodal eye disease screening.

Keywords: Uncertainty estimation, Eye disease, Multi-modality learning

*Corresponding authors: Xuedong Yuan (yxdongdong@163.com); Haoyu Chen (drchen-haoyu@gmail.com); Huazhu Fu (hzfu@ieee.org)

1. Introduction

Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME) stand as the primary culprits behind permanent vision impairment among individuals of working age [23]. Age-related Macular Degeneration (AMD) is another leading cause of blindness worldwide, with Polypoid Choroidal Vasculopathy (PCV) a subtype of AMD, especially seen in Asians [26, 8]. Driven by the tremendous development of computer vision [21, 14, 45], the screening and continuous monitoring of the above eye diseases under computer-aided detection is imminent.

Retinal fundus image (Fundus) and Optical Coherence Tomography (OCT) are the common 2D and 3D imaging techniques for ophthalmic diseases screening. This motivates the researchers to combine above modalities to improve the performance of ophthalmic diseases screening. After all, multi-modality learning usually provides more complementary information than uni-modality learning [54]. Existing multi-modality learning methods can be roughly classified into early, intermediate, and late fusion according to the fusion stage [3]. For the multi-modality ophthalmic image learning, recent works [52, 16, 24, 40, 48, 15, 36, 6, 25] mainly focused on the early fusion [16, 24, 40] and intermediate fusion stages [52, 48, 15, 36, 6, 25]. Previous researches typically combine features from different eye image modalities directly during fusion. However, this may lead to the collection of misjudged features from the noisy modality, resulting in incorrect prediction results \hat{y}_F , as seen in Fig. 1 (a). To address this challenge, we leverage uncertainty estimation in our method to assess the reliability of uni-modality from the perspective of individual modal distributions. As depicted in Fig. 1 (b), we estimate the prediction and uncertainty of uni-modality $\{\hat{y}_m, U_m\}_{m=1,2}$, and then leverage the distribution fusion of confidence to derive the final prediction and its uncertainty $\{\hat{y}_F, U_F\}$. This endeavor is crucial for ensuring clinical safety and reliability, particularly when dealing with interference from either image type, where uncertainty serves as a dependable metric for integrating multi-modality distributions.

Uncertainty estimations provide an excuse for ambiguous or uncertain network predictions. In particular, when a model encounters data it has never seen before or input tainted by noise, it can express uncertainty with a typo 'I don't know,' and the degree of that uncertainty can be quantified. As stated by [19], uncertainty estimation encompasses two types: aleatoric and epistemic uncertainty. Aleatoric uncertainty is inherent in the observed data and arises from inherent randomness or variability in the underlying processes. In contrast, epistemic uncertainty stems from the limitations of our knowledge or the model, indicating uncertainty that can be reduced or eliminated with additional data or improved models. Current uncertainty estimation methods mainly include the Bayesian neural networks, Deep ensemble (DE), Deterministic-based methods. Bayesian neural networks learn the distribution of network weights by treating them as random variables, using Laplace approximation [30], Markov Chain Monte Carlo [34] and variational inference techniques [38]. Affected by the challenge of convergence, these methods have a large amount of computations, until the introduction of dropout into the network has been alleviated to a certain extent [19]. Rather than to learn the distribution, a alternative and simple way to estimate the uncertainty is to learn an ensemble of deep networks [22]. To alleviate computational

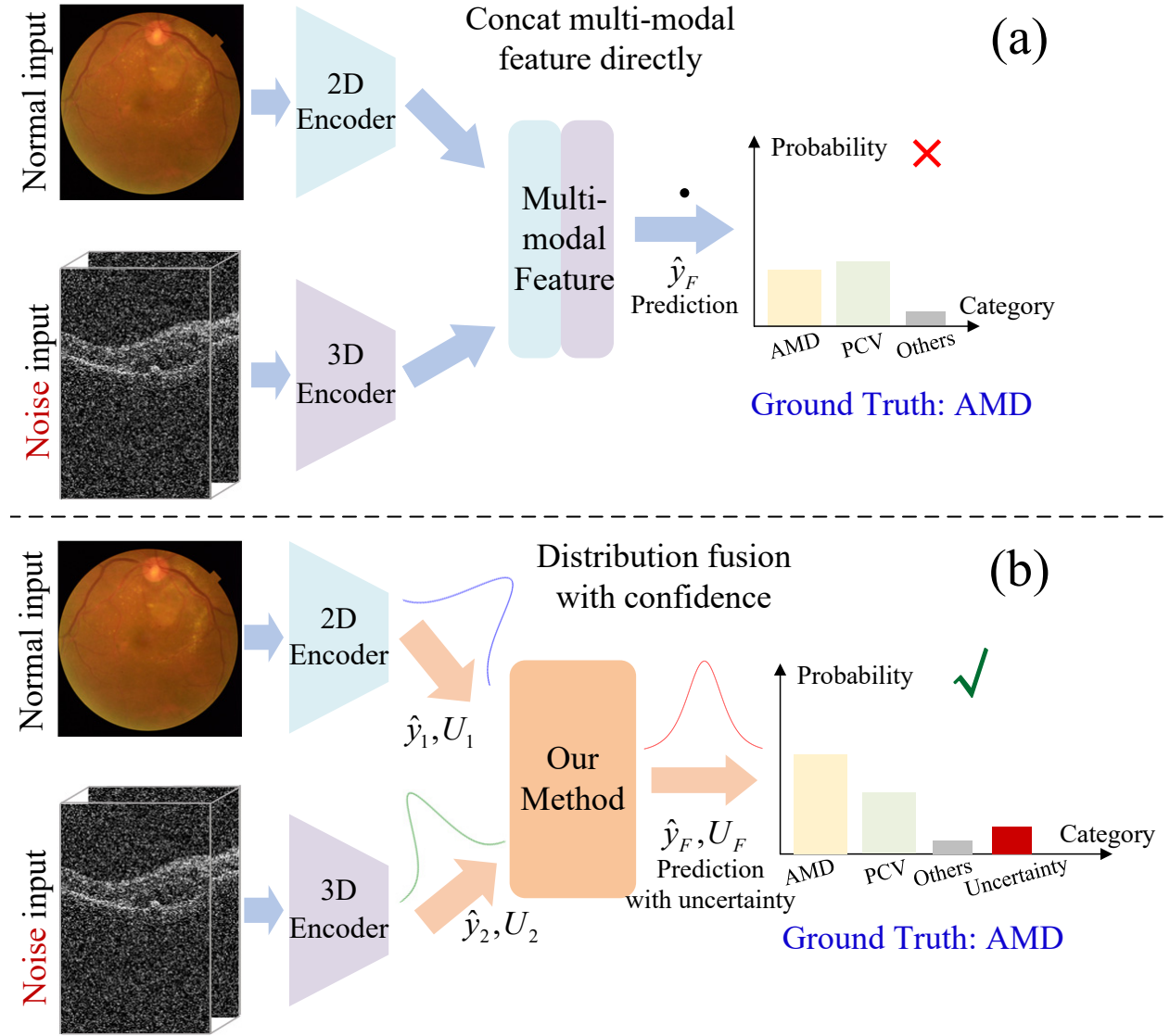


Figure 1: Comparison multi-modality classification methods for eye disease screening. (a) Traditional multi-modality eye disease screening. (b) Our confidence-aware multi-modality learning for eye disease screening. \hat{y} and U denote the prediction and its uncertainty, respectively.

complexity and overconfidence [42, 47], many deterministic-based methods [42, 31, 44, 27] have been designed to directly output uncertainty in a single forward pass through the network. Most of above methods are focused on the single-modality with uncertainty estimation. How to be aware of multi-modal uncertainty and fuse them in principle remains to be studied. Recently, an uncertainty-aware multimodal learner for estimating uncertainty through cross-modal stochastic network prediction is proposed by [46]. Notably, relying solely on cross-attention for multimodal feature fusion may not optimize post-fusion performance, especially with the presence of noisy modality, as depicted in Fig. 1. Differently, our approach provides confidence scores for each modality and elegantly integrates multi-modal

information using a multi-distribution fusion perspective.

In this paper, we present EyeMoSt+, a novel confidence-aware multi-modality eye disease screening method aimed at promoting reliable fusion of Fundus and OCT modalities. Our approach utilizes Normal-inverse Gamma (NIG) prior distributions over pre-trained models to learn both aleatoric and epistemic uncertainty for uni-modality. By solving the NIG prior analytically as a Student’s t distribution, we transform it into a mixed Student’s t distribution fusion problem. To endow the model with global uncertainty and robustness, we introduce a confidence-aware fusion strategy for Mixture of the Student’s t (MoSt) distribution. To prevent an escalation of global confidence in the presence of modal noise, we utilize a novel confidence-aware ranking-based regularization approach. To validate the effectiveness of our proposed method, we conduct extensive experiments on three datasets, covering various eye diseases, such as glaucoma grading, AMD, PCV, DR, and DME. These experiments underscore the reliability and robustness of EyeMoSt+ in multimodal screening for eye diseases, highlighting its effectiveness in processing noisy inputs, identifying missing patterns, and handling unseen data. In summary, the contributions of this paper are mainly included:

- (1) We propose a novel confidence-aware multi-modality eye disease screening method, which providing a new evidential multi-modality paradigm for classification with reliability and robustness.
- (2) To integrate different modalities, a novel MoSt is designed to be dynamically aware of heavy-tailed and confidence for each modality with uncertainty, which promisingly provides significantly robustness as well and promotes reliable decision.
- (3) To address the confidence relationship between uni-modality and fusion modality, we propose a novel confidence-aware ranking regularization term for multi-modality eye disease screening.
- (4) We conducted comprehensive experiments on both public and internal datasets encompassing various eye diseases to thoroughly validate the accuracy, robustness, and reliability of our model, including its performance on Out-of-distribution (OOD) test samples such as those with Gaussian noise, missing modality, and unseen data. ¹

We compared three methods [2, 41, 56] directly related to our research, outlined below:

- (1) Compared with the evidential deep regression method [2], our work extends this framework to the domain of ophthalmic classification, introducing confidence-aware evidential multi-modality fusion. Leveraging [2] method, we employ the evidence prior distribution NIG to characterize confidence for different modalities (Eqs. 1 to 4). However, the original method lacks a solution for integrating the prediction and confidence of different modalities. To address this limitation, we propose MoSt, a dynamic fusion approach that merges predictions and uncertainties from diverse modalities (Eq. 7). Additionally, to ensure that the confidence level of the fusion modalities consistently exceeds the confidence level of uni-modality, we introduce a novel confidence-aware ranking regularization term tailored for multimodal eye disease screening (Eq. 12).
- (2) In comparison with [41], where fusion involves two Student-t distributions, we enhance

¹Our code has been released in <https://github.com/Cocofeat/EyeMoSt>.

this process and propose a ranking regularization term for confidence perception. Building on [41], our confidence-aware fusion strategy (Eq. 7) is an improved version. We assume that multiple Student’s t distributions remain approximate Student’s t distribution after fusion, with the degrees of freedom ν_F and Σ_F aligning with [41]. The calculation of u_F introduces the confidence \mathcal{C} of both modalities. Furthermore, we construct an evidence prior distribution NIG for different modalities and transform it into two Student’s t distributions for fusion (Eq. 1). Besides, a new confidence-aware ranking regularization term for multi-modal eye disease screening is introduced to establish the ranking relationship between the confidence of the fusion modality and the confidence of each single modality (Eq. 12).

(3) Compared with our prior conference version [56], we further enhance the fusion process of the mixture of Student’s t distributions (Eq. 7), introducing an innovative confidence-aware multi-modal learning ranking component. Our contributions also encompass more robust validation, OOD data detection, and missing modality experiments in practical applications. We refine the fusion modality of the mixture of Student’s t distribution, incorporating Eqs. 10 to 12 and Eq. 18 to propose a new confidence perception ranking regularization term for multimodal eye disease screening. Additionally, we enrich the robust experimental verification in Sec. 4.3 and 4.4 and perform OOD data verification in different scenarios. Finally, the addition of missing modality experiments in Sec. 4.5 aim to comprehensively verify the robustness and reliability of the proposed algorithm.

2. Related Works

In this section, we first briefly review multi-modality learning for eye disease screening. Then, different uncertainty quantification methods are introduced.

2.1. Multi-modality learning for eye disease screening

According to integrate of multi-modality fusion at different stages, existing multi-modality image methods for typical ophthalmic diseases screening can be divided into methods that fuse in early, intermediate, and late stages [3]. Early fusion-based approaches integrated multiple modalities directly at the data level, usually by concatenating the raw or preprocessed multi-modality data. Rodrigues et al. [40] prone to use the complementary features based on grayscale and vessel connectivity attributes in the early fusion stage. The following methods tend to fuse special raw data early rather than stitching multimodal raw images directly. Hua et al. [16] combined the preprocessed Fundus image and wide-field swept-source Optical Coherence Tomography Angiography (OCTA) at the early stage and then extracting representational features for DR recognition. Li et al. [24] obtained synthesized FFA data through CycleGAN [55], and then feeds into a convolutional neural network (CNN) with paired FFA and Fundus data. They tried to learn both modality-invariant features and patient-similarity features for retinal disease diagnosis. The early fusion stage methods can preserve the original image information to the greatest extent, and most people currently perform multi-modality ophthalmic image fusion at the intermediate stage for disease screening.

The intermediate fusion strategies allow multiple modalities to be fused at different intermediate layers of the network. Yoo et al. [52] first attempted to diagnose AMD from multi-modality images at the intermediate fusion stage. They used pre-trained VGG-Net model to extract features, then aggregated them and diagnosed AMD by random forest classifier. Different from directly aggregating the features extracted by the pre-trained model, Wang et al. [49, 48] trained end-to-end two-stream CNN with class activation mapping and then concatenated information from the Fundus and OCT streams. Similarly, Ou et al. [36] and He et al. [15] extracted the different modality features with CBAM [50] and modality-specific attention mechanisms, then concatenated them to realize the multi-modality fusion for retinal image classification. Cai et al. [6] does well in capturing domain-specific features embedded in ophthalmic images in the early and intermediate fusion stages to achieve classification. Above methods on the early and intermediate fusion stage are too simple and lack of exploiting the complementary information between Fundus and OCT modality. Therefore, Li et al. [25] combined features across multiple dimensions of the network and explored the relation between them by a hierarchical fusion strategy. However, in the later stage of fusion, the features of each modality and the features of hierarchical fusion are only concatenated for eye diseases classification. While the identification of DR diseases using Ultra-WideField Color Fundus Photography (UWF-CFP) imaging and OCTA was undertaken by [9], a manifold mixup strategy was incorporated to enhance the generalization of concatenated features. In the late fusion stage, more attention should be paid to how to combine the predictions of these multiple models robustly and reliably. Therefore, in this paper, we try to focus on adaptive fusion based on uncertainty estimation in the late fusion stage for multi-modality eye disease screening. Our aim is to integrate information from various modalities using a multi-distribution fusion approach, particularly emphasizing Student’s t distribution fusion. This methodology, although previously explored in medical image registration [11, 39] and segmentation [35], offers promising avenues for advancing eye disease screening. For instance, pixel similarity within MoSt algorithm was introduced for the rigid registration of multimodal medical images [11]. Building upon this foundation, Ravikumar et al. [39] proposed group-wise similarity registration to enhance correspondence and align shapes more robustly. Inspired by the aforementioned methods, we propose the MoSt within a confidence-aware fusion framework to effectively integrates different modalities for robust and reliable eye disease diagnosis.

2.2. Uncertainty estimation

Uncertainty quantification provides reliable predictions and confidence levels, which are critical for advancing explainable deep neural networks (DNNs) [1]. Bayesian neural networks (BNNs) [4, 32, 17] models uncertainty by learning a distribution of deterministic parameters. Commonly used techniques include Laplace approximation [30], Markov Chain Monte Carlo [34] and variational inference [38]. Although BNNs are robust to overfitting problems, they can be unacceptably computationally intensive. To address this problem, Kendall et al. [19] introduced a straightforward method, leveraging Bayesian deep learning with Monte Carlo Dropout (MCDO), to model both aleatoric and epistemic uncertainty in the context of computer vision. DE [22] trained and integrated multi deep learning mod-

els to produce uncertainty. However, there is still a certain consumption of memory and computing costs.

Recently, deterministic-based methods [27, 44] are designed to estimate the uncertainty by a single forward pass without much sampling and time cost. Van Amersfoort et al. [44] proposed to measure the distance between the test sample and the prototype as a deterministic uncertainty based on the idea of building an radial basis function network. Furthermore, it was contended by [18] that uncertainty estimation for multimodal data remained a challenge. They introduced multimodal neural processes incorporating several innovative and principled mechanisms designed to address the specific characteristics of multimodal data. Quality-aware multimodal fusion was introduced by [53] to attain robust multimodal fusion. Nevertheless, they did not explicitly characterize the aleatoric and epistemic uncertainty of each modality, potentially limiting the model’s ability to effectively perceive and differentiate data quality. In this paper, our model is extended the deep evidential regression [2] to classification for multi-modality fusion. We focus on the modality-specific uncertainty estimation and how to fuse multi-modality estimation more confidently and reliably.

3. Proposed method

In this section, we introduce the overall framework of EyeMoSt+, which efficiently estimates the aleatoric and epistemic uncertainty for each modality and integrate the fusion modality in principle adaptively. As shown in Fig. 2, we first employ the pretrained CNN or transformer encoders to capture different modality features. Then, we place multi-evidential heads after the trained networks and to model the parameters of higher-order evidential distributions for each modality. To merge these predicted distributions, we derive the normal inverse Gamma distributions to Student’s t (St) distributions. Particularly, the confidence-aware fusion for mixture of St distributions (MoSt) is introduced to integrate the St distributions of different modalities in principle. Besides, a novel confidence-aware ranking regularization term is proposed to constrain the confidence relationship between unimodality and fusion modality. Finally, we elaborate on the training pipeline for the model evidence acquisition.

3.1. Prediction & uncertainty estimation for each modality

Given a multi-modality ophthalmic dataset $\mathcal{D} = \left\{ \{\mathbf{x}_m^i\}_{m=1}^M \right\}$ and the corresponding label y^i , the intuitive goal is to learn a function that can classify different categories. In the ophthalmic image, OCT and Fundus are common imaging modalities. Therefore, here $M=2$, \mathbf{x}_1^i and \mathbf{x}_2^i represent OCT and Fundus input modality data, respectively. We first load pretrained 2D CNN-based [10] or transformer-based [28] backbone encoder Θ and the 3D CNN-based [7] or transformer-based [13] backbone encoder Φ to identify the feature-level informativeness, which can be defined as $\Theta(\mathbf{x}_1^i)$ and $\Phi(\mathbf{x}_2^i)$, respectively. We extend the deep evidential regression model [2] to deep multi-modality evidential classification for eye diseases screening. To this end, to model the uncertainty for each modality, we assume that the observe label y^i is drawn from a Gaussian $\mathcal{N}(y^i|\mu, \sigma^2)$, whose mean and variance are

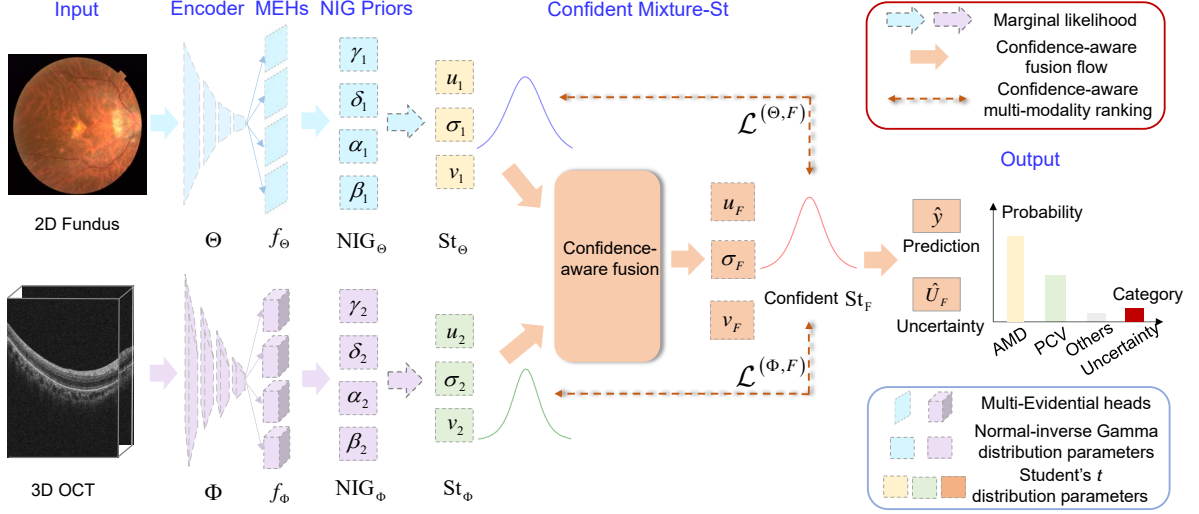


Figure 2: The framework of confidence-aware multi-modality learning for eye disease screening (EyeMoSt+).

governed by an evidential prior named the Normal-Inverse-Gamma (NIG):

$$\text{NIG}(\mu, \sigma^2 | \mathbf{p}_m) = \mathcal{N}\left(\mu | \gamma_m, \frac{\sigma^2}{\delta_m}\right) \Gamma^{-1}(\sigma^2 | \alpha_m, \beta_m), \quad (1)$$

where Γ^{-1} is an inverse-gamma distribution. Specifically, the multi-evidential heads will be placed after the encoders Θ and Φ (as shown in Fig. 2), which outputs the prior NIG parameters $\mathbf{p}_m = (\gamma_m, \delta_m, \alpha_m, \beta_m)$. As a result, the Aleatoric Uncertainty (AU) and Epistemic Uncertainty (EU) can be estimated by the $\mathbb{E}[\sigma^2]$ and the $\text{Var}[\mu]$, respectively:

$$\text{AU} = \mathbb{E}[\sigma^2] = \frac{\beta_m}{\alpha_m - 1}, \quad \text{EU} = \text{Var}[\mu] = \frac{\beta_m}{\delta_m(\alpha_m - 1)}. \quad (2)$$

After that, the Student's t predictive distributions can be derived, which are formed by the interaction of the prior and the Gaussian likelihood of each modality, given by:

$$\begin{aligned} p(y^i | \mathbf{p}_m) &= \frac{p(y^i | \theta, \mathbf{p}_m) p(\theta | \mathbf{p}_m)}{p(\theta | y^i, \mathbf{p}_m)} \\ &= \int_u \int_{\sigma^2} p(y^i | x_m^i, u, \sigma^2) \text{NIG}(\mu, \sigma^2 | \mathbf{p}_m) du d\sigma^2, \end{aligned} \quad (3)$$

When placing a NIG evidence prior on our Gaussian likelihood function, there exists an analytical solution as follows

$$\begin{aligned} p(y^i | \mathbf{p}_m) &= \frac{\Gamma(\alpha_m + \frac{1}{2})}{\Gamma(\alpha_m)} \sqrt{\frac{\delta_m}{2\pi\beta_m(1+\delta_m)}} \left(1 + \frac{\delta_m(y^i - \gamma_m)^2}{2\beta_m(1+\delta_m)}\right)^{-(\alpha_m + \frac{1}{2})} \\ &= \text{St}(y^i; \gamma_m, o_m, 2\alpha_m), \end{aligned} \quad (4)$$

Where $o_m = \frac{\beta_m(1+\delta_m)}{\delta_m\alpha_m}$. Thus, the two modalities distributions are transformed into the Student's t distributions $\text{St}(y^i; u_m, \sigma_m, v_m) = \text{St}\left(y^i; \gamma_m, \frac{\beta_m(1+\delta_m)}{\delta_m\alpha_m}, 2\alpha_m\right)$

3.2. Confidence-aware fusion for Mixture of Student's t distributions

Then, we focus on fusing multiple St distributions from different modalities. How to rationally integrate multiple St s into a unified St is the key issue. To this end, the joint modality of distribution can be denoted as:

$$p(x_1, x_2) = St(y^i; u_F, \Sigma_F, v_F), \quad (5)$$

Then the joint t distribution with:

$$p(x_1, x_2) = St\left(y^i; \begin{bmatrix} u_1^i \\ u_2^i \end{bmatrix}, \begin{bmatrix} \Sigma_1^i \\ \Sigma_2^i \end{bmatrix}, \begin{bmatrix} v_1^i \\ v_2^i \end{bmatrix}\right). \quad (6)$$

In order to preserve the closed Student's t distribution form and the heavy-tailed properties of the fusion modality, the updated parameters are given by [41]. In simple terms, we first adjust the degrees of freedom of the two distributions to be consistent. As described in [41], the smaller values of degrees of freedom has heavier tails, while the larger variance values represent better heavy tails too. Furthermore, considering the variance formula of the Student's t distribution (like Eq. 9), it's important to note that as v increases, the variance decreases, indicating a higher level of confidence. We assume that multiple Student's t distributions are still an approximate Student's t distribution after fusion. Assuming that the degrees of freedom of v_1 are smaller than v_2 , then, the fused Student's t distribution $St(y^i; u_F, \Sigma_F, v_F)$ will be updated as:

$$\begin{cases} v_F = v_1 \\ u_F = C_1 u_1 + C_2 u_2 \\ \Sigma_F = \frac{1}{2} \left(\Sigma_1 + \frac{v_2(v_1-2)}{v_1(v_2-2)} \Sigma_2 \right) \end{cases}, \quad (7)$$

Where C_1 and C_2 denote the confidence from the distribution of uni-modality, which can be defined as:

$$C_1 = \frac{v_1}{v_1 + v_2}, \quad C_2 = \frac{v_2}{v_1 + v_2}. \quad (8)$$

Therefore, the prediction and uncertainty for the fused modality can be estimated by:

$$\begin{aligned} \hat{y}^i &= \int y^i p(y^i | x_F^i, \mathbf{p}_F) dy^i = u_F, \\ U_F &= \Sigma_F \frac{v_F}{v_F-2} = \Sigma_F \left(1 + \frac{2}{v_F-2} \right), \end{aligned} \quad (9)$$

where \mathbf{p}_F is the parameter of St distribution after fusion, which can be denoted as $\mathbf{p}_F = (u_F, \sigma_F, v_F)$. Confidence-aware fusion for $MoSt$ can be seen in Fig. 3 ①.

3.3. Confidence-aware multi-modality ranking

In contemporary multimodal methodologies, the direct fusion of potentially corrupted modalities is a prevalent practice, leading to compromised model reliability and recognition errors. Accurately defining the reliability of each modality poses challenges, particularly when dealing with diverse confidence levels across different modalities for the same sample,

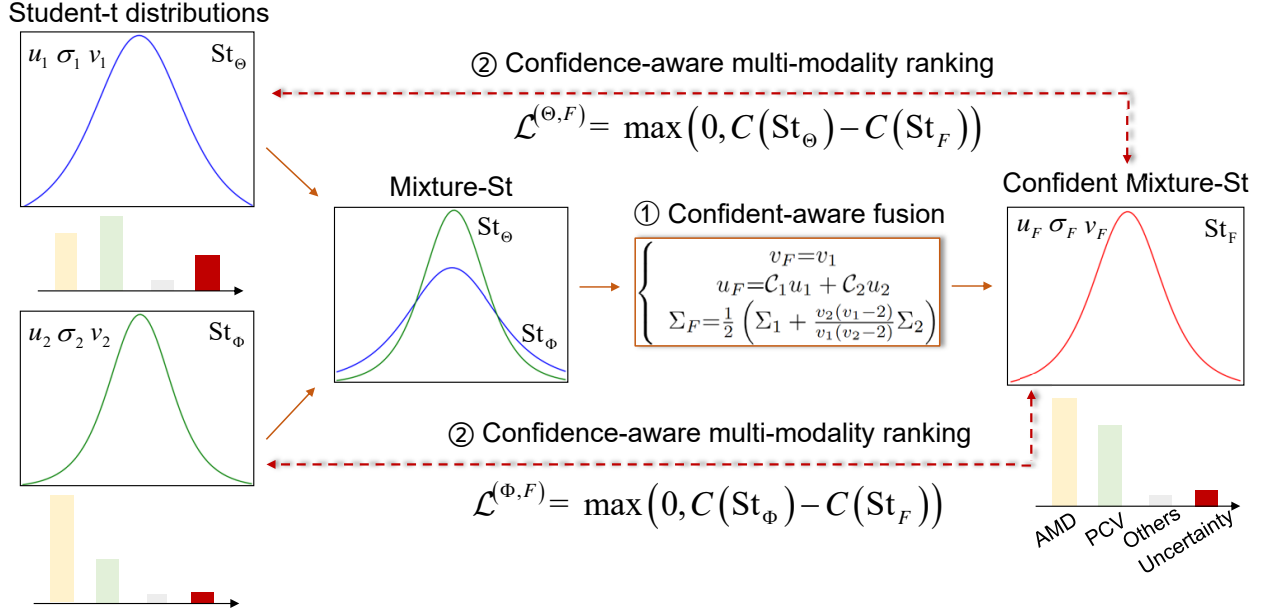


Figure 3: Confidence-aware fusion for Mixture of Student’s t distributions and Confidence-aware multi-modality ranking.

especially in the presence of noisy data. Fortunately, the supervision of confidence estimations can serve as a viable alternative. Drawing inspiration from the informatics principle of “the essence of information is to eliminate uncertainty (Shannon)” [43], where greater information implies reduced uncertainty. That is, for a reliable multi-modality classifier, integrating multimodal information will eliminate uncertain parts of the information, resulting in more confident results. Based on this assumption, we introduce a ranking-based regularization term [29]. This term constrains the relationship between single modality and fused modality, ensuring that the confidence level of the fused modality consistently surpasses that of each individual modality. As a result, the model’s reliability is significantly bolstered.

Specifically, we first directly minimize the confidence difference between the uni-modality and fusion modality as follows:

$$\mathcal{L}^{(m, F)} = \mathcal{C}(St_m) - \mathcal{C}(St_F), \quad (10)$$

where $\mathcal{C}(\cdot)$ represents the confidence of the modality, defined by the logits generated through the softmax layer [29]. Despite the presence of modal contamination, the fused models can, at times, yield accurate predictions. Hence, we solely focus on regularizing the confidence in correct predictions, while avoiding the minimization of confidence for individual modalities. The aforementioned formula can be relaxed for any ophthalmic modality image as follows:

$$\mathcal{L}^{(m, F)} = \max(0, \mathcal{C}(St_m) - \mathcal{C}(St_F)). \quad (11)$$

The proposed confidence-aware multimodal ranking loss is integrated over arbitrary modality

and fused modality pairs for each sample, formalized as follows:

$$\mathcal{L}_C = \sum_{m=1}^M \mathcal{L}^{(m,F)}. \quad (12)$$

The proposed confidence-aware multi-modality regularization term is versatile and can be seamlessly incorporated as an additional loss term into current evidential deep learning framework to constrain their confidence estimates. Confidence-aware multi-modality ranking rule can be seen in Fig. 3 ②. This integration enhances model reliability and boosts performance robustness.

3.4. Multi-modality learning process

Under the evidential learning framework, we expect more evidence to be collected for each modality, thus, the proposed model is expected to maximize the likelihood function of the model evidence. Equivalently, the model is expected to minimize the negative log-likelihood function, which can be expressed as:

$$\begin{aligned} \mathcal{L}_m^{NLL}(\gamma_m, \delta_m, \alpha_m, \beta_m) &= \log \frac{\Gamma(\alpha_m) \sqrt{\frac{\pi}{\delta_m}}}{\Gamma(\alpha_m + \frac{1}{2})} - \alpha_m \log(2\beta_m(1 + \delta_m)) \\ &+ (\alpha_m + \frac{1}{2}) \log((y - \gamma_m)^2 \delta_m + 2\beta_m(1 + \delta_m)), \end{aligned} \quad (13)$$

Then, to fit the classification tasks, we introduce the cross entropy term \mathcal{L}_m^{CE} to Eq. 7:

$$\mathcal{L}_m^{NIG} = \mathcal{L}_m^{NLL} + \lambda_m \mathcal{L}_m^{CE} \times \eta_m, \quad (14)$$

where λ_m is the balance factor and set to be the same as [2]. η_m is the overall model evidence, which can be denoted as:

$$\eta_m = \alpha_m + \delta_m + \frac{1}{\beta_m}. \quad (15)$$

Similarly, for the fused modality, we first maximize the likelihood function of the model evidence as follows:

$$\begin{aligned} \mathcal{L}_F^{NLL}(u_F, \Sigma_F, v_F) &= \log \sqrt{\Sigma_F} + \log \frac{\Gamma(\frac{v_F}{2})}{\Gamma(\frac{v_F+1}{2})} + \log \sqrt{v_F \pi} \\ &+ \frac{1}{2} (v_F + 1) \log \left(1 + \frac{(y_t - u_F)^2}{v_F \Sigma_F} \right). \end{aligned} \quad (16)$$

Then, to achieve better classification performance, the cross entropy term \mathcal{L}_m^{CE} is also introduced into Eq. 16 as below:

$$\mathcal{L}_F^{St} = \mathcal{L}_F^{NLL} + \lambda_F \mathcal{L}_F^{CE}, \quad (17)$$

Where λ_F serves as the balance factor, its optimal selection is identified through the ablation study (Sec 4.6). Totally, the evidential deep learning process for multi-modality screening can be denoted as:

$$\mathcal{L}_{all} = \sum_{m=1}^M \mathcal{L}_m^{NIG} + \mathcal{L}_F^{St} + \lambda_C \mathcal{L}_C, \quad (18)$$

λ_C serves as a crucial hyperparameter that governs the potency of confidence-aware multi-modality learning regularization. Its value is set to 10 based on the insights from [29] and the results of the ablation study (Sec 4.6). This paper primarily focuses on discussing eye diseases using two modalities: OCT and Fundus imaging. Accordingly, the parameter M is set to 2. The process of proposed EyeMoSt+ are shown in Algorithm 1.

Algorithm 1 Confidence-aware multi-modality learning for eye disease screening

Given dataset $\mathcal{D} = \{\{x_i^m\}_{m=1}^M, y_i\}_{i=1}^N$, initialized classifier $\mathcal{F} = \{f^m\}_{m=1}^M$, hyperparameter $\lambda_m = 0.01$, $\lambda_F = 0.5$, $\lambda_C = 10$, and epochs for training the classifier t_e .

for $t = 1, \dots, t_e$ **do**

for $m = 1, \dots, M$ **do**

 Place the NIG prior for each modality with Eq. 1: $\text{NIG}(\mu, \sigma^2 | \mathbf{p}_m) \leftarrow$ each encoder outputs;

 Compute the analytical solution for each modality with Eq. 4: $St(y^i; \gamma_m, o_m, 2\alpha_m) \leftarrow \text{NIG}(\mu, \sigma^2 | \mathbf{p}_m)$;

 Compute the each modality loss with Eq. 14;

end for

 Obtain the fusion modality parameters with Eq. 7;

 Compute the fusion modality loss with Eq. 17;

 Compute the confidence-aware regularization loss with Eq. 11;

 Compute the total loss with Eq. 18;

 Update the parameters of the networks with gradient descent;

end for

return networks parameters.

4. Experiments

4.1. Datasets & Training Details

1) **Experimental Datasets:** This paper conducts a comprehensive evaluation of the performance of the EyeMoSt+ model across the public and private datasets : GAMMA dataset [51], OLIVES [37] and an in-house dataset developed for this study. The different datasets for different diseases are detailed below.

GAMMA Dataset for Glaucoma Recognition: To assess the efficacy of our proposed approach in glaucoma recognition, we assess its performance on the GAMMA dataset [51]. This dataset comprises 100 paired cases, each assigned a three-level glaucoma grading. The original image size for OCT and near-IR Fundus images is $256 \times 512 \times 992$ and 1956×1934 , where 256 is the total number of OCT slices. More details about the original dataset can be found in [51]. The cases are thoughtfully divided into training and test subsets, containing 80% and 20% of the cases, respectively. To mitigate the influence of incidental factors on performance evaluations, a rigorous five-fold cross-validation strategy is employed.

OLIVES Dataset for DR and DME screening: The effectiveness of the proposed algorithm in identifying DR and DME is subsequently verified using the OLIVES dataset [37]. This

dataset comprises paired OCT and near-IR Fundus images from 96 patients over multiple cycles, yielding a total of 3128 paired cases. More specifically, it includes 56 patients with DME and 40 patients with DR at various weeks, resulting in a total of 1837 DME samples and 1291 DR samples. The original image size for OCT and near-IR Fundus images is $48 \times 504 \times 496$ and 768×768 , where 48 is the total number of OCT slices. Additional details about the original dataset can be found in [37]. To ensure reliable experimentation, we partitioned the dataset into training, validation, and test subsets, maintaining an 8:1:1 ratio.

In-house Dataset for AMD and PCV Screening: Finally, our method undergoes rigorous testing using an exclusive in-house dataset obtained from the Shantou International Joint Eye Center at Shantou University, utilizing Topcon 3D OCT-2000 as OCT and Fundus acquisition device. The dataset comprises 149 cases of AMD and 178 cases of PCV, involving a total of 327 patients. Some cases feature both left and right eye images, resulting in a total of 604 paired OCT and Fundus images, including 265 AMD samples and 341 PCV samples. The original image size for OCT and Fundus images is $128 \times 512 \times 885$ and 2100×2000 , where 128 is the total number of OCT slices. Adhering to established practices, we partition the patient cohort into training, validation, and test subsets, maintaining a consistent 8:1:1 patient ratio for reliable experimentation.

2) Training Details: Our proposed method is implemented in PyTorch and trained on NVIDIA GeForce RTX 3090. Adam optimization [20] is employed to optimize the overall parameters with an initial learning rate of 0.0001. The maximum of epoch is 100. The original image size in the GAMMA dataset is comparable to that of the internal dataset. For OOD-related experiments using the GAMMA dataset (as detailed in Section 4.4), we uniformly adjusted the original sizes of its OCT and Fundus images to $128 \times 256 \times 128$ and 256×256 , respectively. Given the size disparity between OLIVES and the original images from GAMMA and the internal dataset, our aim is to optimize the utilization of the NVIDIA GeForce RTX 3090 graphics card memory while maximizing information retention in the original images. Consequently, we adjusted the Fundus and OCT image sizes for the OLIVES dataset to 512×512 and $48 \times 248 \times 248$. The batch size is set to 16. In all the following experiments involving the addition of Gaussian noise, we apply a ten-fold random addition strategy to mitigate any performance improvements resulting from random factors. It should be noted that our proposed EyeMoSt+ can be divided into two versions, EyeMoSt+ (CNN) and EyeMoSt+ (Transformer), depending on the encoders used. Therefore, for the EyeMoSt+ (Transformer) version, to align with the input requirements of the pre-trained models [28, 13], we adjusted the input sizes for Fundus and OCT to 384×384 and $96 \times 96 \times 96$ for all the datasets.

4.2. Compared Methods & Metrics

1) Compared Methods: We compare the following six methods: For different fusion stage strategies, **a) B-CNN** Baseline of intermediate typical fusion method based on CNNs, **b) B-Transformer** Baseline of intermediate typical fusion method based on transformers, **c) B-EF** Baseline of the early fusion [16] strategy, **d) M^2LC** [50] of the intermediate fusion method and the later fusion method **e) TMC** [12] are used as comparisons. B-EF is first

integrated at the data level, and then passed through the same MedicalNet [7]. B-CNN and B-transformer first extract features by the encoders (same with us), and then concatenates their output features as the final prediction. In addition, we compared **f) SmartDSP** [5] and **g) EyeStar** [51], which ranked first and third on the GAMMA dataset. For the uncertainty quantification methods, **h) MCDO** [19] employs the test time dropout as an approximation of a Bayesian neural network. **i) DE** [22] quantifies the uncertainties by ensembling multiple models. In the case of all the baselines and our proposed EyeMoSt+, we selected the best checkpoint for testing based on the validation performance using the Accuracy (ACC) metric.

2) Performance Metrics and Evaluation: In our evaluation, we employ ACC and Kappa metrics, which offer an intuitive basis for comparing our method with other existing approaches. To quantify the effectiveness of ordinal ranking, we utilize area under risk-coverage (AURC). This enables a comprehensive understanding of how our method’s risk estimates align with the actual outcomes. For the calibration, we employ the Expected Calibration Error (ECE) [32] metric.

Table 1: Comparisons with different algorithms on the GAMMA dataset. F and O denote Fundus and OCT modality. The top-2 results are highlighted in **bold** and underlined for our method. Higher ACC and Kappa mean better. Lower ECE means better.

Method	Original			Gaussian noise							
				$\sigma=0.1$ (F)			$\sigma=0.3$ (O)				
	ACC	Kappa	ECE	ACC	Kappa	ECE	ACC	Kappa	ECE	P-value	Time (s)
B-CNN	0.700	0.515	0.340	0.623	0.400	0.530	0.500	0.000	0.740	0.0004	3.79
B-Transformer	0.780	0.641	0.230	0.664	0.459	0.372	0.733	0.574	0.277	0.0305	1.95
B-EF	0.660	0.456	0.350	0.660	0.452	0.360	0.500	0.000	0.740	0.0004	3.70
M^2 LC	0.710	0.527	0.290	0.660	0.510	0.352	0.500	0.000	0.740	0.0004	3.83
SmartDSP	0.840	0.743	<u>0.170</u>	0.530	0.323	0.380	0.800	0.679	0.220	0.0270	15.17
EyeStar	0.860	0.774	0.150	0.650	0.439	0.380	0.740	0.583	0.250	0.0050	26.79
MCDO	0.758	0.636	0.253	0.601	0.341	0.494	0.530	0.000	0.740	0.0004	27.79
DE	0.710	0.539	0.330	0.666	0.441	0.385	0.530	0.000	0.730	0.0004	31.14
TMC	0.810	0.658	0.230	0.430	0.124	0.919	0.580	0.045	0.700	0.0360	4.27
EyeMoSt	<u>0.850</u>	0.754	<u>0.170</u>	0.663	0.458	0.368	<u>0.830</u>	<u>0.716</u>	0.210	-	3.90
EyeMoSt+T	0.820	0.732	0.180	0.751	0.579	0.294	0.827	0.710	<u>0.193</u>	-	<u>2.18</u>
EyeMoSt+C	0.860	<u>0.761</u>	0.150	<u>0.675</u>	<u>0.464</u>	<u>0.350</u>	0.850	0.764	0.175	-	4.09

4.3. Robustness validation

1) GAMMA dataset: We first reported our algorithm with start-of-the-art methods on the GAMMA dataset in Tab. 1. We conducted a performance comparison of various methods under the normal condition, including the top-ranking SmartDSP (1st place) and EyeStar (3rd place) methods from the GAMMA Challenge. Based on the three metrics of ACC, Kappa, and ECE under the normal condition in Tab. 1, we observed that our proposed EyeMoSt+ (CNN) method achieved comparable performance, securing the second position. To assess the robustness of the proposed method, we introduced Gaussian noise with $\sigma = 0.1$ or $\sigma = 0.3$ to the Fundus modality or the OCT modality, respectively, during testing. As

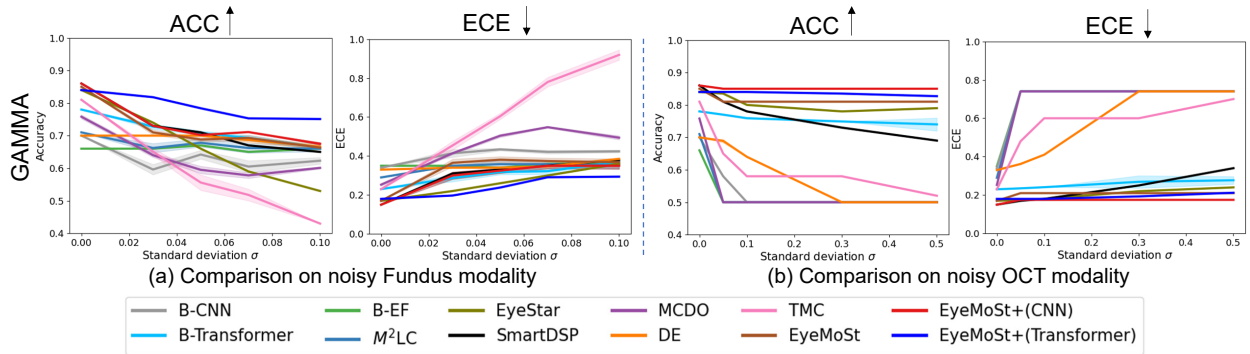


Figure 4: Accuracy and ECE performance of different algorithms in noisy single modality with different levels of noise on GAMMA dataset. (a) ACC and ECE for various algorithms in the presence of noise at different levels in Fundus modality. (b) ACC and ECE for various algorithms in the presence of noise at different levels in OCT modality. Higher ACC and Lower ECE mean better.

Table 2: Comparisons with different algorithms on the OLIVES dataset. F and O denote Fundus and OCT modality. The top-2 results are highlighted in **bold** and underlined for our method. Higher ACC and Kappa mean better.

Method	Original		Gaussian noise					
			$\sigma=0.5$ (F)		$\sigma=0.3$ (O)		P-value	Time (s)
	ACC	Kappa	ACC	Kappa	ACC	Kappa		
B-CNN	1.000	0.840	0.669	0.000	0.377	0.047	$\ll 0.001$	11.01
B-Transformer	1.000	1.000	1.000	1.000	0.669	0.000	$\ll 0.001$	<u>10.67</u>
B-EF [16]	0.986	0.968	<u>0.979</u>	<u>0.953</u>	0.331	0.000	$\ll 0.001$	9.13
M^2 LC [50]	1.000	1.000	0.957	0.905	0.389	0.059	$\ll 0.001$	11.86
MCDO [19]	1.000	1.000	1.000	1.000	0.373	0.035	$\ll 0.001$	34.12
DE [22]	1.000	1.000	1.000	1.000	0.331	0.000	$\ll 0.001$	36.73
TMC [12]	1.000	0.835	0.775	0.389	0.557	0.253	$\ll 0.001$	12.96
EyeMoSt [56]	0.932	0.838	1.000	1.000	<u>0.932</u>	<u>0.838</u>	-	12.08
EyeMoSt+ (Transformer)	1.000	1.000	1.000	1.000	0.775	0.451	-	11.35
EyeMoSt+ (CNN)	<u>0.993</u>	<u>0.984</u>	1.000	1.000	0.981	0.957	-	12.20

illustrated in Tab. 1, the addition of $\sigma = 0.1$ Gaussian noise to the Fundus modality led to a significant reduction in the performance of all methods. However, our methods, EyeMoSt+ (CNN) and EyeMoSt+ (Transformer), remained comparable, with EyeMoSt+ (Transformer) exhibiting the best performance. When $\sigma = 0.3$ Gaussian noise was added to the OCT modality, rendering almost all methods ineffective, our proposed EyeMoSt+ method maintained a high recognition accuracy. In the context of hypothesis testing, we computed the significance differences, as indicated by p-value, between all methods under noisy conditions and the optimal results obtained by our proposed method, as shown in Tab. 1. Based on the p-value in Tab. 1, our proposed method exhibited distinctions compared to TMC [12], EyeStar [51], and B-transformer. Notably, it demonstrated pronounced differences when compared to the remaining methods.

Furthermore, in a more general scenario, we demonstrated the ACC and ECE metrics under different noise conditions ($\sigma = 0.1, 0.2, 0.3, 0.4, 0.5$) for the Fundus or OCT modalities, as depicted in Fig. 4. As noise increased, both of our proposed methods exhibited optimal performance, highlighting the effectiveness of our fusion approach. Specifically, under noisy conditions in the Fundus modality, EyeMoSt+ (Transformer) demonstrated superior performance, while under noisy conditions in the OCT modality, EyeMoSt+ (CNN) also exhibited the best performance. This variation could be attributed to differences in the pretrained encoders. Overall, EyeMoSt+ (CNN) achieves the best performance under normal and noisy conditions.

2) OLIVES and in-house datasets: We further compared our algorithm with different methods on the OLIVES and in-house datasets in Tab. 2 and Tab. 3. We compare these methods under the clean multi-modality eye data. Our method obtained competitive results in terms of ACC and Kappa. It should be noted that on the OLIVES dataset, due to the abundance of data and easy classification, most methods can achieve a perfect accuracy rate. Then, to verify the robustness of our model, we added Gaussian noise to Fundus or OCT modality ($\sigma = 0.5/0.3$) on the dataset. We discovered that when all methods encountered Fundus modality affected by Gaussian noise, they managed to maintain their performance to a certain degree. However, once Gaussian noise was introduced to OCT modality, their performances were noticeably affected as shown in Tab. 2 and Tab. 3. Specically, when compared to early fusion B-EF [16], intermediate fusion methods B-CNN, B-Transformer and M^2 LC [50], EyeMoSt [56] demonstrates enhanced or maintained classification accuracy in both noisy OCT and Fundus modalities. Simultaneously, our proposed EyeMoSt+ method exhibits the best performance, consistently ranking first or second across various conditions. In comparison to the late fusion method TMC [12], our EyeMoSt+ demonstrates comparable performance under normal condition and superior performance in noisy Fundus or OCT modality. We also computed the significance differences between the optimal results of our proposed method and other methods under noisy conditions on the OLIVES and in-house datasets. The results of p-value in Tab. 2 and Tab. 3 indicate significant distinctions compared to other methods.

More generally, we added different Gaussian noises ($\sigma = 0.1, 0.2, 0.3, 0.4, 0.5$) to Fundus or OCT modality, as depicted in Figure 5 (a) and (b), to showcase their effects on ACC and Kappa metrics. The same conclusion can be drawn from Fig. 5 that our EyeMoSt+

demonstrates superior performance in both noisy Fundus or OCT modality. In addition, we found that noisy OCT modality exert a significantly greater influence on performance compared to Fundus modality. Based on the above experiments, we can draw a conclusion that our EyeMoSt+ remains unaffected by any noisy modality and achieves comparable performance under normal condition. This resilience can be attributed to the confidence-aware distributional fusion and the multi-modality ranking loss, which enables robust fusion under noisy modality. The visual comparisons of original and different noises to the Fundus or OCT modality on the in-house dataset can be shown in Fig. 5 (c).

Table 3: Comparisons with different algorithms on the In-house dataset. F and O denote Fundus and OCT modality. The top-2 results are highlighted in **bold** and underlined for our method. Higher ACC and Kappa mean better.

Method	Original		Gaussian noise					
			$\sigma=0.5$ (F)		$\sigma=0.3$ (O)			
	ACC	Kappa	ACC	Kappa	ACC	Kappa	P-value	Time (s)
B-CNN	0.800	0.581	0.457	0.002	0.443	0.000	$\ll 0.001$	<u>3.20</u>
B-Transformer	<u>0.814</u>	0.612	0.457	0.002	0.443	0.000	$\ll 0.001$	2.94
B-EF [16]	0.829	<u>0.643</u>	0.814	0.615	0.443	0.000	$\ll 0.001$	5.90
M^2 LC [50]	<u>0.814</u>	0.607	0.703	0.417	0.443	0.000	$\ll 0.001$	6.46
MCDO [19]	0.786	0.549	0.457	0.023	0.429	0.204	$\ll 0.001$	5.01
DE [22]	0.829	0.646	0.814	0.615	0.626	0.033	$\ll 0.001$	8.28
TMC [12]	0.829	<u>0.643</u>	0.729	0.448	0.443	0.000	$\ll 0.001$	3.98
EyeMoSt [56]	0.829	0.646	<u>0.800</u>	0.575	0.829	0.646	-	3.22
EyeMoSt+ (Transformer)	<u>0.814</u>	0.612	0.787	0.543	<u>0.671</u>	0.345	-	3.25
EyeMoSt+ (CNN)	0.829	0.641	0.814	<u>0.612</u>	0.829	<u>0.641</u>	-	3.61

3) Uncertainty estimation & and Inference time: To further quantify the reliability of uncertainty estimation, we compared different uncertainty estimation algorithms [22, 19, 12] using the ECE indicator on the GAMMA dataset. As shown in Tab. 1 and Fig. 4, our proposed algorithm shows comparable performance in clean modality and more robust performance in case of noised uni-modality. The more comparisons of ECE and AURC on the OLIVES and in-house datasets can be seen in the Fig. 5 (a-b). Similar experimental conclusions were observed from this observation. Finally, we compared the average inference time of a single test sample on different data sets for different algorithms, as shown in Tab. 1, Tab. 2 and Tab. 3. As shown these tables, our EyeMost+ has improved little running time compared to methods without uncertainty estimation, but provides more accurate and robust performance. Compared to uncertainty estimation methods, our EyeMoSt+ (CNN) exhibits faster processing speeds than other uncertainty estimation methods (MCDO [19], DE [22], and TMC [12]), although it is slightly slower than the previous version, EyeMoSt [56]. It is worth noting that EyeMoSt+ (Transformer) achieves the optimal processing speed compared to other uncertainty estimation methods, attributed to the reduced input image size. Overall, EyeMoSt+ (CNN) achieves more robust accuracy and reliable uncertainty estimation. Therefore, in the following sections, we will only present the results of EyeMoSt+ (CNN).

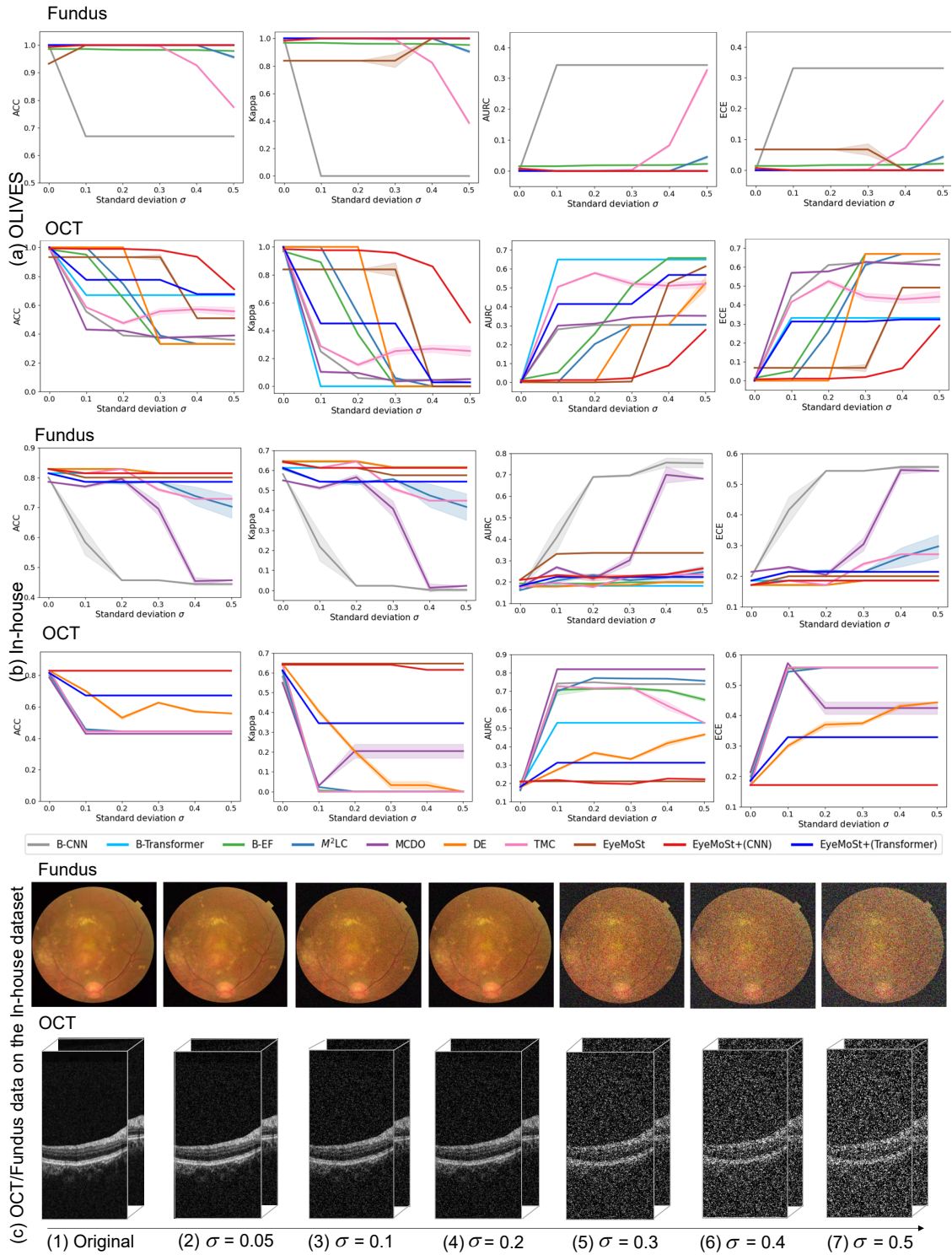


Figure 5: Results from Robust validation. (a-b) Performance metrics including ACC, Kappa, AURC, and ECE for various algorithms in the presence of noise at different levels in single modalities. Results are shown for both OLIVES and in-house datasets. (c) Comparisons of original and noisy OCT/Fundus data on the In-house dataset.

4.4. Out-of-distribution detection

According to [33], OOD data can be categorized into two main groups: shifted samples, which exhibit visual differences but semantic similarities compared to in-distribution (ID) data, and near-OOD samples, which share perceptual similarities but possess distinct semantics relative to ID data. To replicate these scenarios, we introduced noise to create shifted samples and substituted in-house fundus images with fundus images from the GAMMA dataset in our experimental setup to generate near-OOD samples. To advance our pursuit of uncertainty estimation in multi-modality ophthalmic clinical applications, we conducted uni- and multi- modality uncertainty analyses on eye data.

1) Uncertainty analysis for shifted eye data: In our first analysis, we introduced varying levels of Gaussian noise to the uni-modality data (Fundus or OCT) in both the OLIVES and in-house datasets to simulate shifted OOD data. The original samples without noise were labeled as in-distribution (ID) data. Fig. 6 (a) illustrates a significant correlation between uncertainty and OOD data. Uncertainty in uni-modality images increases proportionally with added noise. This observation underscores the role of uncertainty as a metric for assessing the reliability of uni-modality eye data. Additionally, we examined the uncertainty density of uni-modality and fusion modality before and after introducing Gaussian noise. Fig. 6 (b) provides an example by adding noise with $\sigma = 0.3$ to the Fundus modality or OCT modality on the in-house dataset. After noise introduction, fused uncertainty increases, leading to a rightward shift in the entropy distribution map. Notably, the distribution of the fusion modality aligns more closely with that of the modality without noise (Fig. 6 (b) (2)-(3)). Therefore, our proposed method can serve as a valuable tool for evaluating the reliability of modalities in ophthalmic multi-modality data fusion.

2) Confidence analysis for near OOD eye data: Furthermore, we replaced the fundus modality in the in-house dataset with fundus modality images from the previously unseen GAMMA [51] dataset to simulate near-OOD data. The fundus and OCT image inputs of the in-house dataset were considered as ID data. As indicated in Tab. 4, the performance of various methods declined to varying degrees when compared to in-distribution data, whereas our method maintained robust performance. We also analyzed the change in confidence distribution for data within and outside the fundus modality and fused modality divisions. As depicted in Fig. 7, we observed that confidence decreased for the fundus modality and the fusion modality on OOD data. However, the confidence decline of the fusion modality was less pronounced when compared to the fundus modality. Consequently, our proposed method can serve as an effective OOD detector, facilitating reliable and robust decision-making in multi-modality eye disease screening.

4.5. Missing modality experiments

In real-world clinical diagnoses, datasets containing paired fundus and OCT images are often limited. Consequently, we compared our method with other methods on the in-house dataset featuring a missing modality scenario. To simulate the absence of one modality, we set the input for the missing modality to $\mathbf{0}$. Tab. 5 shows that even when the Fundus modality is missing, most algorithms, including ours, maintain a certain level of performance. Conversely, in the absence of the OCT modality, most algorithms experience a notable

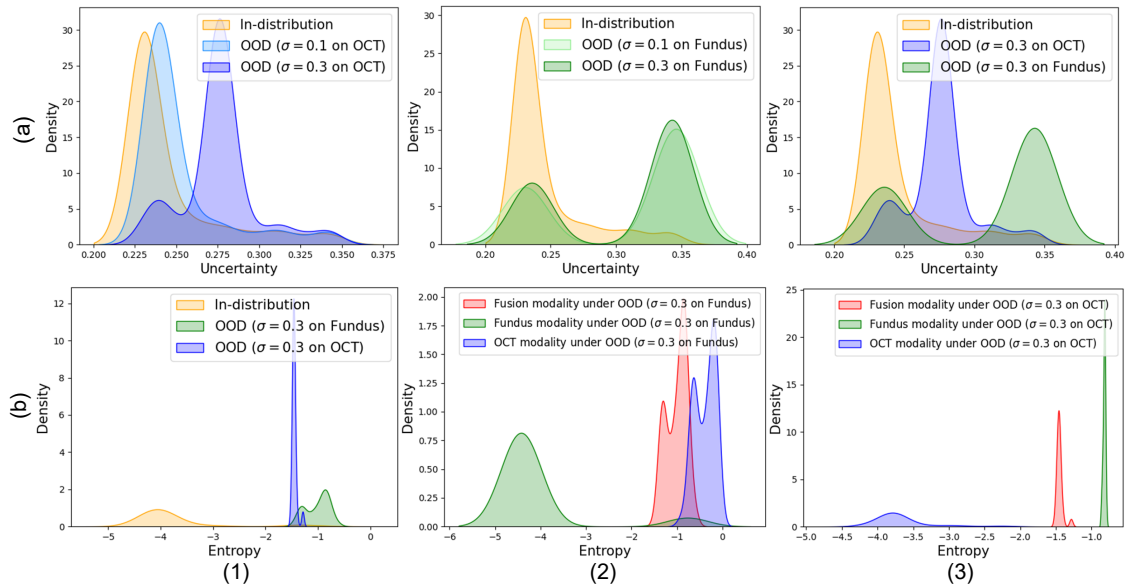


Figure 6: (a) Uncertainty density of uni-modality eye data on the OLIVES dataset. (1-2) ID and OOD under various levels of noise in OCT/Fundus data. (3) ID and OOD under different noise levels in either OCT or Fundus data. (b) Uncertainty density of uni-modality and multi-modality eye data on the in-house dataset. (1) ID and OOD under a noise level of $\sigma = 0.3$ in either OCT or Fundus data. (2-3) Uncertainty density for uni-modality and fusion modality under OOD data (noisy Fundus or OCT modality).

decline in performance. Notably, our proposed algorithm exhibits sustained performance. This observation suggests a tendency among most algorithms to over-rely on a specific modality, such as OCT modality, in the absence of other modalities, leading to a natural decline in performance. Our proposed algorithm dynamically incorporates more dependable modalities by evaluating the confidence and reliability of each modality. By doing so, our method addresses the limitations observed in other algorithms and demonstrates robust performance in scenarios with missing modalities.

Table 4: Accuracy, Kappa, and ECE performance of different algorithms under adding unseen Fundus modality. (BLUE) means indicates the performance of near OOD eye data.

Methods	Metrics		
	ACC \uparrow	Kappa \uparrow	ECE \downarrow
B-CNN	0.629 (0.171)	0.271 (0.310)	0.371 (0.171)
M^2 LC [50]	0.743 (0.071)	0.472 (0.135)	0.257 (0.071)
MCDO [19]	0.676 (0.110)	0.357 (0.192)	0.324 (0.110)
DE [22]	0.771 (0.058)	0.537 (0.109)	0.229 (0.058)
TMC [12]	0.786 (0.043)	0.559 (0.084)	0.214 (0.043)
EyeMoSt [56]	0.771 (0.058)	0.521 (0.125)	0.186 (0.015)
EyeMoSt+ (CNN)	0.814 (0.015)	0.607 (0.034)	0.186 (0.015)

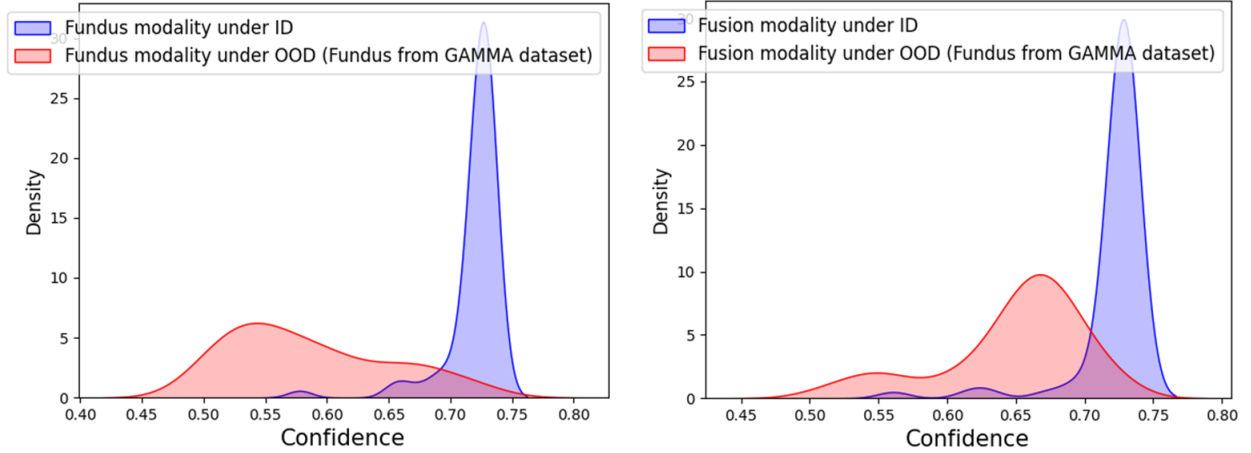


Figure 7: Confidence density of uni-modality and multi-modality eye data on the near-OOD condition.

Table 5: Accuracy, Kappa, and ECE performance of different algorithms under missing Fundus or OCT modality condition on the in-house dataset.

Methods	Missing Fundus modality			Missing OCT modality		
	ACC \uparrow	Kappa \uparrow	ECE \downarrow	ACC \uparrow	Kappa \uparrow	ECE \downarrow
B-CNN	0.443	0.000	0.557	0.443	0.000	0.557
M^2 LC [50]	0.443	0.000	0.557	0.443	0.000	0.557
MCDO [19]	0.794	0.574	0.206	0.443	0.000	0.557
DE [22]	0.786	0.543	0.214	0.443	0.000	0.557
TMC [12]	0.743	0.479	0.257	0.443	0.000	0.557
EyeMoSt [56]	0.800	0.584	0.200	0.669	0.000	0.331
EyeMoSt+ (CNN)	0.814	0.612	0.186	0.729	0.448	0.271

Table 6: Parameter selection of λ_F on the in-house dataset.

$\lambda_F =$	0	0.1	0.2	0.5	0.7	1.0
Acc \uparrow	0.800	0.771	0.814	0.829	0.814	0.786
Kappa \uparrow	0.575	0.512	0.606	0.641	0.615	0.543
AURC \downarrow	0.331	0.269	0.263	0.210	0.251	0.336
ECE \downarrow	0.200	0.230	0.186	0.171	0.186	0.214

Table 7: Parameter selection of λ_C on the in-house dataset. (·) denote the Fundus condition with added noise ($\sigma=0.3$).

$\lambda_C =$	0	0.1	0.5	1	5	10	15
Acc \uparrow	0.829 (0.800)	0.843 (0.557)	0.800 (0.800)	0.814 (0.786)	0.814 (0.786)	0.829 (0.814)	0.800 (0.786)
Kappa \uparrow	0.641 (0.575)	0.667 (0.000)	0.578 (0.575)	0.607 (0.543)	0.607 (0.543)	0.641 (0.612)	0.581 (0.543)
AURC \downarrow	0.210 (0.336)	0.269 (0.238)	0.185 (0.208)	0.212 (0.239)	0.233 (0.269)	0.209 (0.228)	0.198 (0.192)
ECE \downarrow	0.171 (0.200)	0.230 (0.443)	0.200 (0.200)	0.186 (0.214)	0.186 (0.214)	0.171 (0.186)	0.200 (0.214)

Table 8: Ablation study for overall learning process on the in-house dataset. (·) denote the Fundus condition with added noise ($\sigma=0.3$).

B	\mathcal{L}_m^{NIG}	\mathcal{L}_F^{St}	\mathcal{L}_C	Acc \uparrow	Kappa \uparrow
✓				0.800	0.581
✓	✓			0.814	0.612
✓	✓	✓		0.829 (0.800)	0.646 (0.575)
✓	✓	✓	✓	0.829 (0.814)	0.641 (0.612)

4.6. Ablation study

1) Hyperparameter selection of λ_F and λ_C : λ_F is the balance factor between the \mathcal{L}_m^{NLL} loss and the \mathcal{L}_m^{CE} loss. In the experiments below, we demonstrate the importance of augmenting training objective with the evidence classifier loss \mathcal{L}_m^{CE} introduced in EyeMoSt. $\lambda_F \in [0, 1]$ represents the importance of \mathcal{L}_m^{CE} loss. We performed parameter validation on the in-house dataset. As shown in the Tab. 6, the performance is improved after introducing \mathcal{L}_m^{CE} loss, and the best value is 0.5. λ_C represents a pivotal hyperparameter governing the regularization of confidence-aware multimodal learning. We conducted parameter selection experiments on the in-house dataset. In alignment with [29], we explored the range $\lambda_C = 0.1$ to 15 to assess its performance. Additionally, to underscore the robustness of this regularization term, we added Gaussian noise ($\sigma=0.3$) to the Fundus modality. As depicted in Tab. 7, the optimal value for λ_C was determined to be 10.

2) Overall learning process: Further, we conduct ablation experiments on Eq. 13, as depicted in Tab. 8. Where B is the baseline of the intermediate typical fusion method B-CNN. B-CNN first extracts features by the encoders (same with us), and then concatenates their output features as the final prediction. \mathcal{L}_m^{NIG} represents pairwise fusion directly after establishing multi-NIG distributions.

3) Uni-modality and multi-modality: Finally, we conducted a comparative analysis between the uni-modal variant of B-CNN and our proposed method on the in-house dataset. Specifically, we examined various uni-modality scenarios, denoted as Uni-B, where only the Fundus modality was used for training. The results, as presented in Tab. 9, reveal that the base method B-CNN can initially achieve performance levels comparable to those of the individual uni-modality Uni-B methods after fusion. However, it becomes susceptible to performance degradation when exposed to noise, occasionally even underperforming the uni-modality methods. In contrast, our proposed method EyeMoSt+ (CNN), described in this paper, incorporates a confidence-based distribution during the fusion process.

5. Conclusion

In conclusion, we introduce EyeMoSt+, a pioneering solution designed to revolutionize multi-modality eye disease screening by seamlessly fusing Fundus and OCT modalities. Drawing upon the principles of NIG prior distributions, we have harnessed aleatoric and epistemic uncertainty embedded in uni-modality data. More importantly, a confidence-aware

Table 9: Comparisons with uni-modality and multi-modality methods on the in-house dataset. Uni-Fundus and Uni-OCT represent the classification of eye diseases using the B-CNN method with only Fundus or OCT, respectively. EyeMoSt+ denotes the EyeMoSt+ (CNN).

Method	Original		Gaussian noise			
			$\sigma=0.3$ (F)		$\sigma=0.5$ (O)	
	ACC	Kappa	ACC	Kappa	ACC	Kappa
Uni-Fundus	0.800	0.581	0.557	0.000	/	/
Uni-OCT	0.786	0.543	/	/	0.557	0.000
B-CNN	0.800	0.581	0.457	0.023	0.443	0.000
EyeMoSt+	0.829	0.641	0.814	0.612	0.829	0.641

fusion for mixture of Student’s t distribution is proposed to establish a robust and reliable disease screening model. Furthermore, our innovative confidence-aware ranking-based regularization form offers a new perspective on fusion integrity, preventing the compromise of outcomes in the presence of noisy modality. Through rigorous validation across a diverse spectrum of eye disease datasets, including Glaucoma recognition, AMD and PCV screening, as well as DR and DME recognition, our method’s reliability and robustness are firmly established. Particularly notable is its effectiveness in handling noisy inputs, identifying missing patterns, and processing unseen data.

In the future, our attention will be directed along two distinct avenues. Firstly, we seek to expand beyond pairwise modality fusion, delving into the realm of comprehensive multimodal fusion. Secondly, we are committed to exploring the real-world implementation of our robust ocular a framework for multimodal screening of eye disease. This strategic initiative holds significant promise in advancing the accuracy and dependability of AI-driven medical decisions, a prospect that resonates strongly with our overarching objectives.

Acknowledgment

This work was supported in part by the Science and Technology Department of Sichuan Province (Grant No. 2022YFS0071 & 2023YFG0273), the China Scholarship Council (No. 202206240082), the National Research Foundation, National Natural Science Foundation of China (Grant No.62376193), the H. Fu’s Agency for Science, Technology and Research (A*STAR) Central Research Fund (“Robust and Trustworthy AI system for Multi-modality Healthcare”), the A*STAR Advanced Manufacturing and Engineering (AME) Programmatic Fund (A20H4b0141), the National Key R&D Program of China (Grant No. 2018YFA0701700), Shantou Science and Technology Program (Grant No. 200629165261641), and 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (Grant No. 2020LKSFG14B).

References

- [1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques,

- applications and challenges. *Information Fusion*, 2021. <https://doi.org/10.1016/j.inffus.2021.05.008>.
- [2] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. <https://dl.acm.org/doi/10.5555/3495724.3496975>.
 - [3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. <https://doi.org/10.1109/TPAMI.2018.2798607>.
 - [4] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. <https://dl.acm.org/doi/10.5555/3045118.3045290>.
 - [5] Z. Cai, L. Lin, H. He, and X. Tang. Corolla: an efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022. <https://doi.org/10.1109/ISBI52829.2022.9761712>.
 - [6] Z. Cai, L. Lin, H. He, and X. Tang. Uni4eye: Unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 88–98. Springer, 2022. https://doi.org/10.1007/978-3-031-16452-1_9.
 - [7] S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. <https://doi.org/10.48550/arXiv.1904.00625>.
 - [8] C. M. G. Cheung, T. Y. Lai, P. Ruamviboonsuk, S.-J. Chen, Y. Chen, K. B. Freund, F. Gomi, A. H. Koh, W.-K. Lee, and T. Y. Wong. Polypoidal choroidal vasculopathy: definition, pathogenesis, diagnosis, and management. *Ophthalmology*, 125(5):708–724, 2018. <https://doi.org/10.1016/j.ophtha.2017.11.019>.
 - [9] M. El Habib Daho, Y. Li, R. Zeghlache, Y. C. Atse, H. Le Boité, S. Bonnin, D. Cosette, P. Deman, L. Borderie, C. Lepicard, et al. Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of uwf-cfp and octa images. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 11–20. Springer, 2023. https://doi.org/10.1007/978-3-031-44013-7_2.
 - [10] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. <https://doi.org/10.1109/TPAMI.2019.2938758>.
 - [11] D. Gerogiannis, C. Nikou, and A. Likas. Robust image registration using mixtures of t-distributions. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. <https://doi.org/10.1109/ICCV.2007.4409127>.
 - [12] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022. <https://doi.org/10.1109/TPAMI.2022.3171983>.
 - [13] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. <https://dl.acm.org/doi/10.1109/WACV51458.2022.00181>.
 - [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. <https://doi.org/10.1109/CVPR.2016.90>.
 - [15] X. He, Y. Deng, L. Fang, and Q. Peng. Multi-modal retinal image classification with modality-specific attention network. *IEEE transactions on medical imaging*, 40(6):1591–1602, 2021. <https://doi.org/10.1109/TMI.2021.3059956>.
 - [16] C.-H. Hua, K. Kim, T. Huynh-The, J. I. You, S.-Y. Yu, T. Le-Tien, S.-H. Bae, and S. Lee. Convolutional network with twofold feature augmentation for diabetic retinopathy recognition from multi-modal images. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2686–2697, 2020. <https://doi.org/10.1109/JBHI.2020.3041848>.
 - [17] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference

- for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020. <https://doi.org/10.48550/arXiv.1907.07504>.
- [18] M. C. Jung, H. Zhao, J. F. Dipnall, B. J. Gabbe, and L. Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. In *Neural Information Processing Systems*, 2023. <https://api.semanticscholar.org/CorpusID:257921153>.
- [19] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5580–5590, 2017. <https://dl.acm.org/doi/10.5555/3295222.3295309>.
- [20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. <https://doi.org/10.48550/arXiv.1412.6980>.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. <https://doi.org/10.1145/3065386>.
- [22] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. <https://dl.acm.org/doi/10.5555/3295222.3295387>.
- [23] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng. Canet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, 39(5):1483–1493, 2019. <http://dx.doi.org/10.1109/TMI.2019.2951844>.
- [24] X. Li, M. Jia, M. T. Islam, L. Yu, and L. Xing. Self-supervised feature learning via exploiting multimodal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033, 2020. <https://doi.org/10.1109/TMI.2020.3008871>.
- [25] Y. Li, M. El Habib Daho, P.-H. Conze, H. Al Hajj, S. Bonnin, H. Ren, N. Manivannan, S. Magazzeni, R. Tadayoni, B. Cochener, et al. Multimodal information fusion for glaucoma and diabetic retinopathy classification. In *Ophthalmic Medical Image Analysis: 9th International Workshop, OMIA 2022, Held in conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings*, pages 53–62. Springer, 2022. https://doi.org/10.1007/978-3-031-16525-2_6.
- [26] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong. Age-related macular degeneration. *The Lancet*, 379(9827):1728–1738, 2012. [https://doi.org/10.1016/S0140-6736\(22\)02609-5](https://doi.org/10.1016/S0140-6736(22)02609-5).
- [27] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512, 2020. <https://dl.acm.org/doi/10.5555/3495724.3496353>.
- [28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. <https://dl.acm.org/doi/10.1109/ICCV48922.2021.00986>.
- [29] H. Ma, Q. Zhang, C. Zhang, B. Wu, H. Fu, J. T. Zhou, and Q. Hu. Calibrating multimodal learning. In *International Conference on Machine Learning*, pages 23429–23450. PMLR, 2023. <https://doi.org/10.48550/arXiv.2306.01265>.
- [30] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. <https://doi.org/10.1162/neco.1992.4.3.448>.
- [31] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. <https://dl.acm.org/doi/10.5555/3327757.3327808>.
- [32] J. Maronas, R. Paredes, and D. Ramos. Calibration of deep probabilistic models with decoupled bayesian neural networks. *Neurocomputing*, 407:194–205, 2020. <https://doi.org/10.1016/j.neucom.2020.04.103>.
- [33] J. Mukhoti, T.-Y. Lin, B.-C. Chen, A. Shah, P. H. Torr, P. K. Dokania, and S.-N. Lim. Raising the bar on the evaluation of out-of-distribution detection. *arXiv preprint arXiv:2209.11960*, 2022. <https://doi.org/10.48550/arXiv.2209.11960>.
- [34] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. <https://doi.org/10.1007/978-1-4612-0745-0>.
- [35] T. M. Nguyen and Q. J. Wu. Robust student’s-t mixture model with spatial constraints and its

- application in medical image segmentation. *IEEE Transactions on Medical Imaging*, 31(1):103–116, 2011. <https://doi.org/10.1109/TMI.2011.2165342>.
- [36] Z. Ou, W. Chai, L. Wang, R. Zhang, J. He, M. Song, L. Yuan, S. Zhang, Y. Wang, H. Li, et al. m^2lc -net: A multi-modal multi-disease long-tailed classification network for real clinical scenes. *China Communications*, 18(9):210–220, 2021. <https://doi.org/10.23919/JCC.2021.09.016>.
- [37] M. Prabhushankar, K. Kokilepersaud, Y.-y. Logan, S. Trejo Corona, G. AlRegib, and C. Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022. <https://doi.org/10.48550/arXiv.2209.11195>.
- [38] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014. <https://doi.org/10.48550/arXiv.1401.0118>.
- [39] N. Ravikumar, A. Gooya, S. Çimen, A. F. Frangi, and Z. A. Taylor. Group-wise similarity registration of point sets using student’s t-mixture model for statistical shape models. *Medical image analysis*, 44:156–176, 2018. <https://doi.org/10.1016/j.media.2017.11.012>.
- [40] E. O. Rodrigues, A. Conci, and P. Liatsis. Element: Multi-modal retinal vessel segmentation based on a coupled region growing and machine learning approach. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3507–3519, 2020. <https://doi.org/10.1109/JBHI.2020.2999257>.
- [41] M. Roth, E. Özkan, and F. Gustafsson. A student’s t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5770–5774. IEEE, 2013.
- [42] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3183–3193, 2018. <https://dl.acm.org/doi/10.5555/3327144.3327239>.
- [43] J. Soni and R. Goodman. *A mind at play: how Claude Shannon invented the information age*. Simon and Schuster, 2017.
- [44] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020. <https://dl.acm.org/doi/10.5555/3524938.3525836>.
- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. <https://dl.acm.org/doi/10.5555/3295222.3295349>.
- [46] H. Wang, J. Zhang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Uncertainty-aware multi-modal learning via cross-modal random network prediction. In *European Conference on Computer Vision*, pages 200–217. Springer, 2022. https://doi.org/10.1007/978-3-031-19836-6_12.
- [47] M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian, et al. Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications*, 14(1):6757, 2023. <https://doi.org/10.1038/s41467-023-42444-7>.
- [48] W. Wang, X. Li, Z. Xu, W. Yu, J. Zhao, D. Ding, and Y. Chen. Learning two-stream cnn for multi-modal age-related macular degeneration categorization. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4111–4122, 2022. <https://doi.org/10.1109/JBHI.2022.3171523>.
- [49] W. Wang, Z. Xu, W. Yu, J. Zhao, J. Yang, F. He, Z. Yang, D. Chen, D. Ding, Y. Chen, et al. Two-stream cnn with loose pair training for multi-modal amd categorization. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 156–164. Springer, 2019. https://doi.org/10.1007/978-3-030-32239-7_18.
- [50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. https://doi.org/10.1007/978-3-030-01234-2_1.
- [51] J. Wu, H. Fang, F. Li, H. Fu, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023. <https://doi.org/10.1016/j.media.2023.102938>.
- [52] T. K. Yoo, J. Y. Choi, J. G. Seo, B. Ramasubramanian, S. Selvaperumal, and D. W. Kim. The possibility of the combination of oct and fundus images for improving the diagnostic accuracy of deep learning

- for age-related macular degeneration: a preliminary experiment. *Medical & biological engineering & computing*, 57:677–687, 2019. <https://doi.org/10.1007/s11517-018-1915-z>.
- [53] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769, 2023. <https://doi.org/abs/2306.02050>.
- [54] T. Zhou, S. Ruan, and S. Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019. <https://doi.org/10.1016/j.array.2019.100004>.
- [55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. <http://dx.doi.org/10.1109/ICCV.2017.244>.
- [56] K. Zou, T. Lin, X. Yuan, H. Chen, X. Shen, M. Wang, and H. Fu. Reliable multimodality eye disease screening via mixture of student’s t distributions. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 596–606, 2023. https://doi.org/10.1007/978-3-031-43990-2_56.