AMERICAN ACADEMY OF OPHTHALMOLOGY®

Advancing Glaucoma Diagnosis: Employing Confidence-Calibrated Label Smoothing Loss for Model Calibration

Midhula Vijayan, PhD, Deepthi Keshav Prasad, PhD, Venkatakrishnan Srinivasan, MTech

Objective: The aim of our research is to enhance the calibration of machine learning models for glaucoma classification through a specialized loss function named Confidence-Calibrated Label Smoothing (CC-LS) loss. This approach is specifically designed to refine model calibration without compromising accuracy by integrating label smoothing and confidence penalty techniques, tailored to the specifics of glaucoma detection.

Design: This study focuses on the development and evaluation of a calibrated deep learning model.

Participants: The study employs fundus images from both external datasets—the Online Retinal Fundus Image Database for Glaucoma Analysis and Research (482 normal, 168 glaucoma) and the Retinal Fundus Glaucoma Challenge (720 normal, 80 glaucoma)—and an extensive internal dataset (4639 images per category), aiming to bolster the model's generalizability. The model's clinical performance is validated using a comprehensive test set (47 913 normal, 1629 glaucoma) from the internal dataset.

Methods: The CC-LS loss function seamlessly integrates label smoothing, which tempers extreme predictions to avoid overfitting, with confidence-based penalties. These penalties deter the model from expressing undue confidence in incorrect classifications. Our study aims at training models using the CC-LS and comparing their performance with those trained using conventional loss functions.

Main Outcome Measures: The model's precision is evaluated using metrics like the Brier score, sensitivity, specificity, and the false positive rate, alongside qualitative heatmap analyses for a holistic accuracy assessment.

Results: Preliminary findings reveal that models employing the CC-LS mechanism exhibit superior calibration metrics, as evidenced by a Brier score of 0.098, along with notable accuracy measures: sensitivity of 81%, specificity of 80%, and weighted accuracy of 80%. Importantly, these enhancements in calibration are achieved without sacrificing classification accuracy.

Conclusions: The CC-LS loss function presents a significant advancement in the pursuit of deploying machine learning models for glaucoma diagnosis. By improving calibration, the CC-LS ensures that clinicians can interpret and trust the predictive probabilities, making artificial intelligence-driven diagnostic tools more clinically viable. From a clinical standpoint, this heightened trust and interpretability can potentially lead to more timely and appropriate interventions, thereby optimizing patient outcomes and safety.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100555*

Glaucoma, often described as the "silent thief of sight," stands as a predominant cause of irreversible vision impairment across the globe. As of now, a staggering 80 million individuals grapple with this condition. Disturbingly, projections suggest that by 2040, this number could surge to an estimated 111 million,[1] underscoring the escalating clinical and public health challenges associated with glaucoma.

In the realm of ophthalmology, early and precise detection of glaucoma remains pivotal. Traditional diagnostic methods, encompassing intraocular pressure measurements, visual field testing, and optic nerve head evaluation, offer valuable insights. However, they come with inherent limitations-subjectivity in interpretation, variability in measurements, and sometimes late-stage detection.

Enter the realm of machine learning in medical diagnostics—a promise of a paradigm shift. In ophthalmic imaging, sophisticated algorithms can dissect intricate patterns in retinal photographs or OCT scans, with some models even rivaling human experts in their diagnostic accuracy. However, the journey of machine learning in clinical diagnostics doesn't culminate at accuracy. The actual clinical setting demands something more nuanced: calibration.

In the specialized field of medical image classification, neural network-based models are gaining significant traction as vital tools for diagnostic evaluations.[2] Given the life-critical nature of medical diagnoses, it becomes imperative that these models go beyond mere predictive accuracy to exhibit a robust degree of calibration in their outputs. Calibration,[3] in this intricate context, signifies the model's

https://doi.org/10.1016/j.xops.2024.100555
ISSN 2666-9145/24

capability to align its outputted probability scores with the actual confidence or uncertainty inherent in its predictions. For example, diagnostic images that elicit low-confidence scores from the model are optimally routed for subsequent examination by medical professionals.[4]

While much of the recent scholarly and engineering endeavors have concentrated on enhancing the precision of predictive models, the issue of model calibration has somewhat fallen by the wayside. Despite advancements in predictive accuracy, a growing body of literature indicates that the calibration quality of contemporary neural networks often leaves much to be desired.[5] For instance, multiple studies have pointed out that as these models grow larger and become more precise, they often suffer from calibration issues.[6]

Recognizing the pressing need for calibrated models in glaucoma classification, our study introduces the "Confidence-Calibrated Label Smoothing" (CC-LS) loss function. Traditional loss functions, while optimizing for accuracy, can inadvertently produce models exuding excessive confidence in their predictions, a trait particularly risky in medical settings. The CC-LS loss function is innovatively crafted to mitigate this. It melds label smoothing, which curbs extreme predictions to ensure robustness against overfitting, with confidence-based penalties that penalize undue certainty, especially in erroneous predictions.

By championing both accuracy and calibration, the proposed loss function aspires to redefine standards for training machine learning models in medical diagnostics, particularly for glaucoma classification. Through this paper, we delve deep into its mechanics, and potential impact in this specialized domain.

The main contributions of the work are listed below:

- **Domain-Specific Calibration Enhancements:** The approach has been tailored to meet the specific challenges of the domain, introducing a unique integration of label smoothing loss and confidence penalty techniques. Optimized specifically for the application, this combination not only addresses the unique characteristics and challenges but also significantly enhances model performance and reliability. The methodology demonstrates notable improvements over traditional calibration methods, effectively boosting both calibration accuracy and overall performance.
- **Empirical Validation:** Presenting extensive empirical evidence showcasing that our method not only improves calibration but also maintains or enhances model performance across various metrics. This empirical validation underlines the practical benefits and novelty of our approach in clinical settings.
- **Theoretical Contributions:** Offering theoretical insights into the impact of label smoothing and confidence penalty on model calibration and generalization. These contributions advance the understanding of how these techniques can be effectively applied in concert with binary cross-entropy (BCE) loss,[7] providing a foundation for future research in the field.

Literature Review

In recent times, image processing technologies have become increasingly integral to medical diagnostic procedures, particularly in the field of ophthalmology.[1] Retinal imaging serves as a crucial tool for assessing the health of the visual system. Automated systems not only offer a cost-effective alternative for large-scale screenings but also significantly reduce the likelihood of human error. Moreover, such systems have the potential to facilitate eye care in remote or rural locations where specialized medical professionals are scarce. This ensures that a greater number of patients can be diagnosed in a timely manner. There has been sustained research focus over the past few decades aimed at developing these automated methodologies.

In this literature survey, we delve into 2 distinct but interrelated streams of research within the domain of ophthalmic diagnostics. The first stream focuses on the advancements in glaucoma classification, detailing the cutting-edge deep learning models that have significantly improved the accuracy and efficiency of glaucoma detection.[8] The second stream centers around calibration models, investigating how well these deep learning models not only classify but also accurately represent the confidence level of their predictions. Both streams are integral to creating a holistic automated system for effective and reliable glaucoma diagnosis.

Advances in Deep Learning Models for Glaucoma Detection

Artificial intelligence (AI)-based methods for glaucoma identification can be categorized into 2 paradigms[9]: the monolithic approach and the sequential approach. In the monolithic framework, also known as an end-to-end or 1-step method, glaucoma is directly identified through intricate deep learning architectures that operate as "black-box" models. In contrast, the sequential or 2-step methodology initially employs AI algorithms to delineate the optic disc (OD) and cup contours, subsequently leveraging this anatomical information to derive automated diagnostic rules for glaucoma detection. For the scope of this literature review, our focus is confined to the monolithic approaches, as our own work adopts this 1-step strategy.

The field of glaucoma detection has benefitted immensely from advancements in deep learning techniques, as illustrated by several key studies. Researchers in[10] utilized deep learning models on retinal fundus images from the Online Retinal Fundus Image Database for Glaucoma Analysis and Research, Retinal Image Database for Optic Nerve Evaluation, and Retinal Image Dataset for Optic Nerve Head Segmentation databases. They integrated the results from various architectures to achieve an area under the curve of 94%, demonstrating the method's efficacy for glaucoma detection. Building on similar themes of effectiveness, the team in[11] employed a binary classification deep learning algorithm, optimized for a mixed dataset consisting of 5716 images from both Asian

and White populations, and reported an area under the curve of 94%.

Expanding the possibilities of diagnostic sensitivity, the work in[12] proposed a deep ensemble network with an attention mechanism. They utilized stereo images from Tan Tock Seng Hospital in Singapore to achieve a high sensitivity rate of 95.48%. On a different note, the authors in[13] customized a 3-dimensional convolutional neural network that could function efficiently even without data augmentation, showcasing its potential in practical clinical applications.

Taking a specialized approach, Aamir et al in[14] developed a multilevel deep convolutional neural network for diagnosing glaucoma in retinal fundus images. This dual-phase model excels in both initial detection and subsequent categorization of the disease's severity, achieving high sensitivity and accuracy rates.

Broadening the scope of feature extraction, Liao et al in[15] introduced multi-layers average pooling, a methodology that aggregates features from multiple scales to enhance diagnosis. This technique creates glaucoma activations, linking global diagnostic data with precise spatial localization. Similarly focused on feature extraction, Nawaz et al in[16] applied the EfficientNet-B0 architecture to extract deep features, which were further processed for precise glaucoma localization.

In terms of fine-tuning existing models, Diaz-Pinto et al in[17] modified well-known architectures like VGG16, VGG19, and others, specifically observing the performance metrics of the Xception model. Finally, the study in[18] introduces the Classification of Glaucoma Network, a highly accurate algorithm for diagnosing glaucoma, with an accuracy rate of 93.5% and an area under the curve of 0.99.

Together, these studies demonstrate the versatility and robustness of deep learning models in the realm of glaucoma detection and classification. While they offer various strategies and architectures to address this critical medical issue, it is worth noting that the focus remains largely on achieving high accuracy, with less emphasis on model calibration.

Model Calibration in Medical Image Classification

The landscape of medical imaging analysis has undergone a significant transformation owing to groundbreaking strides in deep learning technologies.[19,20] A plethora of sophisticated deep learning frameworks have emerged, elevating the capabilities of this domain. While the academic community primarily emphasizes enhancing the precision of convolutional neural networks, the equally vital aspect of uncertainty quantification often remains unfathomed. In medical contexts, where automated decision-making systems are increasingly prevalent, the accurate gauging of a model's uncertainty is indispensable. Failing to address this can yield imprecise confidence or probability metrics, consequently posing the risk of considerable diagnostic inaccuracies.[21]

Recent studies have examined how image classifiers perform when the data changes, also known as robustness.[3] In addition, some research has focused specifically on calibration. A notable example is a study by Guo et al,[6] which revealed that modern neural networks are generally not well-calibrated, especially when they are larger in size. They also observed that even as these networks get better at classifying images, their calibration tends to get worse. These observations have been supported by other research as well.[22,23]

In the realm of medical imaging, researchers are actively exploring various techniques for model calibration to enhance predictive accuracy and reliability. For instance, a study by Carneiro et al[24] employed the monoparametric variant of Platt scaling to improve probability calibration in a multiclass polyp categorization task. Their calibrated model demonstrated reductions in both expected calibration error and maximum calibration error, thereby boosting interpretability. Building on this theme, researchers in study[25] adopted a similar Platt scaling method for calibrating probabilities in renal biopsy image classification. Despite achieving lower expected calibration error values, they observed that this calibration adversely affected the model's overall accuracy. Expanding the scope to different imaging modalities, the study in[26] took a comprehensive approach, systematically examining the impact of calibration on performance metrics across chest x-rays and fundus photographs. Various deep learning classifiers were employed to assess the effects, making it a multifaceted inquiry into the role of calibration in medical imaging.

In the quest for achieving not only high accuracy but also reliable uncertainty quantification in deep learning models, recent advancements such as label smoothing and confidence penalty mechanisms have emerged as pivotal strategies. Label smoothing, a technique aimed at preventing the model from becoming overconfident in its predictions, has shown promise in enhancing the generalization of models by softening the targets during training.[27,28] This method mitigates the issue of overfitting to the hard labels and encourages the model to be more robust to input variations. On the contrary, the confidence penalty approach directly addresses the calibration of neural networks by penalizing overly confident predictions.[29,30] This regularization strategy discourages the model from assigning extreme probabilities to its predictions, thereby promoting a more calibrated and interpretable output. Both label smoothing and confidence penalty are especially relevant in the medical imaging domain, where the cost of misinterpretation can be high, and the demand for calibrated probability estimates is critical. By incorporating these methods, deep learning models can achieve a delicate balance between accuracy and reliability, ensuring that the predictions are not only precise but also reflect true confidence levels. The integration of these approaches into glaucoma classification models could potentially bridge the current gap in research, offering a novel perspective on how to achieve calibrated and trustworthy predictions in ophthalmic diagnostics.

In the existing literature, it is noteworthy that there appears to be a gap with respect to calibration studies

specifically focused on glaucoma classification. Despite the plethora of research on model calibration in various medical imaging contexts, glaucoma diagnosis remains an area yet to be explored in this regard. To fill this research void, the present study introduces a calibrated model tailored for glaucoma classification, employing a CC-LS function to enhance both predictive accuracy and reliability. This work aims to contribute a calibrated approach to glaucoma detection, offering a new dimension to the ongoing discourse in the field.

Approach

Significance of Calibration in Medical Image Classification and the Imperative for Innovative Loss Functions

Calibration pertains to the alignment between a model's asserted confidence in its predictions and the actual likelihood of those predictions being correct. In medical image classification, where decisions can influence clinical outcomes and patient care trajectories, this alignment is especially paramount.

Why Calibration Matters:

1. **Patient Safety:** Inaccurate confidence estimations could lead to overdiagnosis or missed diagnoses. For instance, a high-confidence misclassification might bypass further necessary investigations, potentially compromising patient safety.
2. **Clinician Trust:** For physicians to trust and integrate AI tools into their workflow, they need more than just an accurate prediction. They require an understanding of how certain the model is about its decision, enabling them to make informed clinical judgments.
3. **Resource Allocation:** Health care resources, both in terms of time and finance, are precious. A calibrated model ensures that interventions and further diagnostics are reserved for patients who truly need them, promoting efficient resource allocation.

The Demand for Customized Loss Functions:

Traditional loss functions in deep learning, such as BCE loss,[7] focus predominantly on optimizing model accuracy. However, these conventional loss functions may inadvertently overlook the model's calibration.

1. **Bridging the Calibration-Accuracy Gap:** While many high-accuracy models exist, they sometimes suffer from miscalibration, revealing a disconnection between accuracy and calibration. Novel loss functions can be designed to bridge this gap, ensuring models are both accurate and well-calibrated.
2. **Direct Calibration during Training:** Instead of post hoc calibration methods, which attempt to recalibrate models after training, a specialized loss function, like the one introduced in this study, directly incorporates calibration during the training phase. This preemptive approach ensures the model learns to make calibrated predictions from the outset.
3. **Inherent Model Robustness:** A calibrated model tends to generalize better to unseen data,[31] especially in cases where there might be slight deviations from the training distribution. This robustness is crucial in medical imaging, where patient data can vary widely.

Decisively, in the evolving domain of medical imaging powered by deep learning, the need for calibration is not just a technical requirement but a clinical imperative. As we advance our methods and tools, introducing novel loss functions that prioritize calibration will be pivotal in shaping a future where AI-assisted diagnostics seamlessly merge with the broader tapestry of patient care.

Technical Description of the CC-LS

The CC-LS loss function is formulated to counteract model overconfidence while ensuring calibrated and generalized predictions. It comprises 2 core components: label smoothing and confidence-based penalties.

Label Smoothing Loss: This component prevents extreme confidence in predictions by adjusting the hard 0 and 1 labels of the binary classification task. The adjusted targets are computed as:

$$t = (1.0 - s) \times L + s \times (1 - L) \quad (1)$$

In this equation, t represents the new or smoothed target label, designed to mitigate the model's overconfidence in its predictions and improve generalization. The term s serves as the smoothing factor, playing a crucial role in adjusting the original labels. The s values ranging from 0.01 to 0.1 were explored to identify the most effective degree of label smoothing. This range was selected based on preliminary tests aimed at enhancing the model's robustness without compromising prediction accuracy. The value of $s = 0.05$ was ultimately chosen, as it optimally reduced overconfidence and improved the model's generalization capabilities. On the other hand, L symbolizes the original, or ground truth, labels, which could be 0 or 1 in the case of binary classification. The equation operates by attenuating the original label L through the factor $1.0 - s$ and introducing an additive term that nudges the label toward its opposite class, scaled by s. This subtle modification to the labels prevents the model from becoming overly sure of its predictions, thereby fostering better adaptability to unseen data.

After obtaining the smoothed labels, the Label Smoothing Loss L is calculated as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[t_i \log\left(\frac{1}{1+\exp(-o_i)}\right) + (1-t_i)\log\left(1-\frac{1}{1+\exp(-o_i)}\right)\right] \quad (2)$$

Here is a breakdown of the terms in the equation:

- L is the Label Smoothing Loss.
- N is the number of samples.
- t_i is the target label after label smoothing for the 'i^{th}' sample.
- o_i is the output from the neural network for the 'i^{th}' sample.

Confidence-Based Penalty Mechanism: The essence of the CC-LS loss is to penalize instances where the model is highly confident but incorrect. This is achieved through:

Overconfidence Masks. For each prediction, the probability score (*probs*) is compared against predefined thresholds (T_{normal} and $T_{diseased}$). These thresholds are hyperparameters set to distinguish between overconfident and appropriately confident predictions for each class. For example, T_{normal} could be set to 0.9, meaning any predicted probability >0.9 for a sample being normal would be considered overconfident. Similarly, $T_{diseased}$ could be set to 0.1, indicating that any predicted probability <0.1 for a sample being diseased would be deemed overconfident.

$$(\text{overconfident\_mask\_normal} = (probs > T_{normal}) \quad (3)$$

$$(\text{overconfident\_mask\_diseased} = (probs < T_{diseased}) \quad (4)$$

Penalty Computation. By applying the boolean masks *combined_mask_normal* and *combined_mask_diseased,* the model assigns penalties to predictions that are both overconfident and incorrect. These combined masks function as filters that isolate instances where the model is mistaken despite its high confidence. These masks likely synthesize information from both the model's confidence levels, indicated by the overconfidence masks, and the model's accuracy based on the ground truth labels. The severity of the penalty depends on the degree to which the predicted probability strays from the established threshold.

$$\text{penalty\_normal} = \text{combined\_mask\_normal} \times |probs - T_{normal}| \quad (5)$$

$$\text{penalty\_diseased} = \text{combined\_mask\_diseased} \times |T_{diseased} - probs| \quad (6)$$

Here is a breakdown of the terms in the equations:

- *penalty_normal* and *penalty_diseased* are the penalties computed for overconfident predictions in the normal and diseased classes, respectively.
- *combined_mask_normal* and *combined_mask_diseased* are boolean masks that identify instances that are both overconfident and incorrect for each respective class.
- *probs* are the predicted probabilities produced by the model.
- T_{normal} and $T_{diseased}$ are the thresholds that define what is considered an overconfident prediction for each class.

These penalties are then combined to yield an overall confidence-based penalty for each prediction.

Final Loss Computation: The label smoothing loss (L) and the confidence-based penalties are combined to compute the final CC-LS loss:

$$CC - LS = label\_smoothing\_loss + \lambda \times confidence\_penalty \quad (7)$$

Here, λ is a weighting factor that determines the confidence-based penalty's importance in the total loss. It allows balancing between generalization (from label smoothing) and calibration (from the confidence penalty). To empirically determine the optimal balance, a range of values for λ from 0.1 to 1.0 was explored. This range was chosen based on preliminary experiments aimed at maximizing model performance while preventing overfitting. Ultimately, $\lambda = 1.0$ was selected as it provided the best compromise between generalization and calibration, as evidenced by the validation set performance.

Model Architecture and Training Protocols

Prior to elaborating on the specifics of the training process, it's pertinent to mention that we have chosen the EfficientNet-B3[32] model as the backbone architecture for our deep learning model. This architecture was chosen because of its state-of-the-art performance in image classification tasks and its efficient utilization of computational resources, making it well-suited for the task of glaucoma classification.

In the study, a transfer learning strategy was employed, beginning with a model pretrained on the expansive ImageNet dataset.[33] This initial phase allowed the model to learn generic features from a broad spectrum of images, laying a solid foundation for specialized tasks. The pretrained model was then fine-tuned on our glaucoma-specific dataset. The training dataset was divided into portions for actual training and validation to focus the model's learning on distinguishing between glaucomatous and non-glaucomatous fundus images. This division allows for a comprehensive assessment of the model's accuracy and the reliability of its predictions. Further details on the dataset composition and division are described in section Dataset Details.

The fine-tuning phase was meticulously executed to enhance the model's ability to accurately detect glaucoma, ensuring that predictions are both precise and accompanied by reliable confidence intervals. This approach guarantees that the model not only achieves high accuracy in identifying the presence or absence of glaucoma but also supports clinical decision-making by providing confidence measures alongside its predictions. By adopting this transfer learning methodology, our model is optimized to assist in the early detection and accurate categorization of glaucoma, demonstrating its potential to significantly aid in medical diagnostic processes.

To facilitate model calibration, we incorporate a customized loss function, referred to as CC-LS, during the training regimen. The Adam optimizer[34] is employed to train the model across 50 epochs, using a batch size of 16 and a learning rate of $1e^{-4}$. To circumvent the pitfalls of model overfitting, we implement several data augmentation techniques, including but not limited to geometrical transformations like scaling, flipping, and rotating. These strategies are designed to augment the model's robustness and generalizability to new or unseen data. To ensure consistency across our experiments, we applied the same training settings, including a batch size of 16, a learning rate of $1e^{-4}$, and the Adam optimizer, to train all other models tested in our study. This approach allows for a fair comparison of the effectiveness of the CC-LS loss function relative to other loss functions evaluated, while maintaining consistency in our experimental methodology.

Results

In the training and evaluation phases, a varied selection of datasets is utilized, including retinal fundus images from the reputable Online Retinal Fundus Image Database for Glaucoma Analysis and Research[35] and Retinal Fundus Glaucoma Challenge[36] datasets, as well as a specialized internal dataset. This multipronged approach to data sourcing significantly elevates the model's robustness and ensures that it remains applicable and generalizable in diverse clinical scenarios. Subsequent sections will provide a detailed examination of the datasets in use, the experimental setup, and offer both qualitative and quantitative analyses of the results, including calibration curve analysis.

Dataset Details

Training and Validation Dataset. The presented model leverages a diverse set of training data to ensure robust learning. The Online Retinal Fundus Image Database for Glaucoma Analysis and Research dataset[35] includes 482 fundus images classified as normal and 168 as glaucoma. The Retinal Fundus Glaucoma Challenge dataset[36] further contributes with 720 normal fundus images and 80 glaucoma fundus images. Additionally, a specialized internal dataset is utilized, comprising 4639 fundus images for each category—normal and glaucoma. This varied assortment of datasets, featuring different proportions of normal and glaucoma images, facilitates a nuanced training environment. It allows the model to learn from a broad spectrum of examples, enhancing its ability to generalize and accurately evaluate glaucoma presence. Furthermore, 20% of the data from each dataset is reserved as a validation set to assess the model's performance, while the remaining 80% is utilized for actual training, ensuring a comprehensive evaluation and optimization process.

Test Dataset. For evaluation purposes, the analysis primarily utilizes a comprehensive test subset derived from the internal dataset. It is important to note that the test images were selected randomly from this subset, ensuring a diverse representation of data for robust evaluation. This subset consists of 49 542 fundus images, categorized into 47 913 normal and 1629 glaucoma images. Selected for its extensive size and diversity, this test subset accurately reflects the variety of data clinicians are likely to encounter, making it an ideal platform for assessing the model. The evaluation is conducted using both quantitative and qualitative metrics, ensuring a thorough examination of our loss function, CC-LS, and its efficacy in medical diagnostic applications. Thus, this test set serves as the foundation for both quantitative and qualitative evaluations, allowing for an in-depth analysis of the CC-LS loss function's impact on the model's performance in actual medical diagnostic scenarios.

Quantitative Analysis

For an in-depth understanding of the model's performance, 3 variants were analyzed, each distinguished by the loss function used during training. The first model employed the traditional BCE loss function,[7] serving as a baseline for comparison. The second model was configured using focal loss, an adaptation often employed to handle class imbalance. The third and final model utilized the proposed loss function, designed specifically to enhance calibration in glaucoma classification tasks.

These 3 configurations allow for a comprehensive analysis of loss function impact on model performance, particularly in terms of accuracy and calibration. The succeeding subsections delve into the detailed outcomes of these evaluations.

Calibration Metrics: Brier Score Loss. To rigorously assess the calibration performance of the models, the Brier Score Loss[3] was employed as the key metric. It quantifies how well the predicted probabilities match the actual outcomes and is particularly useful for gauging the reliability of probabilistic predictions. A lower Brier Score indicates better calibration, making it an ideal choice for this evaluation.

The calculated Brier Score Loss values for the different loss functions used in model training are depicted in Table 1.

From the obtained results, it is evident that the model trained using the CC-LS function demonstrates the best calibration, as indicated by its lowest Brier Score Loss of 0.098. This is followed by the model trained with focal loss, which has a Brier Score Loss of 0.145. The model trained using BCE loss shows the least optimal calibration, reflected by the highest Brier Score Loss of 0.195.

Table 1. Brier Score Loss Values for Different Loss Functions

Loss Function Used	Brier Score Loss
Binary cross-entropy	0.195
Focal loss	0.145
CC-LS loss	0.098

CC-LS = Confidence-Calibrated Label Smoothing.

These results confirm the efficacy of the presented loss function in improving the model's calibration capabilities compared with traditional loss functions.

Comparative Metrics for Model Evaluation. The key quantitative measures used for assessing the performance of the models are as follows:

- **Sensitivity**, also known as the true positive rate, measures the proportion of actual positive cases (in this context, glaucoma cases) that are correctly identified by the model. It is vital for medical applications, where missing a positive case can have severe implications. Mathematically, sensitivity can be defined as:

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{8}$$

- **Specificity**, also known as the true negative rate, indicates the proportion of actual negative cases (in this context, nonglaucoma cases) that are correctly identified by the model. In medical settings, a high specificity ensures that patients who do not have the condition are not subjected to unnecessary treatments or tests. Mathematically, specificity can be defined as:

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive\ (FP)} \tag{9}$$

- **Weighted Accuracy** is a measure that takes into account both sensitivity and specificity, weighted by the prevalence of each class in the dataset. It can be formally defined as:

$$Weighted\ Accuracy = w_1 \times Sensitivity + w_2 \times Specificity \tag{10}$$

Where w_1 and w_2 are the weights representing the prevalence of the positive and negative classes in the dataset, respectively. These weights should sum to 1 (w_1+$w_2 = 1$). The weights w_1 and w_2 can be calculated as follows:

$$w_1 = \frac{Number\ of\ Actual\ Positive\ Samples}{Total\ Number\ of\ Samples}$$

$$w_2 = \frac{Number\ of\ Actual\ Negative\ Samples}{Total\ Number\ of\ Samples}$$

By using this formula, the weighted accuracy provides a more balanced view of the model's performance across different classes, especially in scenarios where the dataset is imbalanced.

In the Table 2, a comparison is made between the performance metrics of models trained with 3 distinct loss functions: BCE loss, focal loss, and CC-LS function. The proposed loss function shows a clear advantage, achieving 81% in sensitivity, 80% in specificity, and an overall accuracy rate of 80%. Models trained with BCE and focal loss show comparatively lower performance metrics, with the highest scores being 76% in sensitivity and specificity and 75% in accuracy for the model trained with focal loss. The improved performance across all metrics when employing the CC-LS function suggests its effectiveness in enhancing the accuracy and reliability of classification models, thus making it a favorable choice for tasks related to medical image classification.

Table 2. Comparative Evaluation of Model Metrics

Loss Function Used	Sensitivity (%)	Specificity (%)	Weighted Accuracy (%)
Binary cross-entropy	76	78	78
Focal loss	78	77	77
CC-LS loss	81	80	80

CC-LS = Confidence-Calibrated Label Smoothing.

Following the evaluation of sensitivity and specificity, it is pertinent to address the CC-LS model's false positive rate (FPR) in glaucoma detection. The model exhibits a FPR of 20%, indicating that 20% of nonglaucomatous images are erroneously classified as glaucomatous. This performance metric is crucial for assessing the model's clinical efficiency, as a lower FPR minimizes false alarms, thereby alleviating unnecessary patient anxiety and reducing the workload on health care systems. Despite the challenge of minimizing FPR, our model achieves a commendable balance with an 81% sensitivity rate, underscoring its proficiency in accurately detecting glaucomatous conditions.

When compared to other models that employ BCE loss and focal loss, with FPRs of 22% and 23% respectively, the CC-LS model demonstrates a superior balance between sensitivity and specificity. This balance is vital for effective glaucoma screening, aiming to ensure a high detection rate of true positives while maintaining a manageable rate of false positives. The development of the CC-LS model marks a significant advancement in ophthalmic diagnostics, contributing to the improvement of early detection and management strategies for glaucoma.

Qualitative Analysis: Heatmap Visualization for Feature Relevance

In this segment, the qualitative dimension of model performance is scrutinized using heatmap visualizations. These visual aids serve as powerful tools for interpreting the feature importance attributed by the classification model to different regions within the medical images. Significantly, the heatmaps substantiate that the model allocates higher weights to regions around the OD, which is a critical locus for glaucoma diagnosis according to established medical literature.

By systematically attending to the OD region, the model illustrates its adherence to medically relevant criteria. This focus on the OD region not only validates the feature extraction capabilities of the model but also enhances the

interpretability and clinical utility of the diagnostic system. In tandem with the superior quantitative metrics achieved through the implementation of the CC-LS function, these qualitative observations further advocate for the robustness and applicability of this machine learning-based diagnostic approach in a medical imaging context.

Additionally, the heatmaps reveal the nuanced understanding developed by the model during its training phase. While focusing on the OD region is paramount for glaucoma diagnosis, an incorrect or shallow feature extraction could have resulted in false positives or negatives. The heatmaps confirm that this is not the case; they demonstrate a precise and targeted focus on the OD region's distinguishing characteristics, which are critical for accurate glaucoma detection. This targeted feature recognition adds another layer of confidence in the model's diagnostic capabilities. Together with the compelling quantitative performance metrics, the heatmap-based qualitative analysis fortifies the argument for adopting this model as a reliable and insightful tool for medical image classification in the realm of ophthalmology.

Figure 1 showcases heatmaps and zoomed-in views of the OD for 2 distinct sets of glaucomatous input images, highlighting the areas of interest identified by our CC-LS model for the classification of glaucomatous conditions. Organized into 2 subfigures, set 1 and set 2, each set is composed of 4 images that offer varied perspectives on the model's analytical focus, especially on the OD region which is crucial for detecting glaucoma.

In set 1 (Fig 1A), the sequence begins with the original input image (a), followed by a heatmap (b) that visualizes the model's focus areas for classification, underscoring its ability to identify glaucoma-relevant features. The third image (c) provides a closer look at the OD region of the original input, while the fourth image (d) zooms into the OD region within the heatmap. Both the heatmap and its zoomed-in version distinctly highlight the model's precision in recognizing the features indicative of glaucoma within the OD, affirming its classification accuracy.

Set 2 (Fig 1B) mirrors this layout, further validating the model's consistent and effective focus on critical areas for diagnosing glaucoma. The heatmap and its detailed view of the OD region in this set again confirm the model's adeptness at concentrating on the essential aspects for its classification task.

Crucially, the images presented in both sets are confirmed to be glaucomatous, with the heatmaps and zoomed-in views of the OD region serving as visual evidence of the model's capability to accurately detect and classify glaucomatous changes. This focused approach not only demonstrates the model's proficiency in identifying key diagnostic features but also aligns with the clinical diagnosis of glaucoma, showcasing the potential of the CC-LS model as a valuable tool in the early detection and accurate diagnosis of this condition.

Calibration Curve Analysis

In the depicted calibration curve comparison in Figure 2, the graph illustrates the alignment of 3 different loss functions—CC-LS loss, BCE loss, and focal loss—with the perfect calibration line. The CC-LS loss function, represented by the green curve, distinctly demonstrates the most consistent adherence to the ideal calibration line across the full range of predicted probabilities. This superior calibration indicates that the CC-LS loss function is capable of providing exceptionally accurate probability estimates, crucial for applications where precise risk assessments are paramount. Conversely, the BCE loss and focal loss functions, represented by the red and blue curves respectively, show greater deviations from perfect calibration, suggesting that the CC-LS loss-based model is better calibrated compared with the other 2 loss functions.

Discussion

Significance of CC-LS Function

The primary focus of this study was to introduce and evaluate the CC-LS loss function, designed to improve the calibration of deep learning models for glaucoma classification. The quantitative metrics employed—sensitivity, specificity, and accuracy—served as reliable measures for assessing the model's performance. Notably, the proposed loss function outperformed traditional loss functions like BCE and focal lacross all performance metrics. This superiority was most evident in the Brier Score Loss, a key metric for model calibration, where the CC-LS function attained a significant improvement over the existing loss functions.

The CC-LS enhances traditional BCE[7] by emphasizing both accuracy and calibration. While BCE primarily focuses on accuracy, the CC-LS integrates label smoothing and confidence penalty techniques to improve calibration. This ensures that the model's predictions are not only accurate but also appropriately calibrated, especially vital in medical diagnostics. Through empirical evaluations, the effectiveness of the CC-LS is evident, consistently outperforming BCE across various metrics, thereby providing more reliable predictions.

In comparison to focal loss, the CC-LS takes a distinct route to address model calibration and performance. While focal loss aims to mitigate the impact of class imbalance and noisy data during training, the CC-LS directly targets calibration using label smoothing and confidence penalty strategies. By promoting well-calibrated predictions and penalizing overconfidence, the CC-LS enhances uncertainty estimation, leading to more dependable predictions, particularly in challenging scenarios. Empirical results illustrate that the CC-LS achieves comparable or superior accuracy to focal loss while significantly improving calibration, demonstrating its efficacy in medical image classification tasks.

Heatmap Analysis Insights

The heatmap analysis conducted within this study is specifically tied to the model trained using the CC-LS function. This approach distinctively demonstrates the model's proficiency in identifying and prioritizing critical diagnostic features in retinal fundus images for glaucoma detection, with a notable focus on vital areas such as the OD. The

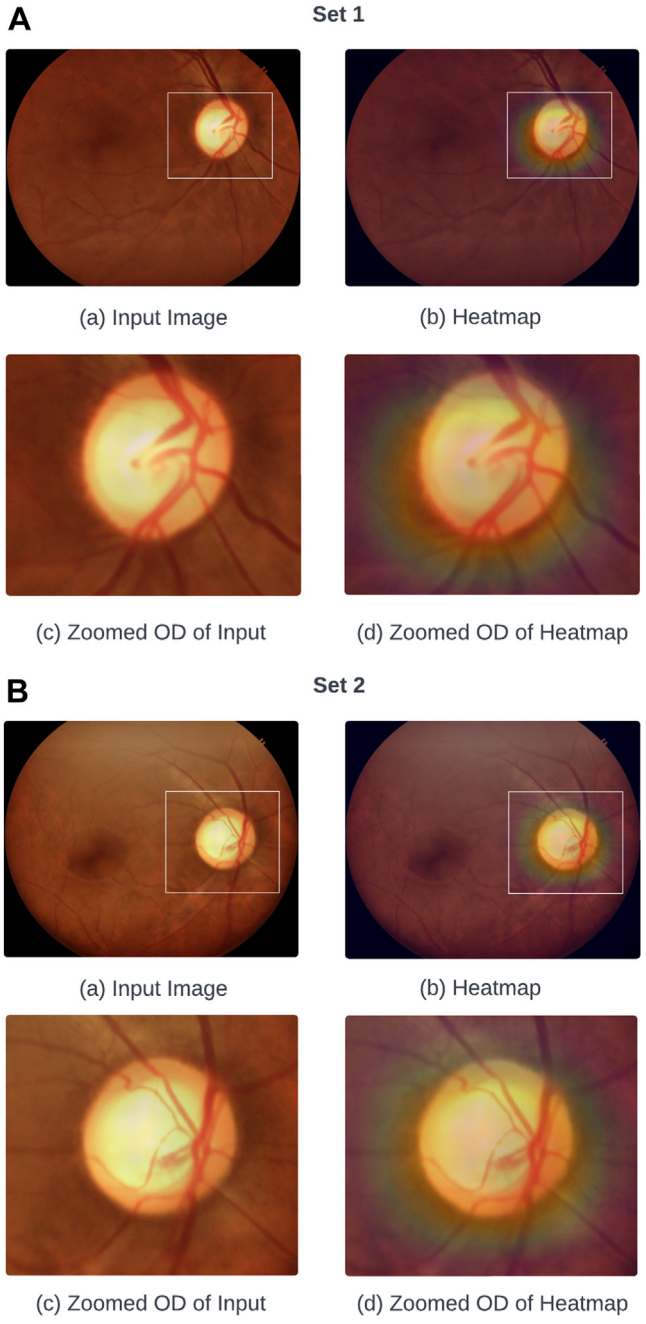

Figure 1. Visualization of heatmap and zoomed OD region for 2 glaucomatous input images (Set 1, **A**, and Set 2, **B**). For each set: (**a**) Input image, (**b**) Heatmap, (**c**) Zoomed OD of input, and (**d**) Zoomed OD of Heatmap. OD = optic disc.

study does not feature direct heatmap comparisons from models trained with alternative loss functions. However, the precision observed in the CC-LS model's focus on relevant diagnostic features, alongside its improved accuracy and calibration, highlights its superior capability in emphasizing significant features for glaucoma diagnosis. Future work is encouraged to include direct comparisons through heatmap analyses of models trained with various loss functions to empirically validate these observations.

Interpretability and Clinical Relevance

A critical aspect that sets this work apart is the focus on model interpretability, a feature often overlooked in machine learning models for health care applications. The qualitative heatmap analysis reinforced the model's clinical relevance. Importantly, the heatmaps confirmed that the model focuses on the OD region when making a classification, a crucial factor for potential clinical applications.

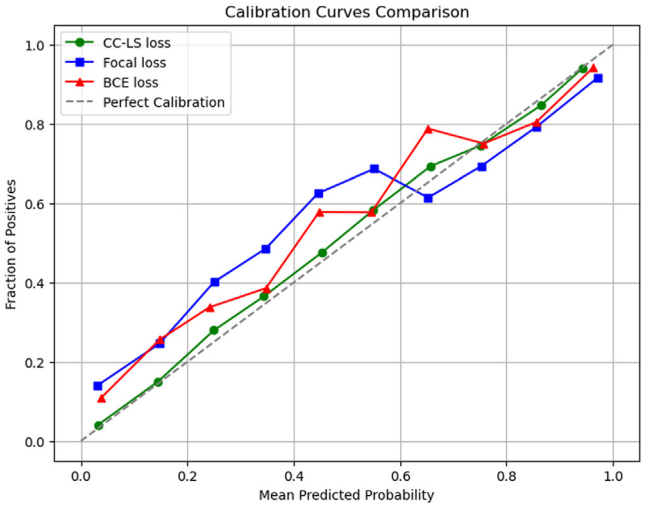

Figure 2. Comparative calibration curves for different loss functions. BCE = binary cross-entropy; CC-LS = Confidence-Calibrated Label Smoothing.

This aligns with medical understanding, providing an additional layer of validation to the quantitative results.

In this study, the test set comprehensively included images representing all stages of glaucoma, alongside borderline cases and more definitive conditions, to thoroughly evaluate the CC-LS model. This diverse inclusion was strategically aimed at assessing the model's capability to accurately identify glaucoma across a broad spectrum of presentations, from early and ambiguous stages to advanced and clear-cut cases. The quantitative analysis, with a particular focus on sensitivity and specificity, has confirmed the model's effectiveness in detecting glaucoma under all conditions presented within these stages. The encouraging results for borderline cases, known for their diagnostic challenges, further highlight the model's robust performance. The ability of the CC-LS model to effectively identify glaucoma across its various stages underscores its potential as a valuable supportive tool for clinicians. It offers a nuanced approach to glaucoma detection, ensuring comprehensive coverage of the full range of clinical presentations encountered in practice.

Ensuring Data Integrity: Measures Against Data Leakage in Model Training

In this study, we utilized mutually exclusive external datasets exclusively for training and validation, ensuring they were not used in the testing phase. The internal dataset, however, was employed across training, validation, and testing. We took strict measures to prevent data leakage between these phases. Initially, we confirmed there were no duplicates in the dataset, then carefully partitioned it to ensure that no data appeared in >1 subset. Each dataset—training, validation, and testing—was processed independently to avoid any unintended leakage of information. Specifically, normalization parameters were calculated using only the training set data. Furthermore, during model development, we fine-tuned hyperparameters solely on the validation set, while the test set was completely isolated and used only for the final evaluation. These measures were crucial to ensure that our results are reliable and truly reflect the model's performance on unseen data.

Limitations and Future Work

While the results are promising, there are a few caveats to consider. Firstly, the study is constrained by its focus on specific types of medical images and the exclusion of other ocular diseases that may cooccur with glaucoma. Although the datasets employed include a range of glaucoma stages, they may not fully capture the breadth of variations seen in actual clinical scenarios. Future studies should aim to incorporate more heterogeneous datasets, including a wider array of glaucoma stages and potentially images from different medical imaging modalities to better reflect the diversity encountered in clinical practice.

The effectiveness of the CC-LS loss function could be further evaluated by applying it to different deep learning architectures. It would also be worthwhile to explore multiobjective loss functions that combine different types of errors to optimize multiple aspects of model performance simultaneously.

Currently, the CC-LS model is designed for binary classification, distinguishing between glaucomatous and nonglaucomatous images without specifying the stage of the disease. As a future direction, there are plans to enhance the model to predict the specific stages of glaucoma, aiming to provide even more detailed insights for clinical assessment and management.

In summary, the proposed loss function, CC-LS, exhibits a compelling case for its adoption in medical image classification tasks, particularly in glaucoma diagnosis. It excels in both calibration and interpretability, crucial aspects for practical health care applications. The research opens several avenues for future investigation, including the application to other medical conditions, integration with different machine learning architectures, and further clinical validation. A prime example is diabetic retinopathy, an ocular condition where our approach could be highly beneficial. Beyond diabetic retinopathy, we foresee potential

applications in other medical areas where diagnostic accuracy is crucial.

This study introduces CC-LS loss function, designed to calibrate glaucoma classification models. The CC-LS loss function demonstrated significant improvements over the conventional loss functions discussed in earlier sections, showing enhanced performance in terms of sensitivity, specificity, and accuracy. Furthermore, the Brier Score Loss metric corroborated the superior calibration characteristics of the proposed loss function.

Beyond quantitative performance metrics, the model's interpretability was verified through heatmap visualizations. These heat maps confirmed the model's focus on the OD region, making the findings particularly relevant for clinical applications. The study thereby addresses a critical gap in the literature by providing a calibrated and interpretable model for glaucoma classification.

Future research could extend the present work by applying the CC-LS loss function to other machine learning architectures and medical imaging modalities. Furthermore, the inclusion of more heterogeneous datasets and the exploration of multiobjective loss functions could provide additional insights into the method's applicability and robustness.

In conclusion, the CC-LS loss function offers an effective, calibrated, and interpretable approach for glaucoma classification, warranting further investigation and clinical validation.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used OpenAI's large language model for linguistic corrections to enhance the clarity of this manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Footnotes and Disclosures

Originally received: November 28, 2023.
Final revision: May 8, 2024.
Accepted: May 17, 2024.
Available online: June 22, 2024. Manuscript no. XOPS-D-23-00307.

Research and Development, Forus Health Pvt. Ltd., Bengaluru, Karnataka, India.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosure(s):

M.V.: Article processing charges — Forus Health Pvt. Ltd., Bangalore, India.

D.K.P.: Article processing charges — Forus Health Pvt. Ltd., Bangalore, India.

S.V.: Article processing charges — Forus Health Pvt. Ltd., Bangalore, India.

HUMAN SUBJECTS: Human subjects were included in this study. Human subjects data were utilized in this study in the form of preexisting medical images. These images were initially collected for clinical assessment and diagnosis with appropriate ethical oversight. This study strictly adheres to the ethical standards outlined in the Declaration of Helsinki. During the original collection of these images, informed consent was appropriately obtained from all patients under protocols that assured the confidentiality and anonymity of their medical data. In cases where patient consent was not required or was waived because of the retrospective nature of the images, this was done in strict accordance with ethical guidelines and with approval from the relevant institutional review board at Forus Health Pvt. Ltd.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Midhula Vijayan

Data collection: Midhula Vijayan, Deepthi Keshav Prasad

Analysis and interpretation: Midhula Vijayan, Deepthi Keshav Prasad, Venkatakrishnan Srinivasan

Obtained funding: Study was performed as part of regular employment duties at Forus Health Pvt. Ltd. No additional funding was provided

Overall responsibility: Midhula Vijayan, Deepthi Keshav Prasad, Venkatakrishnan Srinivasan

Abbreviations and Acronyms:
AI = artificial intelligence; **BCE** = binary cross-entropy; **CC-LS** = Confidence-Calibrated Label Smoothing; **FPR** = false positive rate; **OD** = optic disc.

Keywords:
Glaucoma, Deep learning, Model calibration, EfficientNet, Loss function.

Correspondence:
Midhula Vijayan, Research and Development, Forus Health Pvt. Ltd., Bengaluru, Karnataka 560070, India. E-mail: midhula@forushealth.com.

References

1. Velpula VK, Sharma LD. Multi-stage glaucoma classification using pretrained convolutional neural networks and voting-based classifier fusion. *Front Physiol*. 2023;14:1175881.
2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*. 2017;542(7639):115−118.
3. Minderer M, Djolonga J, Romijnders R, et al. Revisiting the calibration of modern neural networks. *Adv Neural Inf Process Syst*. 2021;34:15682−15694.
4. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med*. 2021;4(1):4.
5. Havasi M, Jenatton R, Fort S, et al. Training independent subnetworks for robust prediction. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2010.06610.
6. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International conference on machine learning*. Sydney, Australia: PMLR; 2017:1321−1330.

7. Ho Y, Wookey S. The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. *IEEE Access*. 2019;8:4806−4813.
8. Camara J, Neto A, Pires IM, et al. Literature review on artificial intelligence methods for glaucoma screening, segmentation, and classification. *J Imaging*. 2022;8(2):19.
9. Coan LJ, Williams BM, Adithya VK, et al. Automatic detection of glaucoma via fundus imaging and artificial intelligence: a review. *Surv Ophthalmol*. 2023;68(1):17−41.
10. Phasuk S, Poopresert P, Yaemsuk A, et al. Automated glaucoma screening from retinal fundus image using deep learning. In: *41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), Germany*. IEEE; 2019:904−907.
11. Parthasarathy DR, Hsu CK, Eldeeb M, et al. Development and performance of a novel offlinedeep learning (DL)-based glaucoma screening tool integrated on a portable smartphone- based fundus camera. *Invest Ophthalmol Vis Sci*. 2021;62(8):1002.
12. Liu Y, Yip LWL, Zheng Y, Wang L. Glaucoma screening using an attention-guided stereo ensemble network. *Methods*. 2022;202:14−21.
13. de Sales Carvalho NR, Rodrigues MdCLC, de Carvalho Filho AO, Mathew MJ. Automatic method for glaucoma diagnosis using a three-dimensional convoluted neural network. *Neurocomputing*. 2021;438:72−83.
14. Aamir M, Irfan M, Ali T, et al. An adoptive threshold-based multi-level deep convolutional neural network for glaucoma eye disease detection and classification. *Diagnostics*. 2020;10(8):602.
15. Liao W, Zou B, Zhao R, et al. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE J Biomed Health Inform*. 2019;24(5):1405−1412.
16. Nawaz M, Nazir T, Javed A, et al. An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization. *Sensors*. 2022;22(2):434.
17. Diaz-Pinto A, Morales S, Naranjo V, et al. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online*. 2019;18:1−19.
18. Juneja M, Thakur S, Uniyal A, et al. Deep learning-based classification network for glaucoma in retinal images. *Comput Electr Eng*. 2022;101:108009.
19. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. *Med Phys*. 2019;46(1):e1−e36.
20. Liang G, Zhang Y, Wang X, Jacobs N. Improved trainable calibration method for neural networks on medical imaging classification. *ArXiv*. 2020. https://doi.org/10.48550/arXiv.2009.04057.
21. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. *J Am Med Inf Assoc*. 2012;19(2):263−274.
22. Thulasidasan S, Chennupati G, Bilmes JA, et al. On mixup training: improved calibration and predictive uncertainty for deep neural networks. *Adv Neural Inf Process Syst*. 2019;32:13888−13899.
23. Wen Y, Jerfel G, Muller R, et al. Combining ensembles and data augmentation can harm your calibration. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2010.09875.
24. Carneiro G, Pu LZCT, Singh R, Burt A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Med Image Anal*. 2020;62:101653.
25. Pollastri F, Maroñas J, Bolelli F, et al. *Confidence calibration for deep renal biopsy immunofluorescence image classification. 2020 25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy (held virtually): IEEE; 2021:1298−1305.
26. Rajaraman S, Ganesan P, Antani S. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLoS One*. 2022;17(1):e0262838.
27. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE; 2016:2818−2826.
28. Müller R, Kornblith S, Hinton GE. When does label smoothing help? *Adv Neural Inf Process Syst*. 2019;32:4696−4705.
29. Pereyra G, Tucker G, Chorowski J, et al. Regularizing neural networks by penalizing confident output distributions. *arXiv*. 2017. https://doi.org/10.48550/arXiv.1701.06548.
30. Kumar A, Sarawagi S, Jain U. Trainable calibration measures for neural networks from kernel mean embeddings. In: *International Conference on Machine Learning*. Stockholm, Sweden: PMLR; 2018:2805−2814.
31. Wald Y, Feder A, Greenfeld D, Shalit U. On calibration and out-of-domain generalization. *Adv Neural Inf Process Syst*. 2021;34:2215−2227.
32. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. Long Beach, USA: PMLR; 2019:6105−6114.
33. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA: IEEE; 2009:248−255.
34. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv*. 2014. https://doi.org/10.48550/arXiv.1412.6980.
35. Zhang Z, Yin FS, Liu J, et al. Origa-light: an online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. Buenos Aires, Argentina: IEEE; 2010:3065−3068.
36. Fu H, Li F, Orlando JI, et al. *REFUGE: retinal fundus glaucoma challenge*. Spain: Springer; 2019.