IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2025.Doi Number

Uncertainty-Aware Selective Prediction of Neovascular Age-Related Macular Degeneration Recurrence Using Artificial Intelligence

Won Tae Yoon[1, ¶], Hun-gyeom Kim[2, ¶], Junhyung Moon[3, ¶], Dat Ngo[4], Jae Hui Kim[1, *], and Baek Hwan Cho[3, 5, *]

[1]Department of Ophthalmology, Kim's Eye Hospital, Seoul, Korea
[2]Department of Biomedical Engineering, Hanyang University, Seoul, Korea
[3]Department of Biomedical Informatics, CHA University School of Medicine, CHA University, Seongnam, Korea
[4]Department of Computer Engineering, Korea National University of Transportation, Chungju, Korea
[5]Institute of Biomedical Informatics, CHA University School of Medicine, CHA University, Seongnam, Korea

* Co-corresponding author: J.H.K. and B.H.C. contributed equally as the co-corresponding authors. E-mail addresses: kjh7997@kimeye.com (J.H. Kim); baekhwan.cho@cha.ac.kr (B.H. Cho)

¶ Co-first author: W.T.Y., H.K., and J.M. contributed equally as the co-first authors. E-mail addresses: gentlewt@kimeye.com (W.T. Yoon); gnsruatkfkd@hanyang.ac.kr (H. Kim); junhyung.moon@chauniv.ac.kr (J. Moon).

Co-author: D. N. contributed as the co-author. E-mail address: datngo@ut.ac.kr (D. Ngo).

ABSTRACT Neovascular age-related macular degeneration (nAMD) is a progressive, sight-threatening disease predominantly affecting older adults and a leading cause of global vision loss projected to rise with aging populations. Although existing artificial intelligence (AI) studies have investigated nAMD recurrence detection, predictive modeling remains underexplored, with limited performance in a few studies relying on a single type of retinal image. Therefore, we aimed to comprehensively explore AI-based recurrence prediction incorporating multiple types of retinal images and clinical data. We collected retinal photographs and clinical data from 399 nAMD patients. The retinal photographs are six images, comprising fundus photographs and optical coherence tomography (OCT) horizontal/vertical scans taken both before and after treatment. The clinical data includes demographic information, such as age, sex and medication type. Using our collected dataset, we proposed an uncertainty-aware selective-prediction model for nAMD recurrence. The proposed model, built on popular CNN architectures, filters out highly uncertain predictions to enhance reliability of the recurrence prediction. Results showed that OCT horizontal post-treatment images achieved an area under the curve (AUC) of 0.662, which increased to 0.836 with Monte Carlo (MC) dropout by excluding uncertain predictions, while the double-threshold method yielded an AUC of 0.744. Incorporating clinical parameters further improved performance, raising the AUC to 0.855, and drug-specific modeling produced particularly high predictive performance, with AUCs of 0.934 for ranibizumab and 0.978 for aflibercept. Our experimental results highlight the potential of AI-based multimodal modeling, selective prediction, and treatment-specific analysis in improving prognostic prediction and guiding therapeutic decision-making for patients with nAMD.

INDEX TERMS Neovascular age-related macular degeneration recurrence prediction; Artificial intelligence; Selective prediction; Monte Carlo dropout; Probabilistic thresholding; Multimodal data

I. INTRODUCTION

Neovascular age-related macular degeneration (nAMD) is a sight-threatening disease affecting primarily older adults and represents a leading cause of visual impairment worldwide [1, 2]. With the global increase in the elderly population, the prevalence of nAMD is expected to rise further [3].

Anti-vascular endothelial growth factor (VEGF) therapy has emerged as the gold standard for treating nAMD and has shown excellent efficacy in maintaining or improving vision. However, the effectiveness of the drug diminishes over time after injection [4, 5], often leading to recurrence, necessitating continued treatment. Early clinical trials using anti-VEGF drugs utilized a fixed-dosing method involving injections at 1-to 2-month intervals [6, 7]. To optimize treatment efficiency, the as-needed regimen [8], which administers injections only upon recurrence, and the treat-and-extend (TAE) regimen [9], which continues injections while adjusting the interval based on treatment response, were introduced.

Methods such as the TAE regimen, which involves continuous injections regardless of recurrence, are known to yield better visual outcomes than the as-needed regimen, which administers injections only upon recurrence [10]. However, considering that a certain proportion of patients may remain recurrence-free with no additional injections after the initial loading injections [11-13], using the TAE regimen may require unnecessary injections in some patients [13]. Therefore, the prediction of recurrence after initial loading injections could significantly contribute to more efficient treatment. Although factors associated with recurrence after loading injections have been investigated [11, 13], no consensus has been reached regarding any specific factor; hence, the prediction of recurrence remains challenging.

The potential of artificial intelligence (AI) in predicting the clinical course and treatment outcomes of nAMD has been extensively investigated. Studies have suggested that AI can match or even surpass the prediction accuracy of human examiners [14-17]. While most existing studies have explored AI-based nAMD detection [18-20], a few works have attempted to predict nAMD recurrence using AI [21, 22]. However, these studies primarily focused on OCT images without attempting to incorporate other diverse modalities, and their models achieved limited performance such as accuracy of 0.602 [21] and AUC of 0.725 [22]. These approaches thus still face the inherent challenge of reliably forecasting recurrence, underscoring the need for further methodological advancements.

Therefore, we propose an uncertainty-aware selective prediction model for nAMD recurrence. Our proposed model predicts the recurrence of nAMD within 12 months following initial loading injections of ranibizumab or aflibercept. To ensure reliable decision-making and improve clinical applicability, the model provides predictions only when confidence is high. Our contributions include (1) To enable reliable nAMD recurrence prediction, we effectively leveraged two existing selective prediction techniques: MC dropout-based uncertainty estimation [23, 24] and double-thresholding [25], both of which help exclude low-confidence cases and enhance model reliability. (2) We comprehensively explored the effectiveness of multimodal data in predicting nAMD recurrence, incorporating six types of retinal images along with clinical data, including demographic information and anti-VEGF medication type. (3) Finally, we employed drug-specific modeling to predict nAMD recurrence for each anti-VEGF agent, yielding clinically relevant insights into recurrence patterns that have not been previously explored.

II. RELATED STUDIES

In this section, we review relevant literature in two key areas: (1) deep learning approaches for nAMD detection and the associated challenges in recurrence prediction, and (2) the role of selective prediction in enhancing clinical reliability.

A. DEEP LEARNING FOR nAMD DETECTION AND CHALLENGES IN RECURRENCE PREDICTION

Deep learning has significantly advanced medical image analysis and has been widely applied to the detection and classification of nAMD using OCT. Most existing studies have focused on distinguishing wet AMD from normal and dry AMD [26, 27] or detecting the presence of the disease from OCT or fundus images [18-20, 28], or performing segmentation of retinal structures such as fluid [29-32]. Chen et al. [33] developed a deep learning-based framework that integrates OCT and fundus imaging for AMD diagnosis, significantly improving diagnostic accuracy and interpretability compared to single-modality approaches. Furthermore, Li et al. in [17] showed that employing a multi-modal, multi-instance learning approach effectively combines diverse imaging data, enhancing both the classification performance for ophthalmic diseases, including AMD, and the robustness of the model. Moreover, recent studies have sought to precisely segment fluid regions and retinal layers—structures associated with nAMD—to facilitate lesion quantification or the development of prognostic models [34] and have integrated eye-tracking technology to highlight on fundus images the areas clinicians most frequently examine to enhance diagnostic efficiency, but these approaches remain primarily focused on detection [35].

Despite these advances in diagnosis, relatively few studies have addressed the prediction of disease recurrence—an essential factor in effective nAMD management. Predicting recurrence remains challenging due to the complex and heterogeneous progression of the disease. Current models for predicting the recurrence of nAMD demonstrate only moderate performance. One line of research utilizes four OCT

images—captured at the onset of nAMD, followed by three anti-VEGF injections and an additional scan one month after the final injection—to incorporate temporal information through LSTM- and attention-based networks. These models typically achieved performance scores ranging from 0.5 to 0.6 [21]. Another study employed a segmentation-based approach that automatically extracts three types of fluid regions from OCT images—intraretinal fluid, subretinal fluid, and pigment epithelial detachment—to predict whether the first recurrence will occur within three months. This method reported AUC values ranging from 0.6 to approximately 0.7 [22]. Nevertheless, these strategies continue to confront limitations in ensuring reliable recurrence prediction, thereby highlighting the need for more further methodological advancements.

B. IMPORTANCE OF SELECTIVE PREDICTION

Selective prediction provides a promising solution by allowing models to "reject" or "defer" predictions when uncertainty is high, thereby presenting only those results deemed sufficiently reliable. For instance, Gal and Ghahramani [36] introduced a Bayesian approximation technique using MC dropout to estimate uncertainty in deep learning models, while Geifman and El-Yaniv [37] proposed a selective classification framework that improves model reliability by rejecting highly uncertain predictions. Similarly, Lakshminarayanan et al. [38] demonstrated that deep ensembles can effectively scale uncertainty estimation, and DeVries and Taylor [39] advanced confidence-learning methods to detect out-of-distribution inputs.

A major limitation of current deep learning models is their limited reliability in clinical settings, where high-confidence predictions are crucial for patient safety and informed decision-making [40]. Traditional approaches could be challenging to explicitly quantify uncertainty, increasing the risk of misdiagnosis. In contrast, selective prediction techniques—such as MC dropout and probabilistic thresholding—allow models to defer uncertain predictions, thereby reducing false positives and enhancing overall reliability. This uncertainty-aware approach has been validated across various applications: for example, uncertainty-aware LSTM networks with a reject option have demonstrated reduced errors in time-series health data [41], and similar frameworks for medical image segmentation have improved decision reliability [42]. Furthermore, selective classification strategies have been shown to boost diagnostic accuracy [37]. In this study, we incorporated selective prediction, including MC dropout and probabilistic thresholding, to enhance the reliability of AI-driven recurrence predictions, thereby bridging the gap between model performance and clinical applicability.

III. METHODS

This retrospective study was approved by the Institutional Review Board of Kim's Eye Hospital (IRB no. 2023-02-004; February 2023) and conducted in accordance with the tenets of the Declaration of Helsinki. The need for informed consent was waived due to the retrospective nature of this study (Kim's Eye Hospital IRB, Seoul, South Korea). Data were collected at Kim's Eye Hospital, and AI model learning was performed at the CHA University School of Medicine (Bundang, South Korea).

C. STUDY PARTICIPANTS AND DATA COLLECTION

This study included treatment-naïve patients diagnosed with nAMD between January 2013 and June 2021, who initially received three loading injections using either ranibizumab (0.5 mg/0.05 mL, Lucentis®; Genentech Inc., San Francisco, CA, USA) or aflibercept (2.0 mg/0.05 mL, Eylea®; Regeneron, Tarrytown, NY, USA). The exclusion criteria were as follows: (1) residual intraretinal fluid (IRF) or subretinal fluid (SRF) after initial treatment; (2) follow-up duration of less than 12 months after initial treatment; (3) a history of vitreoretinal or glaucoma surgery; (4) incomplete labeling information necessary for determining recurrence status; and (5) absence of one or more imaging modalities required for model training and evaluation. After the initial treatment, follow-up examinations were conducted at 1-to 2-month intervals. The development of IRF, SRF, or hemorrhage within the macula, as observed by OCT, fundus photography, or clinical fundus examination, was considered an indication of recurrence.

Finally, a total of 399 patients diagnosed with nAMD (238 males and 161 females; mean age, 70.21 ± 8.38 years) were enrolled. Ranibizumab and aflibercept were administered as loading injections in 197 (recurrence: 114, non-recurrence: 83) and 202 (recurrence: 128, non-recurrence: 74) patients, respectively, as summarized in Table 1. Fundus photo images and OCT images captured before and after the initial loading injections were used in the experiment.

TABLE 1. Baseline demographic and clinical characteristics of the study population. The incidence rate reflects the proportion of patients who developed disease recurrence subsequent to treatment with each anti-VEGF agent during the observation period.

Participant characteristics	Description
Number of participants	399
Gender, n (%)	Male: 238 (59.6%) Female: 161 (40.4%)
Age, mean ± SD	70.21 ± 8.38
Incidence rate by anti-VEGF agent	Ranibizumab: 57.9% Aflibercept: 63.4%

D. DATA PREPROCESSING

Through the medical imaging process, we obtained six images per participant: fundus images and horizontal and vertical OCT images taken before and after treatment (Fig. 1). The

dataset comprised fundus images with a resolution of 940 × 840 pixels and OCT images (horizontal and vertical) with a resolution of 1000 × 650 pixels. All the images were resized to 448 × 448 pixels and normalized via Z-score normalization using the mean and standard deviation values of the ImageNet dataset [43].

Recurrence was used as the classification target for the developed AI models. Among 399 participants, 242 experienced recurrence within 12 months of treatment, whereas 157 did not. Additionally, data augmentation was applied to enhance the diversity of the training dataset. The augmentation techniques, adopted in this study, were horizontal flipping, brightness adjustment (range: 0.9–1.1), and random cropping to retain 95–100% of the original image content, followed by resizing to the original dimensions. These augmentation parameters were selected in consultation with ophthalmologists to simulate real-world clinical variations, such as subtle brightness variations and patient head movement. Crucially, the cropping range was strictly limited to the peripheral areas to preserve the central macular region, ensuring that no pathological features essential for diagnosis were compromised.

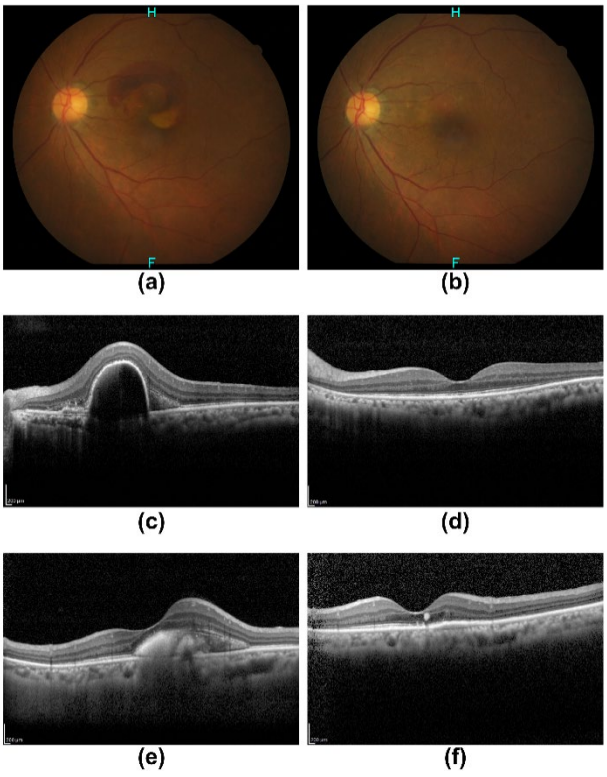


FIGURE 1. Representative set of six images from a single patient. (a) Fundus pre-treatment. (b) Fundus post-treatment. (c) OCT horizontal pre-treatment. (d) OCT horizontal post-treatment. (e) OCT vertical post-treatment. (f) OCT vertical post-treatment.

E. CNN-BASED ARCHITECTURE FOR nAMD RECURRENCE PREDICTION

We systematically investigated nAMD recurrence prediction using single-, dual-, and multi-modality configurations. The single-modality model (Fig. 2(a)) used only one of the six image types to predict the recurrence of nAMD. The dual-modality model (Fig. 2(b)) fused the two highest-performing single modalities (OCT horizontal pre- and post-treatment) via feature map concatenation [44, 45]. Finally, the multi-modality model depicted in Fig. 2(c) integrates all six image types to leverage all the visual information. We employed Inception-v3 [46], EfficientNet-b0 [47], and EfficientNet-v2 [48] as foundational CNN architectures across all modality configurations. These models varied substantially in backbone size, ranging from ~4.0M parameters (EfficientNet-b0) to ~22M (Inception-v3 and EfficientNet-v2). We prioritized CNN-based backbones over Transformer-based architectures to leverage inductive bias for robust feature extraction given the limited dataset size. Furthermore, feature-level concatenation was chosen over recurrent networks (e.g., LSTM) for temporal fusion to mitigate over-parameterization, as the input consists of only two discrete time points.

F. PROPOSED RELIABLE PREDICTION STRATEGIES: DOUBLE THRESHOLDING AND MC DROPOUT

In our CNN-based architecture, we designed novel selective methods for nAMD recurrence prediction by employing conventional schemes, MC dropout, and probabilistic thresholding to enhance the decision-making reliability [23, 24, 36, 37]. Specifically, MC dropout was employed as a computationally efficient Bayesian approximation to estimate uncertainty, avoiding the high resource costs associated with Deep Ensembles.

As a baseline, we employed a conventional classification method based on a single-threshold approach, as shown in Fig. 3 (a), wherein a predefined probability threshold was applied to the model output for class assignment. In our study, we determined the threshold via the Youden index [49], a widely used metric for optimizing the classification performance in medical prediction models.

Compared with the baseline, a naïve prediction approach, we propose two selective prediction strategies. The first approach, which is illustrated in Fig. 3(b), employs a double-threshold strategy [25]. To improve the reliability of probability estimates, temperature scaling [50] was applied as a post-processing calibration technique, adjusting the model's output logits using a temperature parameter T. This calibration mitigates overconfidence in predicted probabilities, ensuring that they more accurately reflect the true likelihood of correctness. The temperature parameter T was treated as a hyperparameter and optimized accordingly.

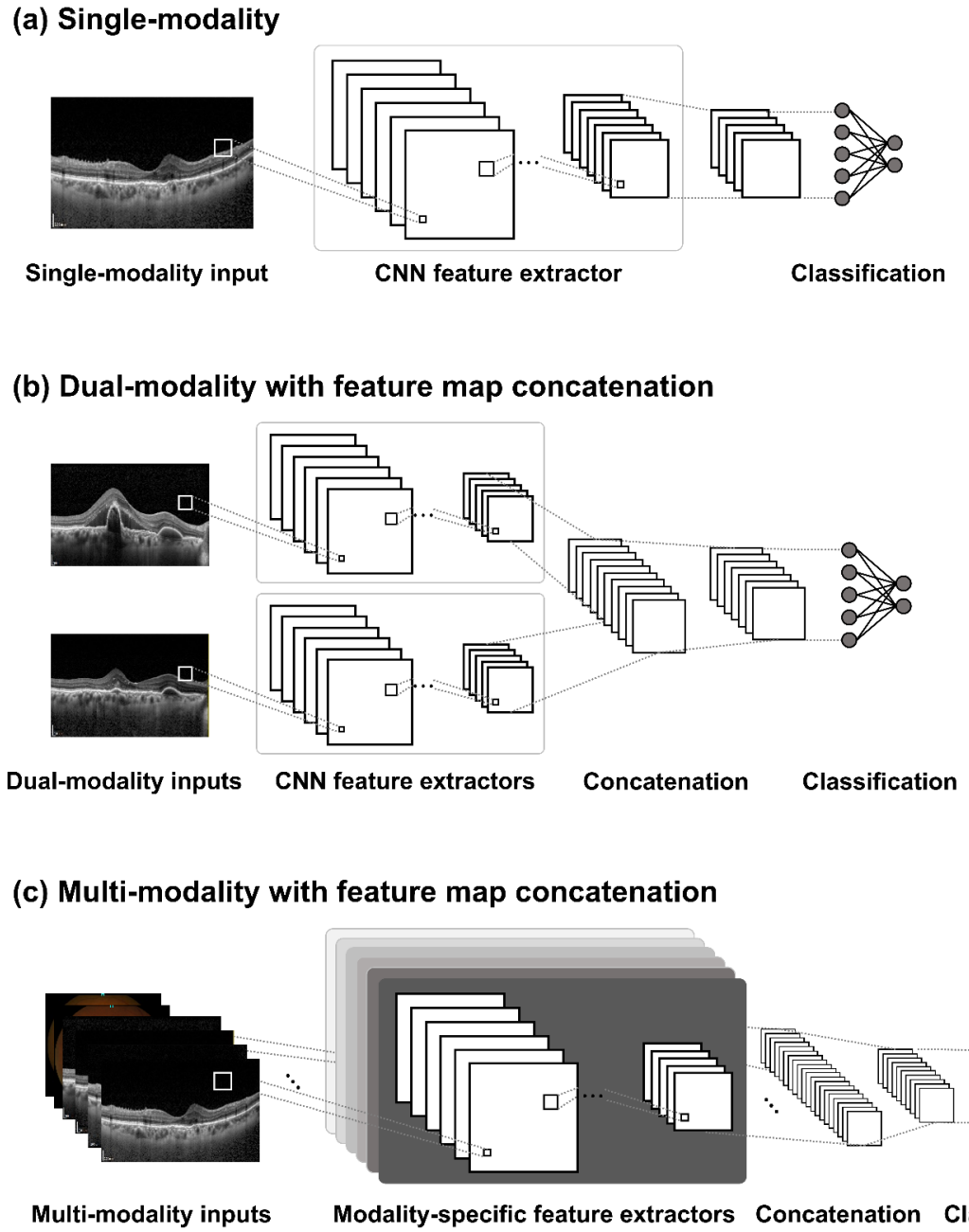


FIGURE 2. **Overview of the model architectures (a) Single-modality approach where a single image is used for recurrence prediction. (b) Dual-modality approach combining features from the two best-performing modalities through feature map concatenation. (c) Multi-modality approach utilizing all six images (Fundus pre-/post-treatment and OCT horizontal/vertical pre-/post-treatment) with feature map concatenation to improve classification performance**

We incorporated temperature scaling into the double-threshold framework to calibrate the model's output probabilities. Without calibration, raw logits often yield poorly calibrated probabilities, necessitating an exhaustive grid search for threshold hyperparameter tuning. Temperature scaling smooths the output distribution, mitigating the model's sensitivity to threshold variations and reducing the computational burden of grid search in cross-validation settings. After calibration, fixed upper and lower probability thresholds were applied to filter predictions; instances with predicted probabilities falling between these bounds were treated as “unknown” and excluded from the final evaluation. This exclusion mechanism helped mitigate the influence of

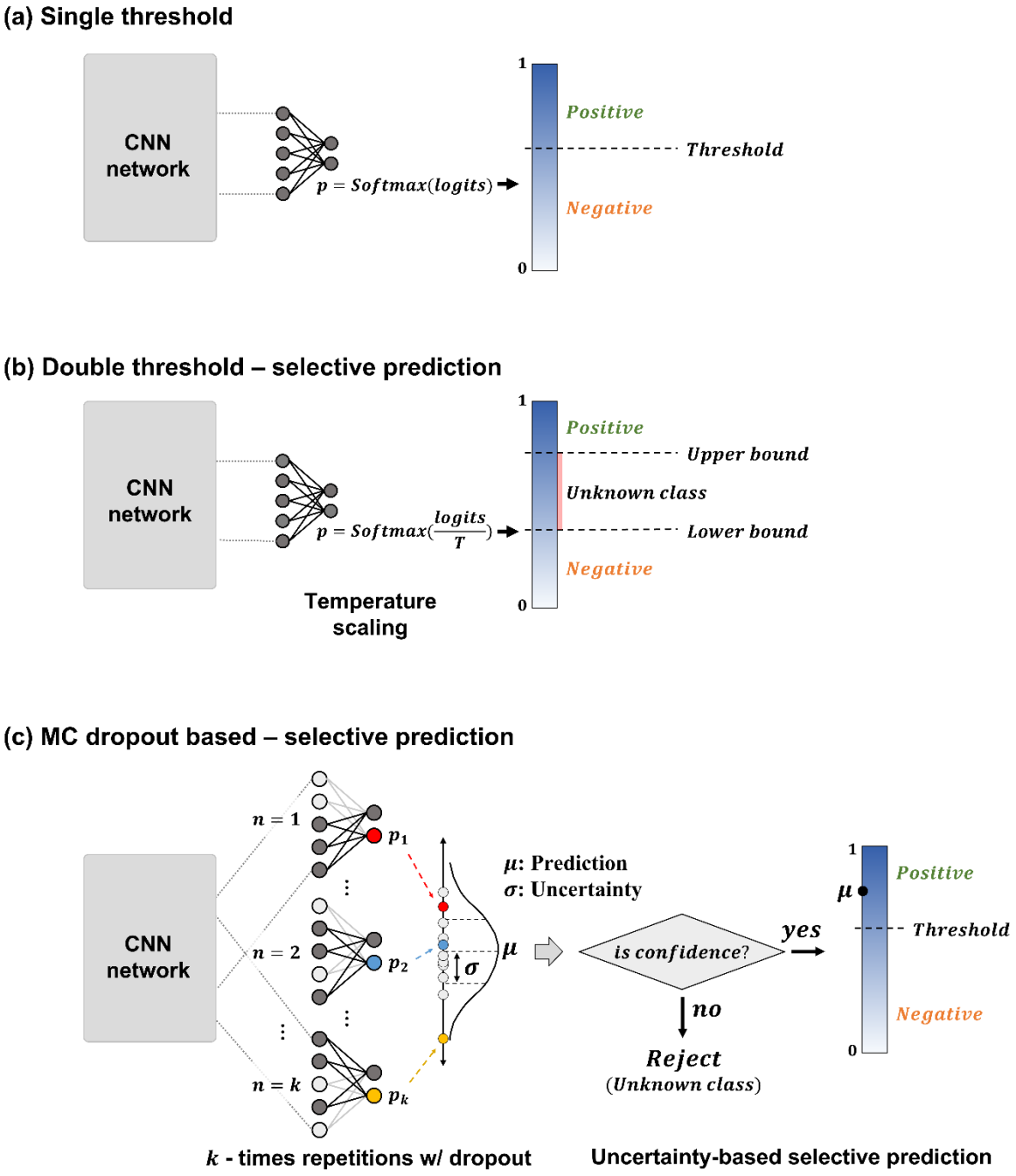


FIGURE 3. **Overview of classification and selective prediction strategies: (a) Conventional classification approach employing a single threshold to classify predictions into positive and negative classes. (b) Selective prediction approach with double thresholds (upper and lower). Predictions falling between the thresholds are classified as unknown and excluded from final evaluation metrics to improve result reliability. (c) MC dropout-based uncertainty estimation approach, each input image undergoes k stochastic forward passes. Predictions whose variance exceeds a specified threshold are excluded, and the mean of the remaining predictions is used to classify the image as positive or negative.**

uncertain predictions and enhanced the overall reliability of the results.

The second approach (Fig. 3(c)) leveraged the MC dropout to quantify the predictive uncertainty. While standard dropout is disabled during inference, MC dropout remains active to generate stochastic predictions for the same input. This controlled variability allows for uncertainty quantification in the decision-making process.

TABLE 2. Evaluation metrics of selective prediction methodologies (double threshold and MC dropout) applied to ophthalmic imaging datasets for wet AMD recurrence detection. Metrics include AUC, accuracy, sensitivity, specificity, F1 score, and exclusion rates (% and count). Results are averaged over 5-fold cross-validation

Input	Dataset	AUC	Accuracy	Sensitivity	Specificity	F1 score
Single-modality	Fundus pre-treatment	0.624 (±0.032)	0.587 (±0.036)	0.468 (±0.120)	0.772 (±0.130)	0.569 (±0.088)
	Fundus post-treatment	0.635 (±0.059)	0.614 (±0.065)	0.694 (±0.265)	0.489 (±0.254)	0.655 (±0.151)
	OCT horizontal pre-treatment	0.653 (±0.057)	0.439 (±0.106)	0.142 (±0.283)	0.900 (±0.169)	0.236 (±0.334)
	OCT horizontal post-treatment	0.662 (±0.046)	0.592 (±0.071)	0.571 (±0.285)	0.624 (±0.259)	0.588 (±0.171)
	OCT vertical pre-treatment	0.638 (±0.039)	0.629 (±0.039)	0.834 (±0.187)	0.314 (±0.259)	0.722 (±0.067)
	OCT vertical post-treatment	0.626 (±0.036)	0.572 (±0.093)	0.508 (±0.297)	0.667 (±0.266)	0.656 (±0.078)
Dual-modality	OCT horizontal pre-/post-treatment	0.631 (±0.020)	0.586 (±0.065)	0.652 (±0.293)	0.486 (±0.314)	0.614 (±0.181)
Multi-modality	All data	0.641 (±0.033)	0.589 (±0.079)	0.538 (±0.229)	0.668 (±0.194)	0.577 (±0.199)

For each input image, the model generated 30 independent predictions, each influenced by different dropout-induced perturbations. The variance of these predictions was computed as a measure of uncertainty. A high variance indicated substantial dispersion among predictions, suggesting greater model uncertainty, whereas a low variance reflected consistent outputs, implying higher confidence. To ensure reliable decision making, we applied a variance threshold (treated as a hyperparameter), excluding predictions that exceeded this threshold as highly uncertain. Predictions exceeding this threshold were deemed highly uncertain and were excluded
from further analyses. For the remaining cases, the final classification was determined by averaging 30 predictions, which yielded a more stable and robust decision. In our implementation, dropout was applied immediately before the final fully connected classification layer, following standard practice in CNN-based image classification models. This placement helps regularize high-level feature representations and mitigate overfitting.

G. EVALUATION OF THE MODEL AND EXPERIMENTAL SETTINGS

The model was trained and evaluated using 5-fold cross-validation, optimizing AUC within each training fold. And thresholds for additional evaluation metrics were determined post hoc using the Youden Index to achieve a balanced trade-off between sensitivity and specificity [49]. To minimize bias and ensure generalizability, a single global threshold—optimized on the aggregate validation results across all folds—was applied uniformly to all test predictions. This strategy avoids fold-specific threshold tuning and aligns with real-world clinical deployment, where a fixed decision threshold is typically required. The performance of the models was evaluated using the AUC.

Cross-entropy loss was used during training. Given that class imbalance was not severe and that cost-sensitive learning with class weights did not lead to meaningful performance improvements, class weighting was not applied. Models were trained using the Adam optimizer (learning rate: 0.01 or 0.001) and a StepLR scheduler (decay factor: 0.5 every 10 epochs). We evaluated batch sizes of 8, 16, 32, and 64. Training ran for 70–100 epochs, with early stopping triggered if the validation AUC did not improve for 25–30 epochs. All models were trained using the PyTorch framework on 64-bit systems. Experiments were conducted on two server configurations: one equipped with a Quadro RTX 8000 GPU paired with an Intel® Xeon® Gold 6226R CPU (2.90GHz, 16 cores) and the other equipped with a GeForce RTX 4090 GPU paired with an Intel® Xeon® Silver 4309Y CPU (2.80 GHz, 8 cores).

IV. RESULT

A. BASELINE PERFORMANCE ANALYSIS ON FULL DATASET (WITHOUT SELECTIVE PREDICTION)

In this subsection, we evaluated the performance of various CNN models on the entire dataset without applying any selective prediction strategies to establish a baseline. Various CNN models have been employed using the PyTorch library, including VGG16 [51], Inception-v3, ResNet50, ResNet152 [5], DenseNet121 [52], MobileNet-v3 [53], EfficientNet-b0, EfficientNet-v2, and ConvNext [54]. Based on five-fold cross-validated AUC performance, EfficientNet-v2 was selected for the Fundus pre-/post-treatment and OCT vertical pre-/post-treatment datasets; Inception-v3 for the OCT horizontal pre-treatment dataset; and EfficientNet-b0 for the OCT horizontal post-treatment dataset. As shown in Table 2, all resulting AUCs exceeded 0.6. Among single-modality inputs, OCT horizontal post-treatment achieved the highest AUC of 0.662 ± 0.046. The dual-modality model combining OCT horizontal pre- and post-treatment images achieved an AUC of 0.631 ± 0.020, while the multi-modality model incorporating all six imaging datasets achieved 0.641 ± 0.033. These baseline results highlight the inherent challenges in nAMD recurrence prediction and serve as a vital reference point for evaluating the effectiveness of the proposed selective prediction methods.

B. IMPACT OF DOUBLE THRESHOLDING AND MC DROPOUT ON PREDICTIVE RELIABILITY IN nAMD RECURRENCE

This section presents the results of applying selective prediction techniques, specifically the double-threshold method and MC dropout, to various ophthalmic imaging modalities for nAMD recurrence prediction. The analysis evaluated the effectiveness of these approaches in enhancing predictive reliability, defined here as the model's ability to focus on high-confidence predictions by excluding uncertain cases, and overall performance.

Table 3 shows that the double threshold method applied to the OCT horizontal post-treatment dataset achieved an AUC of 0.744 ± 0.064, excluding 69.42% of predictions as ambiguous, thereby improving the certainty of retained predictions. The MC dropout method demonstrated the highest overall performance for this dataset, with an AUC of 0.836 ± 0.163. Applying selective prediction methods, such as MC dropout and double thresholding, increased the overall AUC compared to the baseline model without selective prediction.

Among all datasets, the OCT horizontal post-treatment dataset achieved the highest results, underscoring the potential of single-image modalities in conjunction with selective prediction techniques for clinical applications in nAMD recurrence prediction.

C. UMAP ANALYSIS OF CLASSIFICATION CERTAINTY AND UNCERTAINTY WITH MC DROPOUT

Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimensionality-reduction technique that preserves both local and global structures while visualizing high-dimensional data [55]. In this study, we used UMAP to project the MC dropout prediction outcomes onto a two-dimensional plane, enabling a detailed examination of classification confidence and uncertainty. The UMAP visualization was constructed using the ground truth labels of the remaining samples after excluding uncertain predictions identified through MC Dropout.

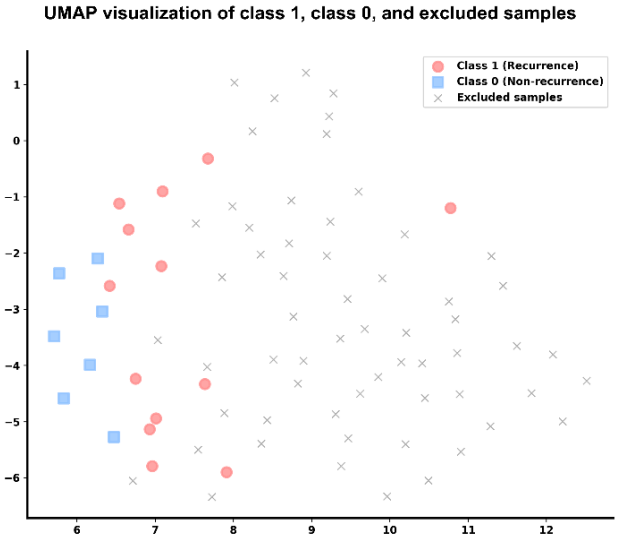


FIGURE 4. UMAP projection of classification. Red circles represent Class 1 (recurrence), blue squares represent Class 0 (non-recurrence), and gray crosses denote samples excluded due to high uncertainty identified by MC Dropout. Classification is based on ground truth labels.

Fig. 4 shows the UMAP projection of the classification outcomes for nAMD recurrence prediction, obtained using the MC dropout method on the Fold 4 validation dataset within a 5-fold cross-validation framework. The visualization reveals that Class 0 samples clustered predominantly to the left, indicating high prediction confidence, whereas Class 1 samples were primarily located to the right, forming a distinct boundary based on prediction certainty. In contrast, the excluded samples were scattered toward the right, indicating an elevated uncertainty. Notably, Fold 4 exhibits a more pronounced separation of predictive uncertainty compared to the other folds. In other folds, a higher proportion of excluded samples and less distinct separation between classes were observed, potentially due to inherent dataset variability or model uncertainty.

D. GRAD-CAM++ VISULIZATION ANALYSIS OF SELECTIVE PREDICTION IN nAMD CLASSIFICATION

TABLE 3. Performance outcomes derived from the highest-performing OCT horizontal post-treatment images (image-only) are compared with those obtained via a multimodal framework incorporating clinical parameters (medication, age, and gender)

Input	Selective prediction methodologies	AUC	Accuracy	Sensitivity	Specificity	F1 score	Excluded rate (%, count)
Fundus pre-treatment	Double threshold	0.703 (±0.081)	0.659 (±0.070)	0.602 (±0.239)	0.712 (±0.344)	0.643 (±0.127)	64.16% (51.20 ± 19.49)
	MC dropout	0.695 (±0.057)	0.644 (±0.057)	0.858 (±0.178)	0.377 (±0.382)	0.729 (±0.028)	60.65% (48.40 ± 7.42)
Fundus post-treatment	Double threshold	0.724 (±0.119)	0.566 (±0.226)	0.313 (±0.391)	0.920 (±0.160)	0.837 (±0.110)	73.93% (59.00 ± 4.38)
	MC dropout	0.707 (±0.149)	0.540 (±0.130)	0.434 (±0.356)	0.733 (±0.332)	0.549 (±0.258)	71.18% (56.80 ± 17.88)
OCT horizontal pre-treatment	Double threshold	0.700 (±0.047)	0.604 (±0.084)	0.602 (±0.341)	0.618 (±0.357)	0.595 (±0.211)	44.11% (35.20 ± 13.72)
	MC dropout	0.777 (±0.184)	0.493 (±0.088)	0.508 (±0.414)	0.550 (±0.452)	0.512 (±0.145)	81.20% (64.80 ± 17.21)
OCT horizontal post-treatment	Double threshold	0.744 (±0.064)	0.690 (±0.176)	0.584 (±0.300)	0.864 (±0.128)	0.773 (±0.072)	69.42% (55.40 ± 11.09)
	MC dropout	**0.836 (±0.163)**	0.717 (±0.182)	0.954 (±0.092)	0.400 (±0.490)	0.812 (±0.114)	78.95% (63.00 ± 11.78)
OCT vertical pre-treatment	Double threshold	0.716 (±0.070)	0.685 (±0.068)	0.651 (±0.337)	0.615 (±0.265)	0.764 (±0.054)	67.17% (53.60 ± 9.52)
	MC dropout	0.723 (±0.063)	0.665 (±0.109)	0.764 (±0.248)	0.531 (±0.327)	0.727 (±0.135)	66.17% (52.80 ± 6.97)
OCT vertical post-treatment	Double threshold	0.719 (±0.094)	0.618 (±0.171)	0.479 (±0.394)	0.806 (±0.177)	0.788 (±0.056)	65.41% (52.20 ± 6.11)
	MC dropout	0.746 (±0.079)	0.616 (±0.060)	0.865 (±0.186)	0.275 (±0.391)	0.731 (±0.045)	60.90% (48.60 ± 5.75)
OCT horizontal pre-/post-treatment	Double threshold	0.737 (±0.078)	0.580 (±0.120)	0.607 (±0.324)	0.619 (±0.344)	0.577 (±0.234)	63.41% (50.60 ± 10.03)
	MC dropout	0.746 (±0.079)	0.636 (±0.112)	0.707 (±0.378)	0.502 (±0.344)	0.780 (±0.031)	75.44% (60.20 ± 5.11)
All 6 images	Double threshold	0.682 (±0.085)	0.656 (±0.038)	0.755 (±0.171)	0.580 (±0.219)	0.688 (±0.045)	73.43% (58.60 ± 6.09)
	MC dropout	0.732 (±0.050)	0.613 (±0.139)	0.349 (±0.259)	0.944 (±0.048)	0.546 (±0.237)	70.43% (56.20 ± 2.93)

The Grad-CAM++ (Gradient-weighted Class Activation Mapping) technique [56] has been widely utilized in deep learning models to visualize salient regions influencing classification outcomes. We employed Grad-CAM++ to investigate which regions of the images were emphasized by the AI model when predicting recurrence, aiming to derive clinically interpretable insights. Fig. 5 illustrates the Grad-CAM++ analysis for the OCT horizontal post-treatment images, which exhibited the highest AUC performance and a sensitivity of 0.954 with MC dropout,
as reported in Table 3. A representative true positive case is presented to demonstrate the regions contributing to the prediction of recurrence. At baseline, the primary focus of the AI model on Grad-CAM++ images were the retina in 171 patients (42.9%), the choroid in 118 patients (29.6%), both the retina and choroid in 67 patients (16.8%), and regions outside

TABLE 4. Performance outcomes derived from the highest-performing OCT horizontal post-treatment images (image-only) are compared with those obtained via a multimodal framework incorporating clinical parameters (medication, age, and gender)

Methodologies	Input	AUC	Accuracy	Sensitivity	Specificity	F1 score	Excluded rate (%, count)
Conventional classification	Image only	0.662 (±0.046)	0.592 (±0.071)	0.571 (±0.285)	0.624 (±0.259)	0.588 (±0.171)	-
	Image + clinical	**0.675 (±0.040)**	0.644 (±0.042)	0.760 (±0.175)	0.465 (±0.207)	0.712 (±0.072)	-
MC dropout-based classification	Image only	0.836 (±0.163)	0.707 (±0.176)	0.954 (±0.092)	0.371 (±0.457)	0.805 (±0.112)	78.95% (63.00 ± 11.78)
	Image + clinical	**0.855 (±0.120)**	0.718 (±0.154)	0.941 (±0.118)	0.356 (±0.441)	0.813 (±0.099)	77.69% (62.00 ± 9.78)

TABLE 5. Comparison of classification performance metrics for anti-VEGF treatments using Ranibizumab and Aflibercept. In the experiment, 399 subjects were divided into two groups (Ranibizumab: 197; Aflibercept: 202) and evaluated using 5-fold cross-validation with OCT horizontal post-treatment images. The table summarizes performance metrics, including AUC, accuracy, sensitivity, specificity, and F1)

Methodologies	Anti-VEGF	AUC	Accuracy	Sensitivity	Specificity	F1 score	Excluded rate (%, count)
Conventional classification	Ranibizumab	0.703 (±0.070)	0.544 (±0.106)	0.389 (±0.383)	0.756 (±0.283)	0.390 (±0.287)	-
	Aflibercept	**0.709 (±0.023)**	0.653 (±0.052)	0.797 (±0.220)	0.411 (±0.359)	0.732 (±0.076)	-
MC dropout-based classification	Ranibizumab	0.934 (±0.044)	0.701 (±0.217)	0.498 (±0.377)	1.000 (±0.000)	0.713 (±0.284)	74.62% (29.40 ± 3.50)
	Aflibercept	**0.978 (±0.044)**	0.582 (±0.269)	0.792 (±0.340)	0.533 (±0.452)	0.644 (±0.269)	86.63% (35.00 ± 3.35)

these structures in 43 patients (10.8%). In post-treatment images, the primary focus of the AI model was the retina in 181 patients (45.4%), the choroid in 65 patients (16.3%), both the retina and choroid in 77 patients (19.3%), and elsewhere in 76 patients (19.0%), respectively.

E. CLINICAL AND MEDICATION DATA INTEGRATION: ADDITIONAL ANALYSIS

We conducted additional experiments to assess the impact of incorporating clinical information on the model performance, extending our analysis based on the OCT horizontal post-treatment model, which yielded the highest classification performance. Clinical variables, specifically medication type, age, and sex, were embedded and integrated into the model architecture. As shown in Table 4, the inclusion of clinical data led to an increase in the AUC compared with the baseline model using imaging data alone. Additionally, the MC

dropout-based analysis demonstrated a slight reduction in the number of excluded predictions and an improvement in the AUC. Notably, previous studies have reported that clinical outcomes may differ depending on the specific anti-VEGF agent administered during treatment [57]. Motivated by these observations, we further explored the role of medications by stratifying the dataset according to treatment type and conducting supplementary experiments.

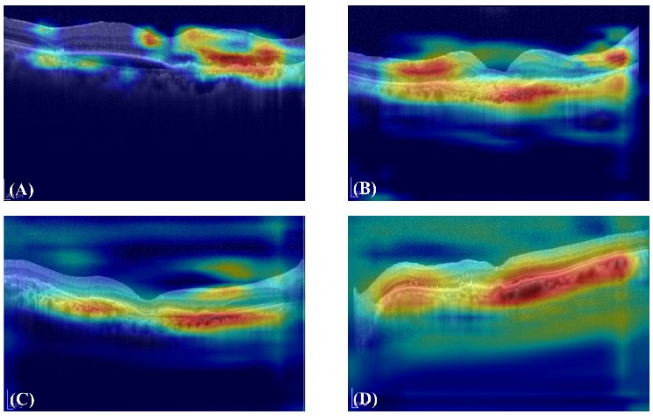


FIGURE 5. Grad-CAM++ visualization of OCT horizontal post-treatment images from the validation set, produced by the model exhibiting the highest AUC in Table 3. True positive predictions of recurrence are shown, with red regions representing the primary evidence the model used to make its decision.

Specifically, the dataset was divided into two subgroups based on the anti-VEGF agents administered: ranibizumab (n = 197) and aflibercept (n = 202). Each subgroup was independently analyzed using the original OCT horizontal post-treatment images, excluding other clinical features. As shown in Table 5, both subgroups exhibited improved predictive performances, with the MC dropout-based models achieving exceptionally high AUCs of 0.934 and 0.978 for the ranibizumab and aflibercept groups, respectively. These findings underscore the potential contribution of medication-specific characteristics to recurrence prediction and suggest that stratified modeling approaches combined with expanded training datasets may offer further performance gains.

TABLE 6. Characteristics of the independent temporal validation dataset (collected in July 2021).

Participant characteristics	Description
Number of participants	47
Gender, n (%)	Male: 26 (55.3%) Female: 21 (44.7%)
Age, mean ± SD	68.47 ± 9.65
Recurrence rate	72.3%

F. CLINICAL VALIDATION AND RETINA SPECIALISTS BENCHMARKING ON AN INDEPENDENT DATASET

To ensure the highest predictive robustness for clinical validation, we utilized the model configuration that achieved the highest AUC in Table 3—specifically the model based on horizontal post-treatment OCT scans. This best-performing architecture was retrained on the entire internal training and internal-validation cohorts to maximize its feature representation capability. The model's generalizability was then evaluated on a newly curated, independent temporal validation dataset (Table 6). This dataset was collected during a distinct period from the training cohort to ensure it functions as a "prospective-style" test set, even if retrospective in nature. This cohort consisted of 47 participants (mean age 68.47 ± 9.65 years; 55.3% male; 72.3% recurrence).

On this dataset, the model initially achieved an accuracy of 59.6% without selective prediction (Fig. 6(a)), already outperforming the senior and junior specialists who recorded 53.2% (Fig. 6(c)) and 53.2% (Fig. 6(d)), respectively. Notably, the experts showed limited ability in identifying non-recurrence cases, whereas the AI model maintained higher specificity. With the selective prediction strategy, performance on the high-certainty subset (N=21) significantly improved to 85.7% accuracy, with the F1-score rising from 0.708 to 0.903 (Fig. 6(b)). The confusion matrices illustrate that the uncertainty-aware mechanism effectively filters out ambiguous cases, reducing misclassifications to only three instances (1 FN, 2 FP). These results confirm that our model provides more robust and reliable predictions than clinical experts in a temporally independent validation setting.

The comparative analysis of performance metrics further underscores the model's robustness (Fig. 7). While clinical specialists achieved accuracies of 53.2% (senior) and 53.2% (junior), the base model reached 59.6%. Crucially, the selective prediction strategy elevated the accuracy to 85.7% and the F1-score to 0.90. As illustrated in the bar chart (Fig. 7), clinical experts exhibited notably low specificity (0.15 for senior and 0.08 for junior), indicating a strong tendency to over-diagnose recurrence. In contrast, the selective AI model achieved a more balanced specificity of 0.67, demonstrating its superior ability to accurately identify non-recurrence cases.

The discriminative power of the proposed framework is further validated by the ROC curves (Fig. 8). The area under the curve (AUC) improved from 0.57 to 0.70 upon implementing the selective prediction strategy. When the operating points of the specialists were plotted, both fell below the selective AI curve, characterized by high false-positive rates (1 - specificity). These results confirm that the uncertainty-aware model, by effectively deferring ambiguous samples, provides a more reliable and clinically discriminative assessment of nAMD recurrence than traditional expert evaluation in this challenging temporal validation cohort.

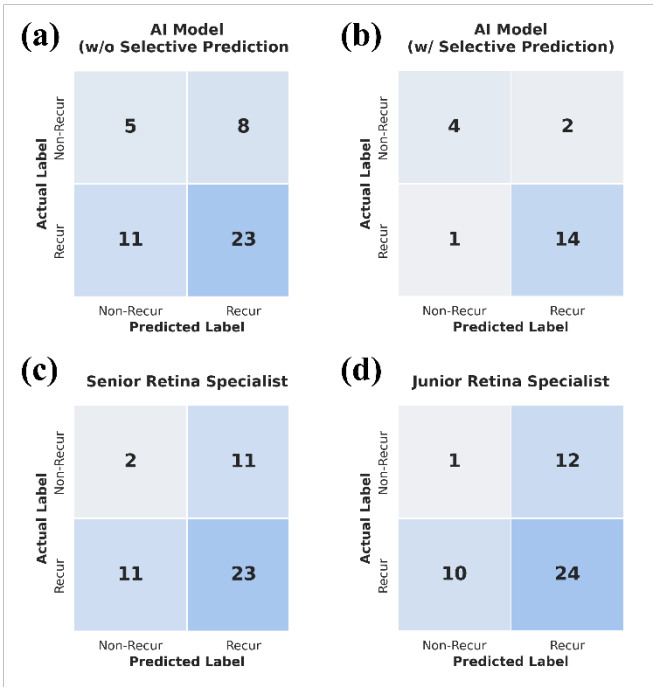


FIGURE 6. Confusion matrices for the independent temporal validation dataset (N=47). (a) AI model without selective prediction. (b) AI model with selective prediction (N=21). (c) Senior retina specialist. (d) Junior retina specialist. "Recur" and "Non-Recur" indicate the presence and absence of nAMD recurrence.

FIGURE 7. Comparison of performance metrics across AI configurations and clinical experts. Bars represent accuracy, sensitivity, specificity, and F1-score. The selective prediction model shows superior diagnostic balance, particularly in specificity, compared to both the base AI model and human specialists.

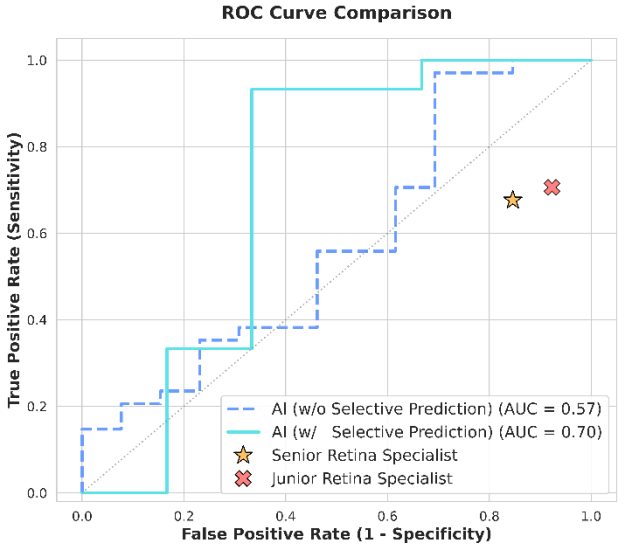


FIGURE 8. ROC curves for recurrence prediction performance. The proposed selective strategy enhances the discriminative power, increasing the AUC to 0.70. Specialist operating points are localized in the lower-specificity region, highlighting the model's advantage in reducing false-positive clinical assessments.

V. DISCUSSION

A. CLINICAL SIGNIFICANCE OF RECURRENCE PREDICTION

The proportion of patients who do not experience recurrence after initial loading injections varies depending on patient characteristics and follow-up duration, but is generally reported to be approximately 20–30% [11-13]. When implementing a TAE regimen, at least seven injections are typically administered during the first year. Thus, patients without recurrence may receive up to four unnecessary injections.

Anti-VEGF therapies are expensive, posing a financial burden, and intraocular injections can cause pain and anxiety [58, 59]. While the risk of complications is low, it cannot be ignored [60]. Since nAMD requires long-term management [61], optimizing treatment burden while maintaining therapeutic efficacy is essential. In this context, predicting recurrence after initial treatment holds considerable clinical value.

B. STRENGTHS AND CONTRIBUTIONS OUR PROPOSED METHOD

Despite the clinical significance of recurrence prediction in nAMD, few studies have tackled this problem using artificial intelligence. For instance, Jang et al. in [22] employed OCT-based fluid region analysis and reported an AUC of 0.725. However, their model was trained on data from multiple anti-VEGF agents and heterogeneous OCT devices, potentially limiting generalizability. Jung et al. in [21], on the other hand, used a DenseNet201-based architecture and achieved a prediction accuracy of 60.2%, only slightly outperforming ophthalmologists (53.3%).

Compared to these prior efforts, the present study offers several methodological advancements tailored for long-term prognosis. A key innovation is the adoption of uncertainty-aware selective prediction. While most existing AI models focus on "detection" (identifying current lesions), predicting "future recurrence" is a significantly more complex task that has consistently shown limited performance in conventional models. Our contribution lies in demonstrating that a well-designed selective prediction framework can bridge this gap, transforming a challenging prognostic task into a reliable clinical tool. By generating multiple stochastic predictions for each input and computing their variance, our model effectively quantifies prediction confidence. When applied to OCT horizontal post-treatment images, this strategy yielded an AUC of 0.836, surpassing those of previous studies and conventional thresholding approaches. Crucially, the performance variation observed across modalities—specifically, the superior results in OCT compared to fundus photography—aligns with clinical expectations regarding information disparity. This confirms that the model effectively prioritizes the depth information essential for fluid detection rather than overfitting to noise, a robustness further corroborated by the consistent trends in 5-fold cross-validation.

Unlike the double-threshold method, which relies on fixed probability cutoffs and lacks explicit uncertainty modeling, MC dropout implements a probabilistic framework that approximates Bayesian inference [36, 38]. This not only improves calibration but also mitigates the overconfidence often observed in deep learning models [50]. To further evaluate the clinical reliability of our proposed framework, we analyzed the probability calibration of our best-performing model (the OCT-based architecture in Table 3). By adopting temperature scaling as a post-hoc calibration technique, the model exhibited a strong alignment between its predictive confidence and the actual recurrence rates, as demonstrated in the reliability diagram (Fig. 9). This refinement yielded an

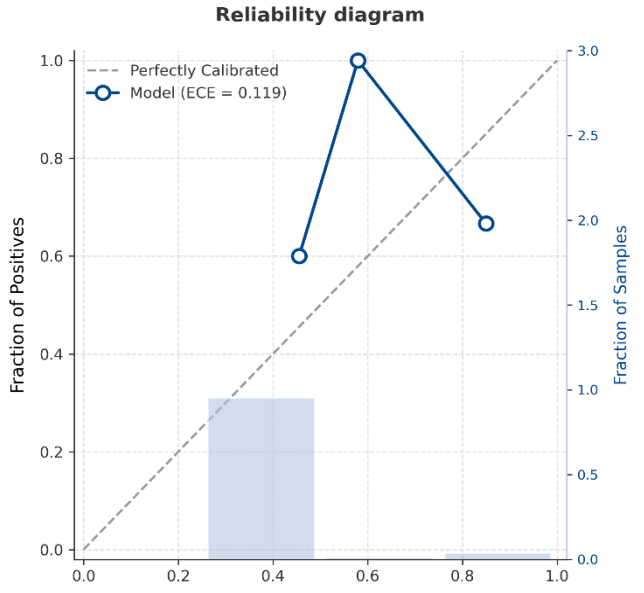


FIGURE 9. Reliability diagram for the best-performing nAMD recurrence prediction model. (a) Calibration curve showing the alignment between the mean predicted probability and the observed fraction of positives (ECE = 0.119). (b) The bottom histogram indicates the sample density across bins. The number of bins was set to 4 to ensure statistical stability given the concentrated distribution of predicted probabilities. The diagonal dashed line represents perfect calibration.

ECE of 0.119 and a Brier score of 0.247. Notably, the majority of samples (80 out of 84) were concentrated in the 0.25–0.50 probability range, where the model showed a cautious under-confidence profile (mean predicted probability of 0.455 vs. observed fraction of 0.600). Since temperature scaling is a monotonic transformation, it strictly preserved the superior discriminative performance—including the AUC, accuracy, and recall reported in Table 3—while ensuring that the output probabilities serve as a trustworthy measure of clinical uncertainty. While previous models reported accuracies and AUCs in the 0.5 to approximately 0.7 range [21, 22], our selective approach enabled robust performance by rejecting low-confidence predictions—a strategy that is especially valuable for clinical applications where decision reliability is critical.

In addition, we explored drug-specific modeling—a dimension rarely addressed in previous AI studies. There are slight differences in the incidence and timing of lesion reactivation between ranibizumab and aflibercept [62, 63]. The VEGF suppression efficacy of aflibercept is known to last longer than that of ranibizumab, extending beyond 4 months in some cases [5]. From a technical perspective, such clinical distinctions introduce statistical heterogeneity into the dataset; thus, drug-specific modeling was implemented to enable the AI to learn specialized feature representations tailored to each treatment's unique dynamics. Previous studies have shown no significant difference in lesion reactivation between ranibizumab and aflibercept after initial treatment, but the incidence of reactivation and the duration until the first reactivation tended to be slightly longer in the aflibercept group. [63] Despite the findings of previous studies, to date, no study investigating AI-based prediction of lesion reactivation has employed drug-specific modeling. Our drug-specific models achieved exceptionally high AUCs of 0.934 for ranibizumab and 0.978 for aflibercept, highlighting the importance of treatment-specific modeling. These findings also imply that models trained on one therapeutic context may not generalize to another, underscoring the need for tailored model development as newer agents like brolucizumab and faricimab enter clinical practice.

Together, these innovations—including selective prediction, uncertainty quantification, and treatment-specific modeling—suggest that the present study may contribute to the development of AI-based recurrence prediction for nAMD.

C. CLINICAL BENCHMARKING AND TEMPORAL GENERALIZABILITY

The benchmarking results against retina specialists (Fig. 7, 8) provide a critical perspective on the gap between AI-driven analysis and human clinical judgment. A notable finding was the remarkably low specificity observed in both senior (0.15) and junior (0.08) specialists. This suggests a significant "over-diagnosis" bias in current clinical practice, where experts tend to predict recurrence to avoid the potential risk of missing a treatment window. While this cautious approach is intended to prioritize patient safety, it inherently contributes to the aforementioned clinical and financial burdens associated with unnecessary anti-VEGF therapy.

Our model, particularly when augmented with the selective prediction strategy, demonstrated a more balanced diagnostic profile. While human experts struggled to identify non-recurrence cases, the selective AI model achieved a specificity of 0.67 on the independent validation set. This objective discriminative power is further highlighted in the ROC analysis (Fig. 8), where the specialists' operating points fell below the AI's curve. Furthermore, the fact that the selective prediction strategy improved the AUC from 0.57 to 0.70 on a temporally separated dataset—which typically presents a "dataset shift" due to evolving clinical environments—proves the robustness of our uncertainty-aware framework. This indicates that the model does not merely memorize training patterns but effectively learns to identify ambiguous cases that are prone to human error.

D. LIMITATION AND FUTURE DIRECTIONS

First, while this study is primarily a single-center, retrospective investigation, we sought to enhance the reliability of our findings by performing independent temporal validation (Table 6). Previous studies have emphasized that reliance on a single dataset with limited diversity can adversely affect the performance and robustness of machine learning models. To address this, we conducted an evaluation using a chronologically distinct dataset, which provided a "prospective-style" assessment of the model's generalizability. Although this temporal validation demonstrated that our

uncertainty-aware framework maintains superior predictive power and specificity compared to clinical experts—even when accounting for potential dataset shifts over time—the model's performance across different institutional protocols and diverse ethnic populations warrants further verification. Therefore, while our results demonstrate the potential of selective prediction as a pilot investigation, the findings should be interpreted within the context of these single-center constraints.

Second, only horizontal and vertical OCT scans were used, omitting raster scans that cover a wider retinal area. the dataset was limited to participants of Korean ethnicity, which may introduce potential demographic bias and restrict applicability across diverse populations.

Third, the current selective prediction framework exhibited relatively high exclusion rates (60–85%) to achieve high predictive reliability. While this "conservative" approach ensures decision safety in high-stakes ophthalmic prognosis, it may limit the model's immediate utility for all patients. However, since the exclusion thresholds are tunable hyperparameters, they can be optimized to balance coverage and accuracy depending on specific clinical requirements. In this study, we addressed model-driven uncertainty by incorporating probability calibration via temperature scaling for our best-performing model, which demonstrated a well-calibrated ECE of 0.119. Future work will focus on further reducing these exclusion rates by incorporating larger, multi-center datasets and exploring even more advanced calibration architectures to build upon the reliable probabilistic framework established here.

Fourth, review of the Grad-CAM++ images showed that the majority of cases demonstrated predominant model attention to the retinal and/or choroidal regions. However, the potential impact of these differences in attention patterns on the model's predictive performance could not be conclusively determined in the present study, and this remains an area warranting further investigation in future studies with more rigorous and targeted study designs.

MC dropout-based uncertainty is also influenced by the model architecture. As shown in Table 3, the reported values should be interpreted as model-specific trends or reference levels, rather than used for absolute comparisons across different backbones.

Future studies should involve multi-center cohorts, expanded imaging modalities, and ethnically diverse populations to validate and enhance the model's robustness.

VI. CONCLUSION

In conclusion, we developed an AI-based model for predicting recurrence of nAMD after initial anti-VEGF loading injections. The incorporation of uncertainty-aware selective prediction and drug-specific modeling was associated with improved predictive performance. However, considering the limitations of this study, further studies are warranted to enhance the generalizability of our findings.

ACKNOWLEDGMENT

FUNDING: This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF2023R1A2C2003577) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02304331, Digital Columbus Project)

CODE AVAILABILITY: The source code used to develop the models and perform the experiments in this study is publicly available at: https://github.com/MIHLabCHA/Wet-AMD-recurrence

REFERENCES

[1] S. R. Flaxman *et al.*, "Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis," (in eng), *Lancet Glob Health,* vol. 5, no. 12, pp. e1221-e1234, Dec 2017.

[2] M. Fleckenstein, S. Schmitz-Valckenberg, and U. Chakravarthy, "Age-Related Macular Degeneration: A Review," (in eng), *Jama,* vol. 331, no. 2, pp. 147-157, Jan 9 2024.

[3] J. Q. Li, T. Welchowski, M. Schmid, M. M. Mauschitz, F. G. Holz, and R. P. Finger, "Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis," (in eng), *Br J Ophthalmol,* vol. 104, no. 8, pp. 1077-1084, Aug 2020.

[4] S. Obata, M. Kakinoki, O. Sawada, I. Kawamoto, M. Murase, and M. Ohji, "Duration of Vascular Endothelial Growth Factor Suppression After Intravitreal Injection of Brolucizumab and Aflibercept in Macaque Eyes," (in eng), *J Ocul Pharmacol Ther,* vol. 39, no. 3, pp. 225-228, Apr 2023.

[5] S. Fauser and P. S. Muether, "Clinical correlation to differences in ranibizumab and aflibercept vascular endothelial growth factor suppression times," (in eng), *Br J Ophthalmol,* vol. 100, no. 11, pp. 1494-1498, Nov 2016.

[6] P. J. Rosenfeld *et al.*, "Ranibizumab for neovascular age-related macular degeneration," (in eng), *N Engl J Med,* vol. 355, no. 14, pp. 1419-31, Oct 5 2006.

[7] J. S. Heier *et al.*, "Intravitreal aflibercept (VEGF trap-eye) in wet age-related macular degeneration," (in eng), *Ophthalmology,* vol. 119, no. 12, pp. 2537-48, Dec 2012.

[8] A. E. Fung *et al.*, "An optical coherence tomography-guided, variable dosing regimen with intravitreal ranibizumab (Lucentis) for neovascular age-related macular degeneration," (in eng), *Am J Ophthalmol,* vol. 143, no. 4, pp. 566-83, Apr 2007.

[9] K. B. Freund *et al.*, "TREAT-AND-EXTEND REGIMENS WITH ANTI-VEGF AGENTS IN RETINAL DISEASES: A Literature Review and Consensus Recommendations," (in eng), *Retina,* vol. 35, no. 8, pp. 1489-506, Aug 2015.

[10] M. Augsburger, G. M. Sarra, and P. Imesch, "Treat and extend versus pro re nata regimens of ranibizumab and aflibercept in neovascular age-related macular degeneration: a comparative study," (in eng), *Graefes Arch Clin Exp Ophthalmol,* vol. 257, no. 9, pp. 1889-1895, Sep 2019.

[11] J. H. Kim, Y. S. Chang, J. W. Kim, C. G. Kim, and D. W. Lee, "Long-term incidence and timing of reactivation in patients with type 3 neovascularization after initial treatment," (in eng), *Graefes Arch Clin Exp Ophthalmol,* vol. 257, no. 6, pp. 1183-1189, Jun 2019.

[12] J. H. Kim, J. W. Kim, and C. G. Kim, "Difference in Lesion Reactivation between Pure Type 2 and Mixed Type 1 and 2 Macular Neovascularization and its Influence on Long-Term Treatment Outcomes," (in eng), *Semin Ophthalmol,* vol. 38, no. 4, pp. 358-364, May 2023.

[13] Y. Kuroda *et al.*, "Factors Associated with Recurrence of Age-Related Macular Degeneration after Anti-Vascular Endothelial Growth Factor Treatment: A Retrospective Cohort Study," (in eng), *Ophthalmology,* vol. 122, no. 11, pp. 2303-10, Nov 2015.

[14] S. Moon *et al.*, "Prediction of anti-vascular endothelial growth factor agent-specific treatment outcomes in neovascular age-related macular degeneration using a generative adversarial network," (in eng), *Sci Rep,* vol. 13, no. 1, p. 5639, Apr 6 2023.

[15] A. Heinke *et al.*, "ARTIFICIAL INTELLIGENCE FOR OPTICAL COHERENCE TOMOGRAPHY ANGIOGRAPHY-BASED DISEASE ACTIVITY PREDICTION IN AGE-RELATED MACULAR DEGENERATION," (in eng), *Retina,* vol. 44, no. 3, pp. 465-474, Mar 1 2024.

[16] M. Rohm *et al.*, "Predicting Visual Acuity by Using Machine Learning in Patients Treated for Neovascular Age-Related Macular Degeneration," (in eng), *Ophthalmology,* vol. 125, no. 7, pp. 1028-1036, Jul 2018.

[17] X. Li *et al.*, "Multi-modal multi-instance learning for retinal disease recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2474-2482.

[18] L. Hamid, A. Elnokrashy, E. H. Abdelhay, M. M. J. N. C. Abdelsalam, and Applications, "A deep learning LSTM-based approach for AMD classification using OCT images," *Neural Computing and Applications,* vol. 36, no. 31, pp. 19531-19547, 2024.

[19] W. Wei, J. Southern, K. Zhu, Y. Li, M. F. Cordeiro, and K. J. S. R. Veselkov, "Deep learning to detect macular atrophy in wet age-related macular degeneration using optical coherence tomography," *Scientific Reports,* vol. 13, no. 1, p. 8296, 2023.

[20] J. H. Tan *et al.*, "Age-related macular degeneration detection using deep convolutional neural network," *Future Generation Computer Systems,* vol. 87, pp. 127-135, 2018.

[21] J. Jung *et al.*, "Prediction of neovascular age-related macular degeneration recurrence using optical coherence tomography images with a deep neural network," (in eng), *Sci Rep,* vol. 14, no. 1, p. 5854, Mar 11 2024.

[22] B. Jang, S. Y. Lee, C. Kim, U. C. Park, Y. G. Kim, and E. K. Lee, "Preliminary analysis of predicting the first recurrence in patients with neovascular age-related macular degeneration using deep learning," (in eng), *BMC Ophthalmol,* vol. 23, no. 1, p. 499, Dec 7 2023.

[23] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. J. M.-R. Fu, "A review of uncertainty estimation and its application in medical imaging," *Meta-Radiology 1.1,* p. 100003, 2023.

[24] A. Kurz *et al.*, "Uncertainty estimation in medical image classification: systematic review," *JMIR Medical Informatics,* vol. 10, no. 8, p. e36427, 2022.

[25] A. Swaminathan *et al.*, "Selective prediction for extracting unstructured clinical data," vol. 31, no. 1, pp. 188-197, 2024.

[26] M. M. Abdullahi, S. Chakraborty, P. Kaushik, and B. S. Sami, "Detection of dry and wet age-related macular degeneration using deep learning," in *2nd International Conference on Industry 4.0 and Artificial Intelligence (ICIAI 2021)*, 2022, pp. 211-214: Atlantis Press.

[27] A. Serener and S. Serte, "Dry and wet age-related macular degeneration classification using oct images and deep learning," in *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)*, 2019, pp. 1-4: IEEE.

[28] C. Dominguez *et al.*, "Binary and multi-class automated detection of age-related macular degeneration using convolutional-and transformer-based architectures," vol. 229, p. 107302, 2023.

[29] X. Xu *et al.*, "Joint segmentation of retinal layers and fluid lesions in optical coherence tomography with cross-dataset learning," vol. 162, p. 103096, 2025.

[30] L. Álvarez-Rodríguez *et al.*, "Fully automatic deep convolutional approaches for the screening of neurodegeneratives diseases using multi-view OCT images," vol. 158, p. 103006, 2024.

[31] B. Fazekas *et al.*, "SD-LayerNet: Robust and label-efficient retinal layer segmentation via anatomical priors," vol. 261, p. 108586, 2025.

[32] F. Li *et al.*, "Global-Local Transformer Network for Automatic Retinal Pathological Fluid Segmentation in Optical Coherence Tomography Images," vol. 266, p. 108772, 2025.

[33] M. Chen *et al.*, "Automated diagnosis of age-related macular degeneration using multi-modal vertical plane feature fusion via deep learning," *Medical Physics,* vol. 49, no. 4, pp. 2324-2333, 2022.

[34] D. Oh *et al.*, "GCN-assisted attention-guided UNet for automated retinal OCT segmentation," *Expert Systems with Applications,* vol. 249, p. 123620, 2024.

[35] H. Jiang, M. Gao, J. Huang, C. Tang, X. Zhang, and J. Liu, "DCAMIL: Eye-tracking guided dual-cross-attention multi-instance learning for refining fundus disease detection," *Expert Systems with Applications,* vol. 243, p. 122889, 2024.

[36] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050-1059: PMLR.

[37] Y. Geifman and R. J. A. i. n. i. p. s. El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems,* vol. 30, 2017.

[38] B. Lakshminarayanan, A. Pritzel, and C. J. A. i. n. i. p. s. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems,* vol. 30, 2017.

[39] T. DeVries and G. W. J. a. p. a. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv,* 2018.

[40] R. Challen *et al.*, "Artificial intelligence, bias and clinical safety," *BMJ quality & safety* vol. 28, no. 3, pp. 231-237, 2019.

[41] B. Nam, J. Y. Kim, I. Y. Kim, and B. H. J. J. M. I. Cho, "Selective prediction with long short-term memory using unit-wise batch standardization for time series health data sets: algorithm development and validation," *JMIR Medical Informatics,* vol. 10, no. 3, p. e30587, 2022.

[42] Y. Ding *et al.*, "Uncertainty-aware training of neural networks for selective medical image segmentation," in *Medical Imaging with Deep Learning*, 2020, pp. 156-173: PMLR.

[43] A. Krizhevsky, I. Sutskever, and G. E. J. A. i. n. i. p. s. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.

[44] K. J. Choi *et al.*, "Deep learning models for screening of high myopia using optical coherence tomography," *Scientific reports,* vol. 11, no. 1, p. 21663, 2021.

[45] H.-g. Kim, S. Song, B. H. Cho, and D. P. J. P. o. Jang, "Deep learning-based stress detection for daily life use using single-channel EEG and GSR in a virtual reality interview paradigm," *Plos one,* vol. 19, no. 7, p. e0305864, 2024.

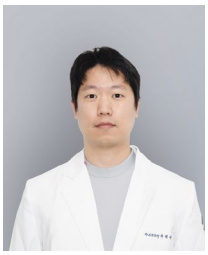
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[47] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105-6114: PMLR.

[48] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, 2021, pp. 10096-10106: PMLR.

[49] E. F. Schisterman, N. J. Perkins, A. Liu, and H. J. E. Bondell, "Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples," *Epidemiology,* vol. 16, no. 1, pp. 73-81, 2005.

[50] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, 2017, pp. 1321-1330: PMLR.
[51] K. J. a. p. a. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv,* 2014.
[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.
[53] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314-1324.
[54] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976-11986.
[55] L. McInnes, J. Healy, and J. J. a. p. a. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv,* 2018.
[56] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018, pp. 839-847: IEEE.
[57] F. Semeraro, F. Morescalchi, S. Duse, E. Gambicorti, A. Cancarini, and C. J. C. d. m. Costagliola, "Pharmacokinetic and pharmacodynamic properties of anti-VEGF drugs after intravitreal injection," *Current drug metabolism,* vol. 16, no. 7, pp. 572-584, 2015.
[58] E. Moisseiev, Y. L. Tsai, and M. Herzenstein, "Treatment of Neovascular Age-Related Macular Degeneration: An Economic Cost-Risk Analysis of Anti-Vascular Endothelial Growth Factor Agents," (in eng), *Ophthalmol Retina,* vol. 6, no. 3, pp. 205-212, Mar 2022.
[59] C. Yiallouridou, J. H. Acton, S. Banerjee, H. Waterman, and A. Wood, "Pain related to intravitreal injections for age-related macular degeneration: a qualitative study of the perspectives of patients and practitioners," (in eng), *BMJ Open,* vol. 13, no. 8, p. e069625, Aug 16 2023.
[60] B. Bodaghi, E. H. Souied, R. Tadayoni, M. Weber, A. Ponthieux, and L. Kodjikian, "Detection and Management of Intraocular Inflammation after Brolucizumab Treatment for Neovascular Age-Related Macular Degeneration," (in eng), *Ophthalmol Retina,* vol. 7, no. 10, pp. 879-891, Oct 2023.
[61] S. Chandra *et al.*, "Ten-year outcomes of antivascular endothelial growth factor therapy in neovascular age-related macular degeneration," (in eng), *Eye (Lond),* vol. 34, no. 10, pp. 1888-1896, Oct 2020.
[62] J. H. Kim, J. W. Kim, C. G. J. J. o. O. P. Kim, and Therapeutics, "Five-year reactivation after ranibizumab or aflibercept treatment for neovascular age-related macular degeneration and polypoidal choroidal vasculopathy," vol. 37, no. 9, pp. 525-533, 2021.
[63] J. H. Kim, Y. S. Chang, D. W. Lee, C. G. Kim, and J. W. Kim, "Incidence and Timing of the First Recurrence in Neovascular Age-Related Macular Degeneration: Comparison Between Ranibizumab and Aflibercept," (in eng), *J Ocul Pharmacol Ther,* vol. 33, no. 6, pp. 445-451, Jul/Aug 2017.



WON TAE YOON received the M.D. degree from Hallym University College of Medicine, Chuncheon, South Korea, in 2012. He is currently a Retina Specialist at Kim's Eye Hospital, Seoul, South Korea, where he is engaged in clinical practice and vitreoretinal surgery. His research interests include age-related macular degeneration and applications of artificial intelligence in ophthalmology. He has published about ten papers in SCI and SCIE-indexed international journals.

Hun-gyeom Kim received the B.S. degree in biomedical engineering from Hankuk University of Foreign Studies, Yongin, Republic of Korea, in 2022, and the M.S. degree in biomedical engineering from Hanyang University, Seoul, in 2024. His research interests include signal processing, medical imaging, and artificial intelligence for medical data analysis. He is particularly interested in multi-modal learning and the application of large language models (LLMs) to healthcare data for disease prediction and clinical decision support.

JUNHYUNG MOON is a Research Professor in the Department of Biomedical Informatics, CHA University School of Medicine, CHA University, Republic of Korea. He received his B.S. and Ph.D. degrees in Computer Science from Yonsei University, Seoul, Republic of Korea. His research focuses on artificial intelligence–based detection, estimation, prediction, and intervention using multimodal data for digital healthcare and precision medicine. He has authored more than 20 papers in peer-reviewed international journals and conference proceedings.

Dat Ngo received the B.S. degree in Computer Engineering from the University of Danang—University of Science and Technology, Danang, Vietnam, in 2016, and the M.S. and Ph.D. degrees in Electronics Engineering from Dong-A University, Busan, South Korea, in 2018 and 2022, respectively. He was a Research Professor with CHA University, Pangyo, South Korea, from 2022 to 2024. Since 2024, he has been with the Department of Computer Engineering, Korea National University of Transportation, Chungbuk, South Korea, where he is currently an Assistant Professor. His research interests include image/video processing, machine learning, deep learning, and FPGA/VLSI/SoC design.

Dr. Jae Hui Kim is the Director of the Clinical Research Center at Kim's Eye Hospital in Seoul, South Korea. He graduated from Seoul National University College of Medicine and obtained his Master's degree from Sungkyunkwan University. He completed his residency and retina fellowship at Samsung Medical Center. His primary research interest is age-related macular degeneration. Through his work in this field, he has received academic awards from the Korean Ophthalmological Society, the Korean Retina Society, and the Alumni Association of the Department of Ophthalmology at Seoul National University College of Medicine. He is an active member of several international societies, including the Macula Society.

Baek Hwan Cho received the B.S. degree in electronics and telecommunication engineering and the M.S. and Ph.D. degrees in biomedical engineering from Hanyang University, Seoul, Republic of Korea, in 1999, 2001, and 2007, respectively. He is currently an Associate Professor with the Department of Biomedical Informatics, CHA University School of Medicine, and the Deputy Director of the Institute of Biomedical Informatics. He also leads the AI and Big Data Research Team at the Data Science Research Center, CHA Medical Research Institute. He was previously an Associate Professor at Samsung Medical Center and a Visiting Professor at the Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University. His research interests include medical data–driven artificial intelligence, disease prognosis, and diagnostic modeling.