Fundus Image-based Visual Acuity Assessment with PAC-Guarantees

Sooyong Jang SOOYONG@SEAS.UPENN.EDU
Kuk Jin Jang JANGKJ@SEAS.UPENN.EDU
Hyonyoung Choi HYOUNCHOI@SEAS.UPENN.EDU
Computer and Information Science, University of Pennsylvania, USA
Yong-Seop Han MEDCABIN@HANMAIL.NET
Department of Ophthalmology, Gyeongsang National University College of Medicine, Republic of Korea
Seongjin Lee INSIGHT@GNU.AC.KR
Jin-hyun Kim JIM.KIM@GNU.AC.KR
Department of AI Convergence Engineering, Gyeongsang National University, Republic of Korea
Insup Lee LEE@SEAS.UPENN.EDU
Computer and Information Science, University of Pennsylvania, USA

Abstract

Timely detection and treatment are essential for maintaining eye health. Visual acuity (VA), which measures the clarity of vision at a distance, is a crucial metric for managing eye health. Machine learning (ML) techniques have been introduced to assist in VA measurement, potentially alleviating clinicians' workloads. However, the inherent uncertainties in ML models make relying solely on them for VA prediction less than ideal. The VA prediction task involves multiple sources of uncertainty, requiring more robust approaches. A promising method is to build prediction sets or intervals rather than point estimates, offering coverage guarantees through techniques like conformal prediction and Probably Approximately Correct (PAC) prediction sets. Despite the potential, to date, these approaches have not been applied to the VA prediction task. To address this, we propose a method for deriving prediction intervals for estimating visual acuity from fundus images with a PAC guarantee. Our experimental results demonstrate that the PAC guarantees are upheld, with performance comparable to or better than that of two prior works that do not provide such guarantees.

Keywords: Visual Acuity Prediction, Fundus Images, Prediction Intervals, PAC Guarantees, Uncertainty Quantification

Data and Code Availability The Visual Acuity dataset was obtained from a previous study (Kim et al., 2022) which consists of fundus images and corresponding visual acuity labels. This dataset is not publicly available due to restrictions on sharing patient images, but it is available from the corresponding author upon reasonable request and permission of the Institutional Review Board. Code is available at this code repository. [1]

Institutional Review Board (IRB) The appropriate approval was acquired for the study by the IRB at the institution where the data was collected. The procedures used in this study followed the principles of the Declaration of Helsinki. The requirement for informed patient consent was waived by the IRB due to the retrospective nature of the study. Additional details will be provided in the camera-ready version.

1. Introduction

In eye health management, early detection and timely treatment are crucial. Deep learning-based models have the potential to support large-scale screening programs, helping to detect abnormalities, support clinicians, and enable earlier diagnosis for individuals. However, the inherent uncertainty in machine learning predictions challenges the effectiveness of such models in safety-critical applications like large-scale screening and clinician decision-support systems.

A common approach to addressing uncertainty is to predict a set (or interval) of labels with a Probably Approximately Correct (PAC) guarantee (Valiant,

1. https://github.com/precise-ai4oph/va_pred_pac

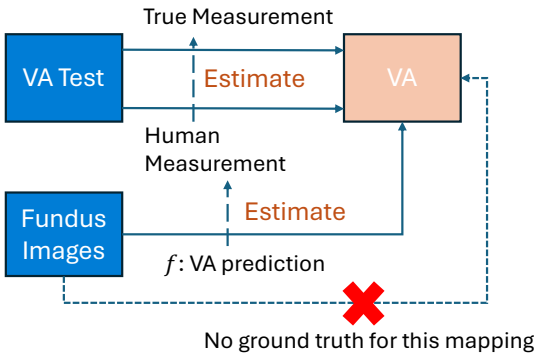


Figure 1: Uncertainties in VA prediction: Although VA prediction is based on fundus images, the ground truth acuity measurements were not obtained from these images. The VA test (e.g., using a Snellen chart) measures visual acuity, and the VA regressor estimates value measured by a human. Additionally, the human measurement process itself introduces uncertainty.

1984) or to employ Conformal Prediction (Vovk et al., 2005), both of which provide formal guarantees on the reliability of predictions. Machine learning approaches for analyzing fundus images—a type of retinal imaging—have been limited in providing formal guarantees compared to the vast amount of prior work in that field, which lacks such assurances (Ayhan et al., 2020; Zhou et al., 2022; Rahaman et al., 2021; Ge and Wang, 2021; Filos et al., 2019; Leibig et al., 2017; Ayhan et al., 2022; Gulshan et al., 2016).

The situation is similar for visual acuity (VA) prediction, where the goal is to predict a subject's visual acuity based on fundus images. The VA prediction task introduces uncertainties from multiple sources, as illustrated in Figure 1. First, the ground truth acuity value is obtained from a separate VA test conducted interactively with a human examiner and not derived directly from fundus images. Additionally, the VA test itself introduces variability and deviations from the true VA level (Siderov and Tiu, 1999; Veselỳ and Synek, 2012). Given these various uncertainties, it is crucial to use prediction sets (or intervals) with guarantees for this task. However, to the best of our knowledge, no previous studies have addressed this issue.

Two previous studies (Kim et al., 2022; Paul et al., 2023) propose models to estimate visual acuity from fundus images. Both studies evaluate their models on their respective datasets and demonstrate good performance. However, neither study provides any guarantees, which limits their clinical applications, such as in the aforementioned screening programs.

In this work, we address this limitation by constructing a prediction interval and providing a Probably Approximately Correct (PAC) guarantee in our model's predictions. In detail, we train a regressor to estimate the visual acuity from fundus images, and derive a prediction interval $C(x_i)$ on example (x_i, y_i) with a PAC guarantees on coverage, *i.e.*,

$$\mathbb{P}_{Z^n \sim \mathcal{D}^n}\left[\mathbb{P}_{(x_i,y_i)\sim\mathcal{D}}\left[y_i \in C(x_i)\right] \geq 1-\epsilon\right] \geq 1-\delta, \tag{1}$$

where Z^n is the validation set, $\mathcal{D}$ is the data distribution, $\mathcal{D}^n = \mathcal{D} \times \cdots \times \mathcal{D}$, and ϵ is an error bound and δ is a significance level. Our experiments demonstrate that our approach achieves the target coverage while maintaining a sufficient average width of the prediction interval. In summary, our contributions are as follows:

1. We propose a method for generating prediction intervals with example-dependent widths, ensuring a PAC guarantee by combining the well-established approach of training a model that outputs a Gaussian distribution with the PAC prediction interval technique.

2. We apply the aforementioned prediction interval method on a visual acuity dataset and empirically demonstrate that for a coverage bound $\epsilon = 0.3$ and a significance level, $\delta = 0.001$ %, we achieve a coverage rate of 71.49 % with average interval width of 3.04.

The remainder of this paper is organized as follows: We discuss related work in Section 2, describe our method in Section 3, and present our experimental results in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

In this section, we review the relevant literature. We begin by discussing two key studies focused on visual acuity prediction from fundus images. Following this, we discuss prior work on prediction guarantees, with an emphasis on conformal prediction and PAC prediction sets.

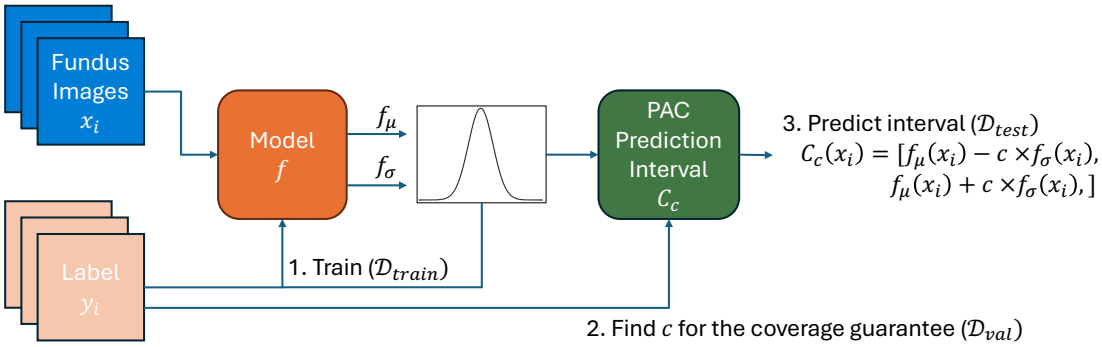


Figure 2: Overall process: First, we train a Gaussian Distribution output model with the training data. Next, we find c for the coverage bound with the validation data. Lastly, we compute the prediction intervals using the test data.

2.1. VA prediction from fundus images

The Visual acuity (VA) prediction task is to estimate VA from fundus images. VA can be represented in various ways, such as a decimal number (*e.g.*, 0.5, 1.0), a fraction (*e.g.*, 20/40, 20/20), or a letter score (*e.g.*, 60, 35).

The VA prediction has been explored in two recent studies (Kim et al., 2022; Paul et al., 2023). In (Kim et al., 2022), the problem was formulated as a classification task. The original eleven visual acuity levels (ranging from 0.0 to 1.0 in increments of 0.1) were mapped into four levels based on an ophthalmologist's guidelines. They then employ a two-stage approach, with different classifiers at each stage. In contrast, Paul et al. (2023) treat the problem as a regression task, leveraging the ordinality of visual acuity labels to predict visual acuity letter scores using various models such as ResNet50, ConvNeXt, EfficientNetV2, and Swin Transformers. Both approaches demonstrate strong performance in their evaluations. However, neither study provides any guarantees on the reliability of their results.

2.2. Conformal Prediction and PAC Guarantees

One approach to providing guarantees on model predictions is to construct prediction intervals that, with high probability, include the true label. This can be achieved using methods such as Conformal Prediction (Vovk et al., 2005) and PAC Prediction sets (Vovk, 2012; Park et al., 2019). The two methods are similar, but offer different types of guarantees. Conformal prediction generates a prediction interval that contains the ground truth for a test data point with high probability. In contrast, the PAC prediction set offers guarantees conditioned on the training data.

In VA prediction, an important consideration when providing guarantees on prediction intervals is the need for varying interval widths depending on the difficulty of each example. Ideally, easy examples (for the model) should have narrow prediction intervals, while more difficult examples should have wider intervals. There has been extensive research in conformal prediction to address this need, using techniques such as scalar estimated uncertainty and locally adaptive conformal prediction (Papadopoulos et al., 2008; Romano et al., 2020; Angelopoulos and Bates, 2021; Gibbs and Candes, 2021; Seedat et al., 2023). These approaches has been employed across various machine learning applications (Straitouri et al., 2023; Ji et al., 2023; Park et al., 2022a).

When applying these techniques to the VA prediction task, a critical aspect is constructing intervals that have clinically useful widths. Specifically, the interval width must be narrow enough to be practical for VA prediction tasks. Although these techniques aim to minimize the interval width (or set size) while ensuring coverage, they do not address the practical requirements for interval width in clinical settings.

3. Method

Our approach first trains a regression model for predicting VA from fundus images. The output is used to derive a prediction interval of the estimate with PAC guarantees. The overall process is illustrated in Figure 2.

3.1. Background - PAC Prediction Interval

The PAC prediction interval (or set) constructs a prediction interval, $C(x)$ with a Probably Approximately Correct (PAC) guarantee, as represented in Equation (1). The PAC guarantee implies that the prediction error is small ("Approximately Correct", described by the inner probability with ϵ), and holds with a high probability ("Probably", represented by the outer probability with δ) as long as the test data follows the same distribution $\mathcal{D}$ as the training and calibration data. In our context, "Approximately" means that the prediction interval includes the true visual acuity with a coverage rate of at least $1-\epsilon$, while "Probably" indicates that this coverage rate bound generally holds as long as all data—training, calibration, and test—are drawn from the same distribution, $\mathcal{D}$.

3.2. Regression model learning

Let $\mathcal{X}$ be the set of fundus images. We train a model $f : \mathcal{X} \rightarrow \mathbb{R}^2$, to predict visual acuity for a given fundus image. We model the VA prediction with a Gaussian distribution, where the mean represents the predicted visual acuity, and the standard deviation indicates the uncertainty of the prediction. The two model output values correspond to the mean ($\mu \in \mathbb{R}$) and standard deviation ($\sigma \in \mathbb{R}_{>0}$). That is, for the i^{th} fundus image x_i, $f(x_i) = (f_\mu(x_i), f_\sigma(x_i))$, where $f_\mu(x_i)$ and $f_\sigma(x_i)$ predict the mean and standard deviation, respectively.

To train this model, we use the Negative Log Likelihood (NLL) loss. The standard deviation represents the prediction uncertainty and will be used to construct the prediction intervals, as illustrated in the following section.

3.3. Prediction Interval with PAC guarantee

From the output, we build a prediction interval for VA, $C(x_i) = [C_l, C_u]$, that should contain the true value with high probability, *i.e.*, $y_i \in C(x_i)$. Our goal is to derive the prediction interval that contains the ground truth with the minimum width, *i.e.*, high coverage and narrow width, that satisfies the following PAC guarantee (Equation (1)).

With a constant c which controls the prediction interval width satisfying Equation (1), we derive a prediction interval $C_c(x_i)$ for (x_i, y_i),

$$C_c(x_i) = [C_{c,l}(x_i), C_{c,u}(x_i)], \tag{2}$$

where $C_{c,l}(x_i) = f_\mu(x_i) - c \times f_\sigma(x_i), C_{c,u}(x_i) = f_\mu(x_i) + c \times f_\sigma(x_i)$, with a constant c which satisfies Equation (1). Our interval uses the estimated standard deviation ($f_\sigma(\cdot)$) for each example, allowing each example to have a different prediction interval based on its estimated uncertainty (standard deviation).

We determine c by solving the following optimization problem as described in (Park et al., 2019):

$$c^* = \arg\min_c c \quad \text{subject to } \underline{c} \geq 1-\epsilon,$$

where $[\underline{c}, \overline{c}]$ is the Clopper-Pearson interval for $W = \{\mathbb{1}(y_i \in C_c(x_i)) | (x_i, y_i) \in Z\}$ with the significance level δ. The prediction interval C_{c^*} satisfies the PAC guarantee (Park et al., 2019, 2022a).

4. Experiment

To evaluate our approach, we train models according to Section 3.2 and compute prediction intervals satisfying the PAC guarantee. The experiments are repeated five times with different seeds, and we report the mean and standard deviations across the repetitions.

4.1. Dataset

For training and evaluation, we use a fundus dataset obtained from existing work (Kim et al., 2022). The dataset consists of 54,781 fundus images labeled with visual acuity levels ranging from 0.0 to 1.0 in increments of 0.1, which were obtained through a visual acuity assessment. The dataset uses categorized integer values from 0 to 10 (0, 1, 2, . . . , 10) to represent visual acuity levels ranging from 0.0 to 1.0, considering visual acuity prediction as a classification task. We utilize these categorized labels for our regression task. The details of the data distribution and related discussion are provided in Appendix A.

The dataset is randomly divided into training, validation, and test sets in a 6:2:2 ratio. To evaluate the robustness, we create five different dataset splits of the dataset using different random seeds.

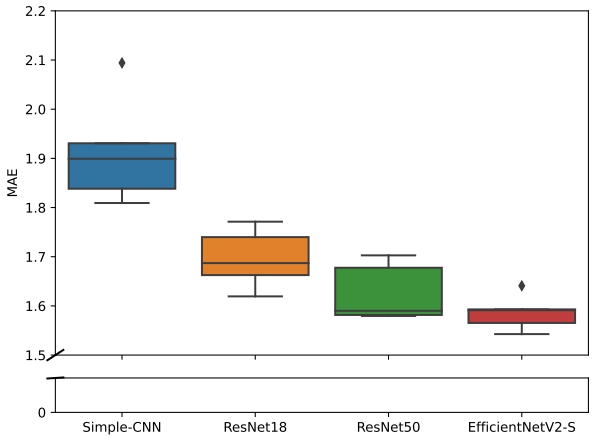


Figure 3: MAE of models over 5 repetitions

4.2. Models

We employ four different base models: Simple-CNN, ResNet18, ResNet50 (He et al., 2016), and EfficientNetV2-S (Tan and Le, 2021). All models have the final fully connected layer with two output nodes for the two Gaussian distribution parameters. The Simple-CNN model that we implemented comprises two convolutional layers followed by three fully connected layers. All layers use ReLU activation functions, except for the final layer. For the other models, we use pre-implemented versions available in PyTorch, modifying the final output layers to match the required dimensions. The ResNet and EfficientNet models are initialized with pre-trained weights provided by PyTorch, while the Simple-CNN is randomly initialized. After the initialization, the models are trained using NLL loss as described in Section 3.

4.3. Results

First, we present the performance of our regression model and with respect to the prediction intervals. In the subsequent section, we compare our results to prior work.

4.3.1. Regression model

The mean absolute error (MAE) of point predictions is displayed in Figure 3. Based on the result, EfficientNetV2-S shows the best result with an average of 1.54, and the Simple-CNN shows the worst with an average of 1.88. While these MAEs might be acceptable considering the inherent inaccuracies in the (manual) assessment of visual acuity (Siderov and Tiu, 1999; Veselỳ and Synek, 2012), it is important to note that the error can vary significantly for individual subjects because the models do not provide any guarantees.

4.3.2. PAC prediction intervals

We compute a prediction interval with PAC guarantee according to Section 3. We utilize a range of $\epsilon \in [0.2, 0.3, 0.4]$ for the coverage guarantee (*i.e.* Coverage bound: $1 - \epsilon$) with a significance level $\delta =$ 0.001 %, and compute the coverage and the average interval width as shown in Figure 4. Our method consistently satisfies the coverage bound for all ϵ values. All models show higher coverage than the given bound for all configurations.

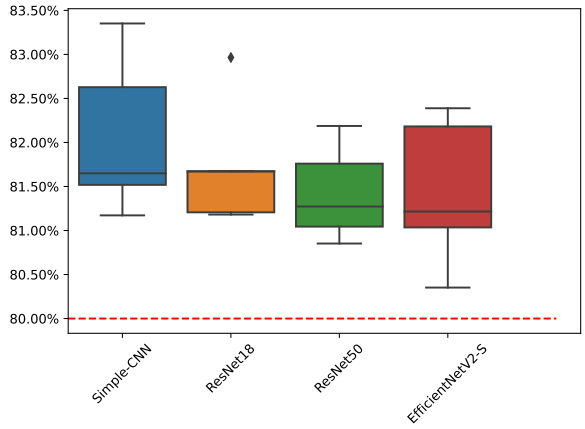
Based on the results, EfficientNetV2-S generates the narrowest prediction intervals, while Simple-CNN produces the widest. This variation may be partially due to differences in model complexity. More complex models, like EfficientNetV2-S, tend to learn intricate patterns in the data, allowing them to make more confident predictions with lower estimated standard deviation. As a result, these models produce narrower intervals. In contrast, simpler models, such as Simple-CNN, may have difficulty capturing such patterns, leading to greater uncertainty and wider prediction intervals.

Specifically, the average width for EfficientNetV2-S is around 3.0, when coverage rate bound is 70 % ($\epsilon = 0.3$), However, for practical usage with this 70 % coverage, we require a slightly narrower width, around 2, which aligns with the variability in VA measurement by humans (Veselỳ and Synek, 2012).

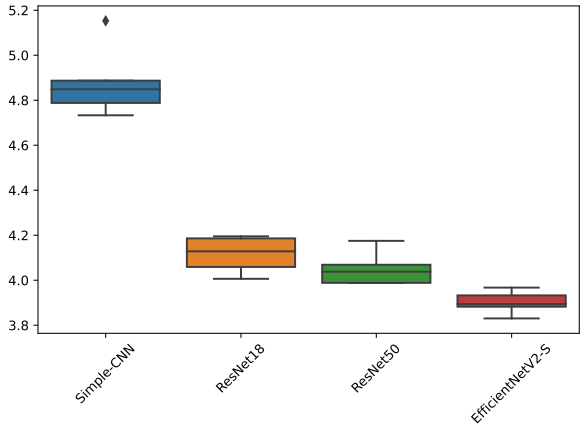
4.4. Comparison to additional baselines

We implement a Bayesian Neural Network (BNN) for ResNet18 and ResNet50 based on the code from the repository (Krishnan et al., 2022) and vanilla conformal prediction as other baselines.

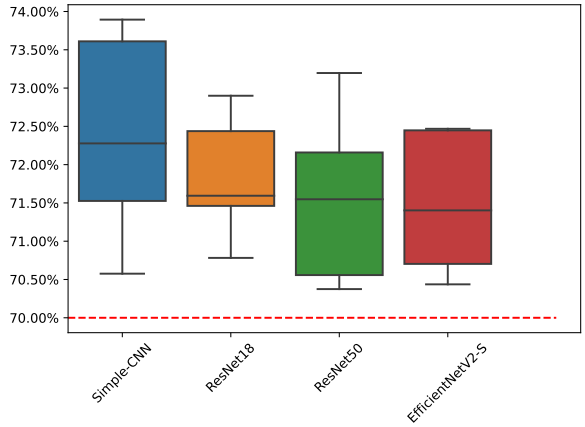
First, we compare the BNN performance for the point estimates with models from Section 3.2. We compute the macro-averaged MAE to account for the dataset imbalance and compare the point estimate performance between our base model and the BNNs. This metric represents the average MAE for each ground truth value. As shown in Table 1, both BNN versions show similar point estimate performance compared to their deterministic counter-



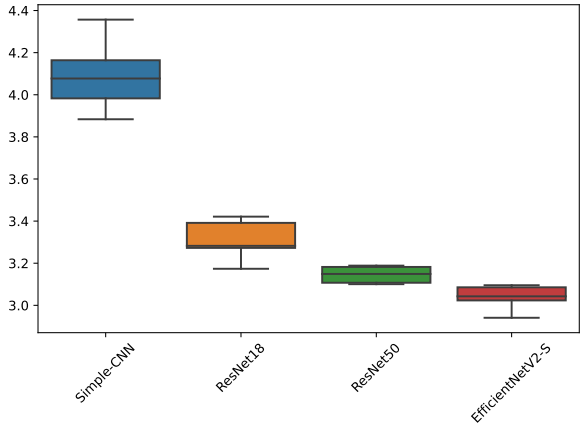
(*a*) Coverage, $\epsilon = 0.2$



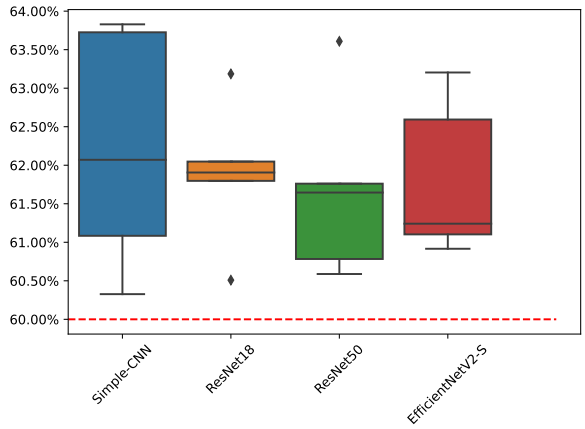
(*b*) Width, $\epsilon = 0.2$



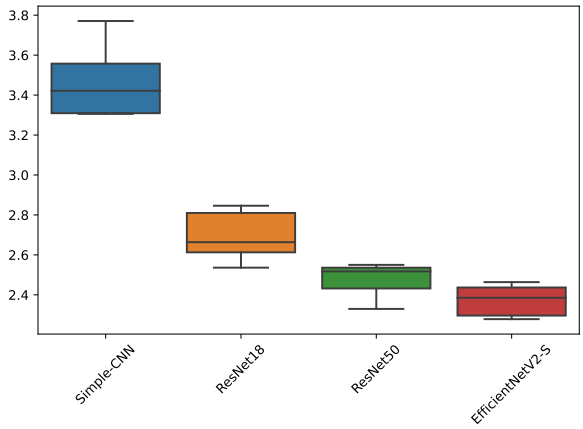
(*c*) Coverage, $\epsilon = 0.3$



(*d*) Width, $\epsilon = 0.3$



(*e*) Coverage, $\epsilon = 0.4$



(*f*) Width, $\epsilon = 0.4$

Figure 4: PAC Guarantee Analysis Results (significance level $\delta = 0.001$ %). The left column displays the coverage, while the right column shows the average width. In the left column figures, the red dotted line represents the coverage bound. For all ϵ and base models, our result show that the coverage bound is satisfied.

Table 1: Point Estimate Performance Comparison

Method	Base Model	Macro-Averaged MAE
Ours	Simple-CNN	2.51 ± 0.13
	ResNet18	2.10 ± 0.06
	ResNet50	2.14 ± 0.07
	EfficientNetV2-S	2.12 ± 0.05
BNN	ResNet18	2.17 ± 0.06
	ResNet50	2.29 ± 0.07

parts. However, it is essential to note that the BNN approach does not provide a probabilistic coverage guarantee like our method. To compare prediction interval performance, we perform Monte Carlo sampling to generate multiple predictions and compute the interval based on the 0.5 and 99.5 quantiles. This enables us to then calculate coverage and interval widths. Since intervals from the BNN model do not ensure a coverage rate bound, the empirical coverage rate is not relevant to the expected quantile percentage (99%).

We also implement a vanilla conformal prediction (VCP) with the nonconformity score, $|\hat{y}_i - y_i|$. The results are presented in Table 2. BNN interval widths are comparable to those of our models; however, they do not provide the guaranteed coverage. VCP, designed to ensure coverage, does not always exceed the target rate of 70% due to small fluctuations, which is expected (Angelopoulos et al., 2023). While the average interval width is comparable to our model's, this approach produces the same interval width for all examples. In contrast, our method provides adaptive widths based on the estimated standard deviation.

4.5. Comparisons to previous studies

We compare our approach with other methods for estimating visual acuity, as described in (Kim et al., 2022; Paul et al., 2023). These two approaches differ in their label scheme and the dataset, making a fair comparison challenging. Nevertheless, we have made every effort to ensure the most accurate comparison possible.

4.5.1. Comparison with (Kim et al., 2022)

There are four main differences in the settings, a) Class scheme (4-level vs. 11-level), b) Dataset split (balanced test-set vs. imbalanced test-set), c) Prediction Type (Point vs. Interval), and d) Repetition (single experiments vs. repeated experiments). Considering the differences, we first map the visual acuity levels into 4 categories, as described by the authors in their paper (see Table 3). Next, we compute the macro-average accuracy (MA-ACC) of our prediction intervals on our imbalanced test set and compare it with the accuracy of their point predictions on the balanced test set. For our MA-ACC calculation, we consider the interval correct when it contains a true label. In addition, we set $\epsilon = 0.2$ for our intervals as per their final accuracy (82.4 %).

The result is shown in Table 4. Our MA-ACC of prediction intervals is comparable with (Kim et al., 2022) with a width of around 4.37. In addition to comparable performance, we have the advantage of providing a guarantee on the coverage. However, we note that a fair comparison is challenging due to differences in settings.

It should be noted that our MA-ACC in the table may be below 80 % (1 - ϵ), because the guarantee is not targeted for this metric. This metric is computed by macro-averaging classwise accuracy for the comparison to the (Kim et al., 2022), while the prediction

Table 2: Vanilla Conformal Prediction Performance

Method	Base Model	Coverage	Width
Ours	Simple-CNN	72.38 ± 1.25	4.09 ± 0.16
	ResNet18	71.83 ± 0.75	3.31 ± 0.09
	ResNet50	71.57 ± 1.04	3.15 ± 0.04
	EfficientNetV2-S	71.49 ± 0.85	3.04 ± 0.06
BNN	ResNet18	49.14 ± 2.69	3.03 ± 0.06
	ResNet50	51.57 ± 4.12	3.11 ± 0.31
VCP	SimpleCNN	70.04 ± 0.71	3.94 ± 0.16
	ResNet18	69.72 ± 0.89	3.24 ± 0.09
	ResNet50	69.69 ± 1.13	3.07 ± 0.09
	EfficientNetV2-S	69.43 ± 1.17	2.96 ± 0.07

Table 3: VA level mapping in (Kim et al., 2022). Top row shows 11-level classes, bottom row shows corresponding 4-level classes.

11-level	0	1	2	3	4	5	6	7	8	9	10
4-level	0	1		2					3		

Table 4: Accuracy comparison with (Kim et al., 2022).

Model	Metric	
	MA-ACC	Width
Simple-CNN	73.09 ± 5.63	6.05 ± 0.17
ResNet18	79.27 ± 1.31	4.72 ± 0.07
ResNet50	80.58 ± 2.66	4.54 ± 0.15
EfficientNetV2-S	80.77 ± 2.03	4.37 ± 0.15
Kim et al. (2022)	82.4	-

Table 5: Error distribution comparison with (Paul et al., 2023). Each number shows the percentage of data in the error range.

Model	Error Range		
	[0,5]	[6,10]	[11,]
Simple-CNN	58.09 ± 1.70	18.41 ± 1.66	23.50 ± 0.14
ResNet18	60.97 ± 1.08	16.16 ± 0.59	22.87 ± 0.88
ResNet50	60.56 ± 0.23	15.19 ± 0.54	24.25 ± 0.67
EfficientNetV2-S	61.55 ± 0.76	15.14 ± 0.27	23.31 ± 0.68
Paul et al. (2023)	33	28	39

interval is computed to satisfy the coverage bound on the whole dataset. Because of this difference, some of the computed MA-ACC in the table can be lower than the guaranteed coverage. However, an appropriate interval could be computed if targeting the same metric.

4.5.2. Comparison with (Paul et al., 2023)

We note two major differences from our settings in terms of a) Dataset and b) Visual Acuity Score System (Letter score vs. Fraction of Snellen chart). Although the datasets are different, we compare the two results based on the error distributions reported in their paper. In (Paul et al., 2023), they present the distribution of errors by computing the percentage of test examples that fall into three error ranges in the VA letter score: between 0 and 5 ([0, 5]), between 6 and 10 ([6, 10]), and greater than 10 ([11,]). We also compute the errors of our predictions using the same VA letter score and compare the error distributions across these ranges. The difference in visual acuity scores was addressed using the conversion formula.[2]

As shown in Table 5, our models have more examples in the small error range([0,5]) compared to the model in (Paul et al., 2023) Again, based on these results, our model is at least comparable to or better than the model presented in (Paul et al., 2023) while providing coverage guarantees.

4.6. Discussion

As shown in Section 4.3.2, our results indicate that the more complex model (EfficientNetV2-S) demonstrates better accuracy and produces narrower prediction intervals compared to less complex models. This suggests that more complex models may offer improved performance in terms of prediction interval width. We believe that models like EfficientNetV2-L and RETFound (Zhou et al., 2023) are promising candidates for future work due to their higher complexity. In particular, as RETFound is a foundation model specifically designed for retinal images, it may be an especially suitable option.

Another point to improve the prediction intervals in terms of the their width - narrower width, we examine the estimated standard deviation ($f_\sigma(x_i)$) of examples as our prediction interval width is derived

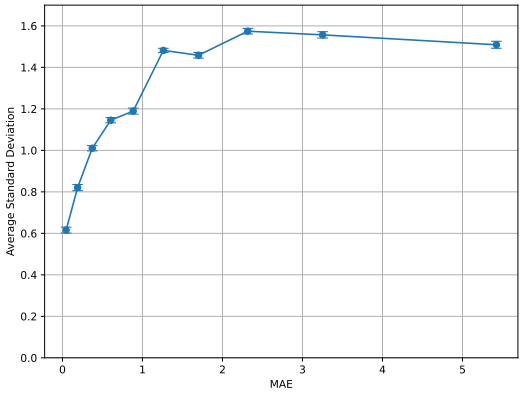


Figure 5: MAE and standard deviations: Absolute errors are binned by equal mass, and average error and standard deviations are computed for each bin, plotted with a 95 % confidence interval.

2. The conversion formula between the letter score (L) and the fraction (F) is: $L = 85 + 50 \times \log_{10} F$.

from the standard deviations ($|C_c(x_i)| = 2 \times c \times f_\sigma(x_i)$). If the estimated standard deviation for each example are lower, we can achieve narrower prediction intervals.

Figure 5 illustrates the relationship between the prediction error and estimated standard deviation for EfficientNetV2-S model using a random seed of 100. We compute both the prediction error ($|y_i - f_\mu(x_i)|$) and estimated standard deviation, then divide the prediction error space into five bins with equal mass, *i.e.*, each bins contains the same number of examples. For each bin, we calculate the average absolute prediction error and the average standard deviation, which are then plotted in the figure.

The figure reveals a positive correlation between prediction error and standard deviation, indicating that examples with lower errors has the lower standard deviations. This observation suggests that for the subset of examples where the given model performs accurately, lower standard deviations can be achieved, resulting in narrower prediction intervals, because the prediction intervals are based on the predicted standard deviation. It may be possible to identify such subsets of examples with low prediction errors, enabling us to provide narrower prediction intervals for the subset.

One significant challenge in applying Conformal Prediction and PAC based approaches in practice is the potential violation of their distributional assumptions. Specifically, the training (or calibration) data and test data may come from different distributions, a phenomenon known as dataset shift (Quiñonero-Candela et al., 2022). To address this, several strategies have been proposed. Detection algorithms (Jang et al., 2022; Liu et al., 2020) can be employed to identify such shifts, enabling model retraining. Alternatively, adaptation algorithms (Si et al., 2024; Park et al., 2022b) can adjust the model to account for the shifted distribution. These approaches can be applied to the visual acuity prediction problem. For instance, the adaptation algorithms in PAC prediction sets could be incorporated into our approach.

Potential applications. Our approach can be extended to applications beyond visual acuity prediction. It can be applied to any regression task. Additionally, since the PAC interval can be used in classification tasks, this approach can be adapted for tasks such as diabetic retinopathy detection and glaucoma detection. In such classification tasks, the predictions will be sets of labels rather than intervals. For both regression and classification tasks, we can provide PAC guarantees for coverage.

Ethical Considerations. Guidelines for this approach can be established, such as specifying the appropriate interval width based on the variability in traditional visual acuity measurements (around 2 within the 0-10 label range) (Veselỳ and Synek, 2012). If the predicted interval exceeds this width, its clinical usefulness may be diminished. Furthermore, previous studies (Straitouri et al., 2023; Babbar et al., 2022) have demonstrated that prediction intervals with probabilistic guarantees help clinicians make more accurate decisions. By engaging in meaningful discussions with clinicians about our approach, we believe that the effectiveness of our methods can be further maximized.

Additional Analysis. We perform qualitative and robustness analyses in Appendix B. We visualize samples categorized by accurate vs. incorrect predictions and narrow vs. wide intervals, along with their activation maps. We also show that our algorithm remains robust under mild blurring conditions.

5. Conclusion

In this work, we developed a regression model that predicts visual acuity, modeled as a Gaussian distribution. From the output, our approach derives prediction intervals with a PAC guarantee on coverage while varying the prediction interval widths. Our empirical results demonstrate that the approach meets the probabilistic guarantee, with an average interval width of 3.04 for a coverage bound of $\epsilon = 0.3$. These outcomes are comparable to, or surpass, those of prior studies. Importantly, our method provides a PAC guarantee on coverage, which is particularly advantageous for clinical applications - an aspect previously unaddressed. This represents a significant step toward creating more trustworthy models for visual acuity prediction.

Future work could further enhance these models by: 1) employing more advanced base architectures, as our findings suggest that increased model complexity yields lower errors, and 2) leveraging the relationship between prediction error and estimated standard deviation, as discussed in Section 4.6. These future directions have the potential to lead to more accurate and clinically useful prediction intervals.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2023R1A2C1006639). Additionally, this research was supported in part by NIH 1R01EY037101 and ARO W911NF-20-1-0080.

References

Or Abramovich, Hadas Pizem, Jan Van Eijgen, Ilan Oren, Joshua Melamed, Ingeborg Stalmans, Eytan Z Blumenthal, and Joachim A Behar. Fundusq-net: A regression quality assessment deep learning algorithm for fundus images quality grading. *Computer Methods and Programs in Biomedicine*, 239: 107522, 2023.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen, and Philipp Berens. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical image analysis*, 64:101724, 2020.

Murat Seçkin Ayhan, Louis Benedikt Kümmerle, Laura Kühlewein, Werner Inhoffen, Gulnar Aliyeva, Focke Ziemssen, and Philipp Berens. Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Image Analysis*, 77:102364, 2022.

Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. *arXiv preprint arXiv:2205.01411*, 2022.

Benton Chuter, Justin Huynh, Christopher Bowd, Evan Walker, Jasmin Rezapour, Nicole Brye, Akram Belghith, Massimo A Fazio, Christopher A Girkin, Gustavo De Moraes, et al. Deep learning identifies high-quality fundus photographs and increases accuracy in automated primary open angle glaucoma detection. *Translational Vision Science & Technology*, 13(1):23–23, 2024.

Aaron S Coyner, Ryan Swan, J Peter Campbell, Susan Ostmo, James M Brown, Jayashree Kalpathy-Cramer, Sang Jin Kim, Karyn E Jonas, RV Paul Chan, Michael F Chiang, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmology retina*, 3(5):444–450, 2019.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Zongyuan Ge and Xin Wang. Evaluation of various open-set medical imaging tasks with deep neural networks. *arXiv preprint arXiv:2110.10888*, 2021.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22): 2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani. Sequential covariate shift detection using classifier two-sample tests. In *International Conference on Machine Learning*, pages 9845–9880. PMLR, 2022.

Xiayan Ji, Hyonyoung Choi, Oleg Sokolsky, and Insup Lee. Incremental anomaly detection with guarantee in the internet of medical things. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, pages 327–339, 2023.

Robert A Karlsson, Benedikt A Jonsson, Sveinn H Hardarson, Olof B Olafsdottir, Gisli H Halldorsson, and Einar Stefansson. Automatic fundus image quality assessment on a continuous scale. *Computers in Biology and Medicine*, 129:104114, 2021.

Jin Hyun Kim, Eunah Jo, Seungjae Ryu, Sohee Nam, Somin Song, Yong Seop Han, Tae Seen Kang, Woongsup Lee, Seongjin Lee, Kyong Hoon Kim, et al. A deep learning ensemble method to visual acuity measurement using fundus images. *Applied Sciences*, 12(6):3190, 2022.

Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. Bayesian-torch: Bayesian neural network layers for uncertainty estimation. https://github.com/IntelLabs/bayesian-torch, January 2022. URL https://doi.org/10.5281/zenodo.5908307.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1): 1–14, 2017.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. Pac prediction sets for meta-learning. *Advances in Neural Information Processing Systems*, 35:37920–37931, 2022a.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. Pac prediction sets under covariate shift. In *International Conference on Learning Representations*, 2022b.

William Paul, Philippe Burlina, Rohita Mocharla, Neil Joshi, Zhuolin Li, Sophie Gu, Onnisa Nanegrungsunk, Kira Lin, Susan B Bressler, Cindy X Cai, et al. Accuracy of artificial intelligence in estimating best-corrected visual acuity from fundus photographs in eyes with diabetic macular edema. *JAMA ophthalmology*, 141(7):677–685, 2023.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Aditya Raj, Anil Kumar Tiwari, and Maria G Martini. Fundus image quality assessment: survey, challenges, and future scope. *IET Image Processing*, 13(8):1211–1224, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10160–10177. PMLR, 2023.

Wenwen Si, Sangdon Park, Insup Lee, Edgar Dobriban, and Osbert Bastani. Pac prediction sets under label shift. In *The Twelfth International Conference on Learning Representations*, 2024.

John Siderov and Annette L Tiu. Variability of measurements of visual acuity in a large eye clinic. *Acta Ophthalmologica Scandinavica*, 77(6): 673–676, 1999.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR, 2023.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Petr Veselỳ and Svatopluk Synek. Repeatability and reliability of the visual acuity examination on logmar etdrs and snellen chart. *Ceska a Slovenska Oftalmologie: Casopis Ceske Oftalmologicke Spolecnosti a Slovenske Oftalmologicke Spolecnosti*, 68 (2):71–75, 2012.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, et al. Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7):12–12, 2022.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Appendix A. Dataset Imbalance

The distribution of the data across each class is provided in Table A.6, and the dataset is imbalanced, with most of the data corresponding to good visual acuity (label = 9 or 10).

Table A.6: Dataset Overview: The dataset comprises 11 classes, with each class containing a varying number of data points.

	0	1	2	3	4	5	6	7	8	9	10	Total
Count	3274	2164	1647	2091	2005	3240	3803	4261	4370	6358	21568	54781

Although the dataset exhibits class imbalance, we did not apply oversampling or weighting techniques, as the PAC coverage still holds. The trade-off is that the intervals may be wider for minority classes due to the model's lower performance on these classes. For comparison, we apply a resampling technique based on class frequency and evaluate the coverage rate and interval width of the EfficientNetV2-S model against those achieved with the vanilla technique. As shown in Table A.7, we do not observe significant improvements from resampling and therefore it is not included in the final version.

Table A.7: Comparison between the vanilla and resampling technique. $\epsilon = 0.3, \delta = 0.001\%$

Technique	Coverage Rate	Width
Vanilla	71.49 ± 0.85	3.04 ± 0.06
Resampling	71.79 ± 1.35	3.21 ± 0.09

Appendix B. Additional Experiment Results

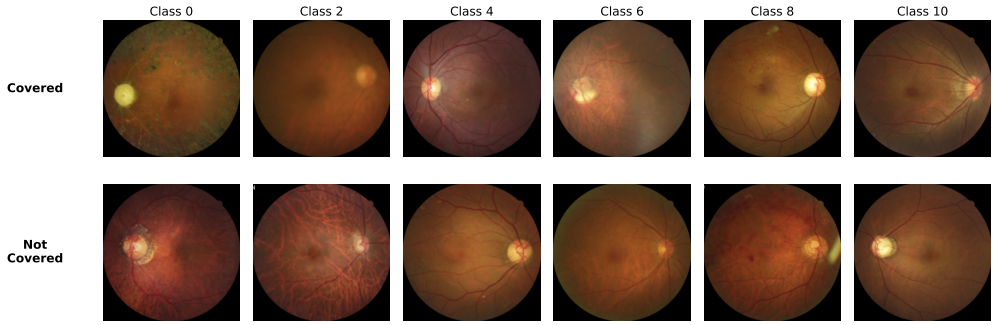
B.1. Qualitative Analysis

We conduct a qualitative analysis of our model's results, specifically sampling images where the model's prediction interval either contains the ground truth or does not, as well as images with both wide and narrow intervals. Additionally, we apply EigenGradCAM to analyze the model's prediction behavior.
Figure B.6 presents sample images with different ground truth values, illustrating cases where the prediction intervals either cover or fail to cover the ground truth. The second figure displays sample images with varying ground truth values, comparing two scenarios: narrow prediction intervals (≤ 2.0) and wide intervals (≥ 5.0). Generally, fundus images with higher visual acuity (higher class) tend to be clearer. Additionally, we plot the activation maps from the neural network layers for deeper analysis.

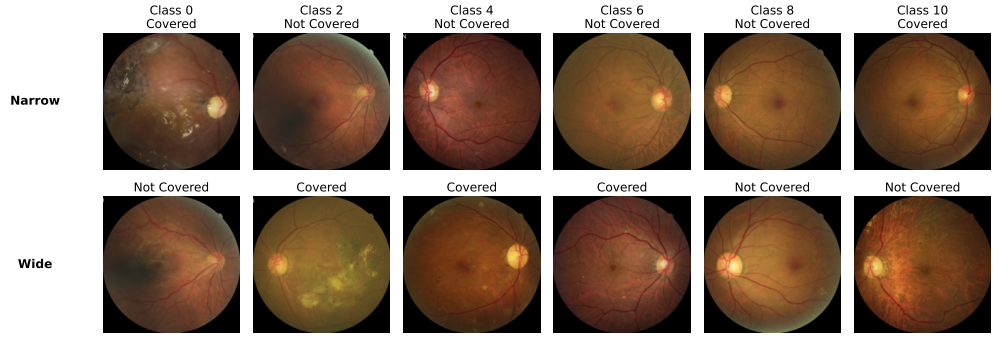
Figure B.7 shows activation maps generated by EigenGradCAM (using code from https://github.com/jacobgil/pytorch-grad-cam?tab=readme-ov-file) for the sample images drawn in the previous figures. These maps indicate that the model mainly focuses on the macula region. Furthermore, when the model evaluates a wider region, the prediction interval tends to include the ground truth, and the intervals are wider.

B.2. Robustness Analysis

We conduct experiments with multiple repetitions using different dataset splits based on various seeds, reporting the average and standard deviation. We believe that this experimental setup demonstrates the robustness of our approach across different participants. Additionally, the PAC interval algorithm provides



(*a*) Covered vs. Not Covered



(*b*) Narrow vs. Wide

Figure B.6: Image Samples for Different Cases.

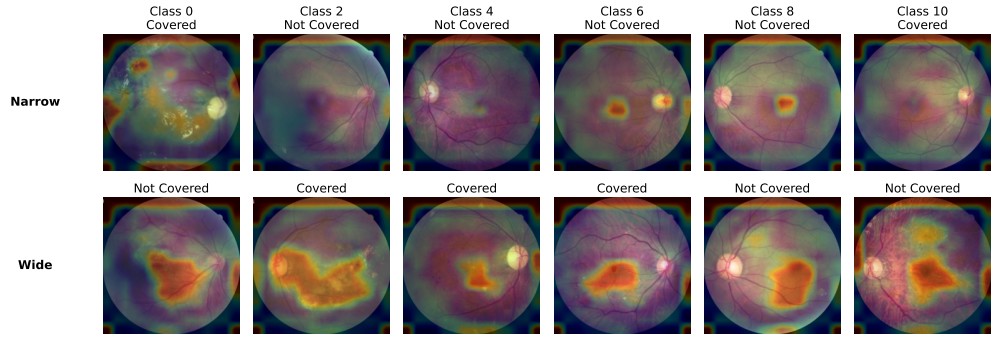
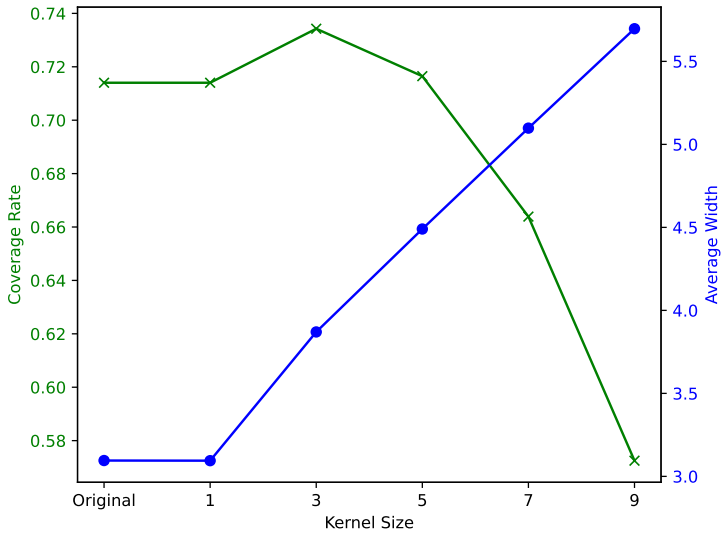


Figure B.7: Activation Maps from EigenGradCAM

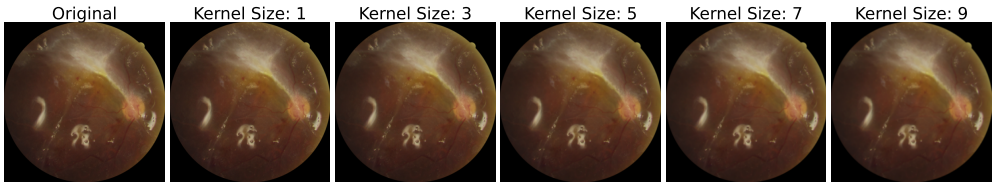
a conditional guarantee based on the training data (as indicated in Equation (1), where the calibration set Z^n is sampled), ensuring coverage regardless of the calibration set.
In terms of image quality, our approach may yield wider intervals for low-quality images, which might not be clinically helpful. Achieving accurate predictions in such cases is quite challenging. However, we believe this issue can be addressed by utilizing automatic image quality detection algorithms (Raj et al., 2019; Coyner

et al., 2019; Karlsson et al., 2021; Abramovich et al., 2023; Chuter et al., 2024). These algorithms can alert users when a captured image is of poor quality and prompt them to retake it. Subsequently, our algorithm will operate on high-quality images.

Since severe image quality degradation can be easily detected by the image quality detection algorithms, we investigate how well our approach handles non-severe low-quality images. We simulate varying image qualities by applying Gaussian blur with different kernel sizes (keeping sigma fixed at 5), as illustrated in Figure B.8(b). The result with EfficientNetV2-S (Figure B.8(a)) indicates that up to a kernel size of 5, the coverage rate remains valid, though the average interval width increases. However, with more severe blurring, our model fails to maintain the coverage rate as the distributional assumptions are violated. These results evidence the robustness of our approach to image quality degradation.



(a) Robustness of Our Model



(b) Blurred mage Samples

Figure B.8: Robustness Analysis Under Varying Levels of Blur