Canthus-Scaled 468-Landmark FaceMesh Framework for Pupillary Distance Estimation Using Nested AutoML Calibration

Mohd Izzuddin Mohd Tamrin[1], Sherzod Turaev[2], Takumi Sase[3],
Mohd Zulfaezal Che Azemin[4*], Tengku Mohd Tengku Sembok[5]
Kulliyyah of ICT, International Islamic University Malaysia, Gombak, Kuala Lumpur, Malaysia[1, 3, 5]
Department of Computer Science and Software Engineering-College of Information Technology,
United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates[2]
Integrated Omics Research Group-Kulliyyah of Allied Health Sciences,
International Islamic University Malaysia, Kuantan, Malaysia[4]

***Abstract*—Pupillary distance (PD) is an important ocular measurement for optical dispensing and vision-related applications, but standard MediaPipe FaceMesh outputs do not provide true pupil-centre or iris-boundary landmarks when only the 468-landmark representation is available. This study proposes a canthus-scaled 468-landmark framework for estimating PD using facial landmarks and Malay young-adult normative palpebral fissure width. A dataset of 44 subjects was used, where each record contained ground-truth PD and 1,404 coordinate values representing 468 FaceMesh landmarks in three dimensions. Since true pupil centres were unavailable, medial and lateral canthus landmarks were used to construct eye-centre proxies and to compute a subject-specific millimetre scale. A direct canthus-scaled proxy was first evaluated as a deterministic baseline, after which canthus-scaled geometric features were used in a nested AutoML calibration framework. Model development used repeated nested cross-validation, with an outer repeated 5-fold design and an inner 4-fold model-selection loop. The direct proxy achieved a mean absolute error (MAE) of 4.26 mm and showed systematic overestimation. The calibrated nested AutoML model improved performance, achieving a subject-level MAE of 3.510 mm, root mean squared error of 4.22 mm, a bias of −0.08 mm, and 75.0% of predictions within ±5 mm. The calibrated nested AutoML model improved overall error and reduced systematic bias compared with the direct canthus-scaled proxy. However, the Bland–Altman limits of agreement remained wide, indicating that the proposed method should be interpreted as an approximate proxy-based estimation approach rather than a substitute for clinical pupillometer- or ruler-based PD measurement. The framework is most relevant for research settings or datasets where only standard 468-landmark FaceMesh data are available, and iris-refined landmarks are absent.**

***Keywords*—Pupillary distance; MediaPipe FaceMesh; facial landmarks; canthus scaling; palpebral fissure width; AutoML; nested cross-validation; computer vision**

I. Introduction

Pupillary distance (PD), commonly discussed as interpupillary distance (IPD), is the horizontal distance between the centres of the two pupils and is an important measurement in optometry, ophthalmic dispensing, binocular vision assessment, and the fitting of ophthalmic lenses. Inaccurate PD measurement may affect optical centration and can contribute to visual discomfort, especially when lenses are prescribed or manufactured using incorrect centration parameters. Conventional PD measurement is commonly performed using clinical instruments such as pupillometers, autorefractors, or manual PD rulers. However, these approaches may require trained personnel, dedicated equipment, or direct clinical access, which motivates the development of low-cost and automated alternatives [1], [2].

Recent studies have investigated smartphone- and computer-vision-based methods for estimating PD or related ocular measurements. Jung and Chu compared modern IPD measurement methods, including clinical and mobile application approaches, showing the growing interest in non-contact digital PD estimation [1]. Han et al. evaluated smartphone applications for IPD measurement, further supporting the feasibility of mobile-based measurement but also highlighting the need for accuracy validation [2]. Zhang et al. proposed a computer-vision method for IPD and pupil-height measurement using ensemble regression trees and segmentation-based techniques, demonstrating that automated ocular measurement remains an active area of applied vision research [3]. Similarly, pupil-centre detection has been investigated using convolutional neural networks and webcam-based systems, indicating that accurate pupil localization is technically challenging but central to reliable ocular measurement [4].

MediaPipe FaceMesh has become a widely used framework for real-time facial landmark extraction because it can estimate dense three-dimensional facial landmarks from ordinary RGB images. The standard FaceMesh model estimates 468 facial landmarks, making it attractive for lightweight and non-invasive facial analysis applications [5]. However, the standard 468-landmark representation does not include true iris-boundary or pupil-centre landmarks. MediaPipe Iris extends the FaceMesh framework by estimating additional iris and eye-contour landmarks, enabling more direct iris- and pupil-related measurements from a single RGB camera [6], [7]. In datasets where only the 468 FaceMesh

*Corresponding author.

landmarks are available, PD cannot be measured directly from true pupil centres; instead, proxy-based estimation using periocular landmarks is required.

Population-specific anthropometric information is important when facial or ocular landmarks are converted from image-space distances into metric units. Ngeow and Aljunid established craniofacial anthropometric norms for young adult Malaysian Malays aged 18–25 years, providing population-relevant reference measurements for craniofacial and orbital regions [8]. Othman et al. further established three-dimensional facial soft-tissue morphology for adult Malay subjects using stereophotogrammetry, reinforcing the importance of population-specific facial measurements in Malaysian samples [9]. Additional Malaysian and regional studies on palpebral fissure and canthal measurements show that periocular dimensions vary across sex and ethnic groups, supporting the use of appropriate normative references when converting facial landmark distances into estimated millimetres [10]–[13].

Although iris-based scaling is preferable when iris landmarks are available, many existing landmark datasets contain only 468 FaceMesh landmarks. In such cases, canthus-based scaling provides a practical alternative. The medial and lateral canthus landmarks can be used to estimate palpebral fissure width, and a normative palpebral fissure width value can then be used as a scale reference. This allows the distance between canthus-derived eye-centre proxies to be converted into an estimated PD value. However, because the eye-centre proxies are not anatomical pupil centres, a direct normative conversion may introduce systematic bias. Machine-learning calibration can therefore be used to learn the relationship between canthus-scaled facial landmark features and ground-truth PD.

The key limitation addressed in this study is that many existing facial-landmark datasets contain only the standard 468 MediaPipe FaceMesh landmarks and do not include iris-boundary or anatomical pupil-centre landmarks. Under this constraint, PD cannot be measured directly from the true pupil centres. This study, therefore, investigates a two-stage framework: first, a geometric canthus-scaled proxy is constructed using medial and lateral canthus landmarks and Malay young-adult normative palpebral fissure width; second, a predictive calibration model is trained to reduce systematic proxy error using nested cross-validation. The contributions of this study are: 1) a 468-landmark canthus-scaled proxy method for approximate PD estimation, 2) an evaluation of direct geometric estimation versus calibrated prediction, 3) a nested AutoML validation strategy for a small biomedical dataset, and 4) an explicit discussion of the clinical and deployment limitations of the proxy-based approach.

II. Related Work

Previous work on pupillary distance estimation can be grouped into four main areas: conventional clinical measurement, smartphone- and computer-vision-based ocular measurement, facial landmark and iris-based estimation, and validation methods for small-sample prediction models. Conventional PD measurement is commonly performed using pupillometers, autorefractors, or manual rulers, which remain more appropriate for clinical dispensing because they are designed to measure ocular centration directly. However, these approaches require equipment, trained personnel, or clinical access, motivating low-cost computer-vision alternatives.

Smartphone and computer-vision methods have shown promise for non-contact PD and related ocular measurements, but their accuracy depends strongly on pupil or iris localization, camera geometry, subject pose, and calibration strategy. Methods that use direct pupil or iris detection are conceptually closer to anatomical PD measurement, whereas approaches based only on facial landmarks must rely on surrogate geometric features. MediaPipe FaceMesh provides dense facial landmarks and is attractive for lightweight facial analysis, but the standard 468-landmark model does not include true pupil-centre or iris-boundary landmarks. MediaPipe Iris extends this representation with additional iris-related landmarks and is therefore preferable when iris landmarks are available.

In contrast, the present study focuses on the more restricted setting where only 468 FaceMesh landmarks are available. Under this condition, canthus landmarks provide a practical periocular reference because they are visible in the standard model and can be linked to palpebral fissure width. Anthropometric scaling is also important because image-space distances must be converted into millimetres. Population-specific normative values may reduce scaling error compared with arbitrary scaling, but they cannot account fully for inter-subject periocular variation. Therefore, this study treats canthus scaling as an approximate population-level calibration rather than a subject-specific physical calibration.

Finally, because the dataset is small and model selection is part of the modelling process, nested cross-validation is used to reduce optimistic performance estimation. This is important because using cross-validation both for model selection and final error estimation can introduce biased performance estimates [14]. In addition, the reporting of the proposed prediction framework was informed by TRIPOD+AI guidance for regression and machine learning-based prediction models, particularly in relation to transparent description of the data, predictors, validation procedure, performance metrics, and limitations [15].

III. Methodology

The proposed framework consists of two conceptually distinct stages. Stage 1 is geometric proxy estimation, in which canthus landmarks are used to estimate palpebral fissure width, construct eye-centre proxies, and convert landmark distances into estimated millimetres. Stage 2 is predictive calibration, in which canthus-scaled geometric features are used to learn the relationship between proxy geometry and reference PD. The first stage is deterministic and interpretable, whereas the second stage is a data-driven correction layer. The calibrated model should therefore be interpreted as a correction of a canthus-derived proxy, not as a direct measurement of anatomical pupil-centre distance.

A. Dataset and Landmark Representation

This study used a subject-level facial landmark dataset for pupillary distance (PD) estimation. Each record contained a subject identifier, a ground-truth PD value in millimetres, and

1,404 coordinate values representing 468 MediaPipe FaceMesh landmarks in three-dimensional form. The landmark values were arranged as 468 x-coordinates, 468 y-coordinates, and 468 z-coordinates. The dataset contained 44 usable subjects after retaining records with valid subject identifiers, ground-truth PD values, and complete landmark coordinates.

The standard 468-landmark FaceMesh representation does not include anatomical pupil centres or iris-boundary landmarks. Therefore, this study estimated PD using canthus-derived eye-centre proxies and a canthus-based normative scaling procedure. The complete methodological workflow is illustrated in Fig. 1.

B. Landmark Pre-processing

The uploaded landmark matrix was first parsed into a structured coordinate tensor. For each subject, the 1,404 landmark values were reshaped into a 468×3 matrix, where each landmark was represented by x, y, and z coordinates. Records of missing or non-numeric landmark values were excluded to ensure complete geometric calculations.

The following FaceMesh landmarks were used as the primary ocular reference points:

L_33and L_133: right-eye lateral and medial canthus region landmarks.

L_263and L_362: left-eye lateral and medial canthus region landmarks.

These landmarks were selected because they are available in the 468-landmark FaceMesh model and correspond to the canthal region required for palpebral fissure width estimation.

C. Canthus-Based Normative Scaling

Since no physical calibration object or iris landmarks were available, image-space landmark distances were converted into millimetres using a population-level canthus-based normative scale. A normative palpebral fissure width (PFW) of 29.100 mm was used as the scaling constant for the Malay young-adult population.

For each subject, the right and left image-space palpebral fissure widths were calculated as:

$$PFW_R = d(L_{33}, L_{133}) \quad (1)$$

$$PFW_L = d(L_{263}, L_{362}) \quad (2)$$

The subject-specific image-space palpebral fissure width was then computed as:

$$PFW_{image} = \frac{PFW_R + PFW_L}{2} \quad (3)$$

The millimetre-per-landmark-unit scale factor was calculated as:

$$S = \frac{PFW_{normative}}{PFW_{image}} \quad (4)$$

where, PFW_normative=29.100mm.

This scale factor was applied to convert selected landmark distances from landmark units into estimated millimetres. This scaling approach assumes that the selected normative palpebral fissure width is representative of the target population. However, it does not account for subject-specific anatomical variation in eyelid aperture, canthal position, sex-related differences, or facial pose. To make this limitation explicit, the subject-specific scale factor and right-left palpebral fissure asymmetry were retained as scale quality-control variables.

D. Eye-Centre Proxy Estimation

Because true pupil centres were not available in the 468-landmark representation, this study did not attempt to measure anatomical PD directly. Instead, each eye was represented by a canthus-derived geometric proxy defined as the midpoint between the medial and lateral canthus landmarks. The right and left eye-centre proxies were calculated as:

$$E_R = \frac{L_{33} + L_{133}}{2} \quad (5)$$

$$E_L = \frac{L_{263} + L_{362}}{2} \quad (6)$$

The direct canthus-scaled proxy PD was then calculated as:

$$PD_{proxy} = d(E_R, E_L) \times S \quad (7)$$

This direct proxy was used as a deterministic baseline for comparison with the machine-learning-based prediction model. This midpoint is expected to differ from the true anatomical pupil centre because the pupil position depends on gaze, eyelid aperture, ocular anatomy, and camera perspective. Therefore, the direct canthus-scaled estimate was treated only as a deterministic baseline and not as a clinical PD measurement.

E. Feature Engineering

Canthus-scaled geometric features were constructed from the 468-landmark coordinate representation. The engineered feature groups included:

- Eye-corner midpoint distance, computed from the distance between the right and left canthus-derived eye-centre proxies.
- Eye-contour centroid distance, computed from the centroid of available eye-contour landmarks.
- Anatomical pairwise landmark distances, including ocular, facial-width, mouth-width, and vertical face-distance proxies.
- Scale quality-control features, including the subject-specific scale factor and the percentage asymmetry between right and left palpebral fissure widths.
- Full landmark representations, including all 468 landmarks and subject-level normalized landmark features.

All distance-based geometric features were converted into estimated millimetres using the subject-specific canthus scale factor. Because the number of subjects was limited, the high-dimensional all-landmark representation was treated cautiously. The candidate model set, therefore, included low-dimensional eye-region and anatomical-distance feature sets, scale quality-control variables, and regularized models. The all-landmark feature set was included only as one candidate representation within the nested selection procedure, not as the default model.

F. Direct Proxy Baseline

A direct canthus-scaled PD proxy was calculated for each subject using only the eye-centre proxy distance and the normative canthus scale factor. This method did not require model training and served as an interpretable baseline for evaluating whether a simple normative conversion was sufficient for PD estimation.

To evaluate whether the added complexity of AutoML calibration was justified, simple statistical calibration baselines were added. These included a linear correction model using the direct canthus-scaled proxy as the predictor and a polynomial correction model using first- and second-order terms of the direct proxy. These models were evaluated using the same outer cross-validation structure as the AutoML framework, so that calibration performance could be compared under the same validation conditions.

G. Nested AutoML Prediction Model

A nested cross-validation framework was used to train and evaluate machine-learning models while avoiding data leakage. The outer loop used repeated 5-fold cross-validation with 10 repeats. In each outer fold, subjects in the test split were held out and were not used during model selection or hyperparameter tuning. The inner loop used 4-fold cross-validation on the training split to select the optimal feature set, estimator, and hyperparameters.

The AutoML candidate set included baseline and regression-based models using canthus-scaled feature sets. Candidate models included dummy regression, linear eye-region estimators, Ridge regression, ElasticNet regression, support vector regression, and partial least squares regression. Model selection was based on minimizing mean absolute error in millimetres within the inner cross-validation loop.

For each outer fold, the best-performing model from the inner loop was refitted on the corresponding outer training set and used to predict PD for the held-out subjects. Repeated outer-fold predictions were averaged to obtain one subject-level prediction per participant.

A landmark perturbation analysis was conducted to evaluate sensitivity to plausible localization error. Key canthus landmarks and eye-contour landmarks were perturbed with zero-mean random noise at predefined magnitudes, and the direct proxy and calibrated predictions were recomputed. Performance was summarized using MAE, RMSE, bias, and Bland–Altman limits of agreement. Because the available dataset contained landmark coordinates rather than raw images, image-level effects such as illumination variation, motion blur, occlusion, and head-pose changes could not be fully evaluated. These image-level robustness factors are therefore identified as priorities for future validation.

H. Evaluation Protocol

The direct proxy baseline and nested AutoML predictions were evaluated against ground-truth PD values. The evaluation metrics included mean absolute error, root mean squared error, bias, standard deviation of error, median absolute error, coefficient of determination, Bland–Altman limits of agreement, and the percentage of predictions within ±2 mm and ±5 mm.

Bootstrap resampling was used to estimate confidence intervals for subject-level performance metrics. The final analysis compared the deterministic canthus-scaled proxy with the calibrated nested AutoML model to assess whether learned calibration improved PD estimation from 468-landmark canthus-scaled features.

I. Workflow Summary

The proposed methodology consists of landmark parsing, canthus-based scale estimation, eye-centre proxy calculation, direct proxy PD estimation, canthus-scaled feature construction, nested AutoML calibration, and final evaluation. This process is summarized in Fig. 1.

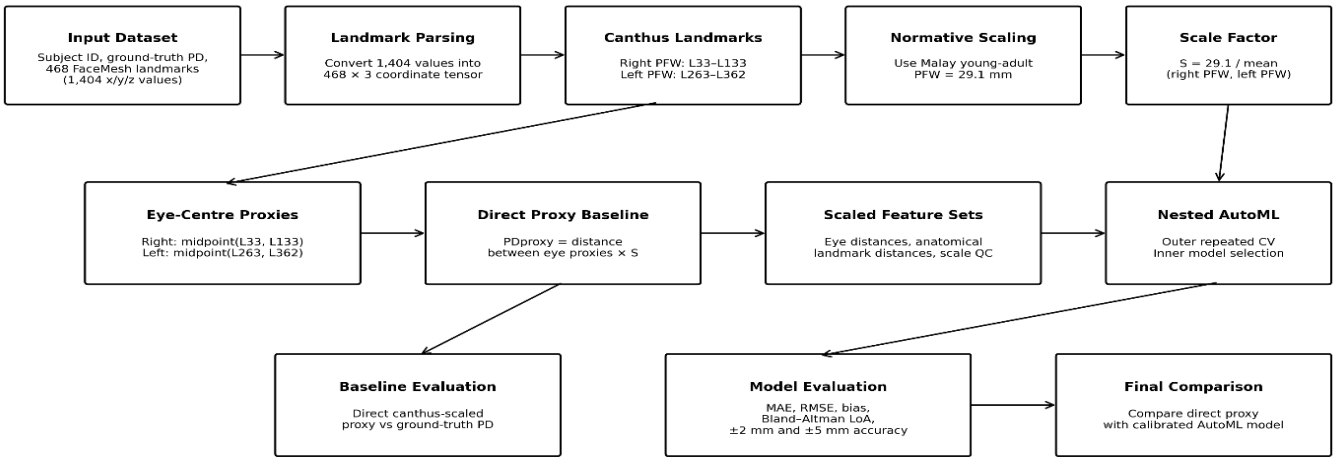

Fig. 1. Overall methodology of the 468-landmark canthus-scaled PD estimation framework.

IV. Results and Discussion

A. Dataset and Experimental Configuration and Acronyms

The final analysis included 44 subjects with complete 468-landmark FaceMesh data and corresponding ground-truth pupillary distance (PD) values. Each subject contributed 1,404 landmark coordinate values, representing 468 landmarks in x, y, and z dimensions. The proposed workflow used a canthus-based normative scaling procedure, where landmarks 33–133 and 263–362 were used to estimate right and left palpebral fissure widths, respectively. The overall experimental configuration is summarized in Table I.

The use of nested cross-validation was appropriate because the dataset was small and model selection was performed during training. In small-sample prediction modelling, non-nested validation may produce unstable or optimistic estimates, whereas nested validation separates model selection from performance estimation [16], [17].

The canthus-based scale factor showed measurable inter-subject variability. The mean scale factor was 1109.74 mm per landmark unit with a standard deviation of 141.41 mm per landmark unit. The mean right-left palpebral fissure width asymmetry was 4.95%, while the maximum asymmetry reached 44.09%. These values indicate that although normative scaling provides a practical conversion from landmark units to millimetres, individual periocular morphology and landmark asymmetry may introduce scaling error.

B. Direct Canthus-Scaled Proxy Baseline

The direct canthus-scaled proxy served as a deterministic baseline. This method estimated PD by calculating the distance between the right and left canthus-derived eye-centre proxies and multiplying it by the subject-specific canthus scale factor. As shown in Table II, the direct proxy achieved an MAE of 4.26 mm and RMSE of 4.91 mm. The bias was +2.83 mm, indicating that the direct proxy generally overestimated the ground-truth PD.

This finding suggests that canthus-based scaling alone is insufficient for accurate PD estimation when true pupil centres are unavailable. The direct proxy is interpretable and simple, but it depends on the assumption that the midpoint between the medial and lateral canthus approximates the pupil centre. Since the 468-landmark representation does not include actual pupil-centre landmarks, systematic geometric bias is expected.

TABLE I. Dataset and Experimental Configuration

Item	Value
N usable subjects	44
Target PD mean (mm)	63.59
Target PD SD (mm)	4.26
Target PD median (mm)	62.50
Target PD min (mm)	55.00
Target PD max (mm)	72.00
Landmark values per subject	1404.00
Interpreted landmarks	468.00
Metric scaling approach	468-landmark canthus-based normative scaling
Normative palpebral fissure width (mm)	29.10
Scale landmarks	Right 33-133 and left 263-362; scale = normative PFW / mean observed PFW
Direct proxy formula	distance(midpoint(33,133), midpoint(263,362)) x canthus scale
Outer CV	Repeated 5-fold CV x 10 repeats
Inner CV	4-fold CV
Primary model selection metric	Mean absolute error (MAE), mm

TABLE II. Performance Comparison Between Direct Proxy and Nested AutoML

Analysis level	MAE (mm)	RMSE (mm)	Bias (mm)	Median AE (mm)	Within +/-2 mm (%)	Within +/-5 mm (%)	LoA lower (mm)	LoA upper (mm)	95% CI
Direct canthus-scaled proxy	4.26	4.91	2.83	4.31	20.45	59.09	-5.12	10.80	
Nested AutoML subject-level	3.51	4.23	-0.08	3.57	31.82	75.00	-8.45	8.30	MAE 2.82-4.21; RMSE 3.47-4.91; Bias -1.34-1.12
Nested AutoML outer-fold	3.61	4.35	-0.08	3.46	32.27	72.05	-8.61	8.46	

TABLE III. AutoML-Selected Candidate Frequency Across Outer Folds

Selected candidate	Selected feature set	Outer folds	Mean inner MAE (mm)	SD inner MAE (mm)	Selection (%)
Eye-centroid estimator	eye_centroid_scaled_mm	23	3.45	0.21	46.00
Eye-corner estimator	eye_corner_scaled_mm	6	3.21	0.35	12.00
Geometry SVR-RBF	geometry_26_scaled_mm	6	3.28	0.17	12.00
Geometry Ridge	geometry_26_scaled_mm	6	3.47	0.22	12.00
Geometry ElasticNet	geometry_26_scaled_mm	5	3.22	0.13	10.00
All-landmarks PLS	all_landmarks_scaled_mm	3	3.39	0.14	6.00
Dummy median baseline	geometry_26_scaled_mm	1	3.396		2.00

C. Nested AutoML Prediction Performance

The nested AutoML model improved the overall prediction accuracy compared with the direct proxy baseline. As reported in Table II, the subject-level nested AutoML prediction achieved an MAE of 3.51 mm, RMSE of 4.22 mm, and bias of −0.08 mm. The proportion of predictions within ±5 mm increased to 75.00%, compared with 59.09% for the direct proxy baseline. The bootstrap 95% confidence interval for subject-level MAE was 2.82–4.21 mm, indicating the expected uncertainty range of the model's average absolute error.

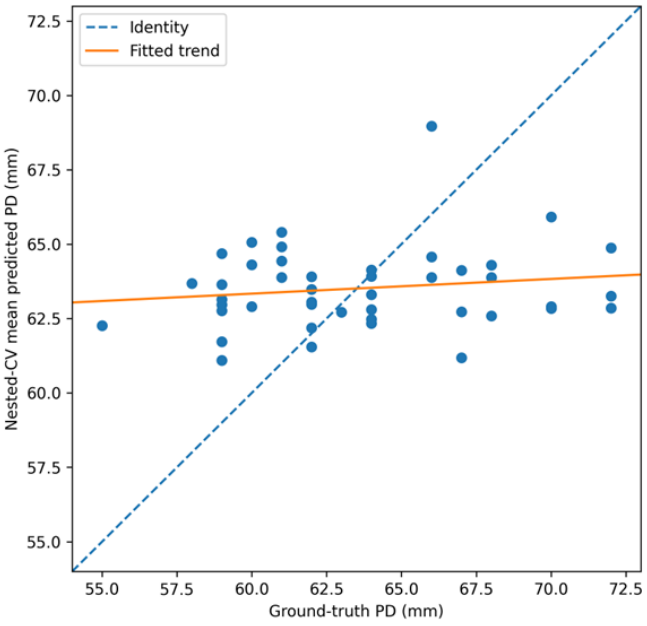

Fig. 2. Predicted versus ground-truth PD using nested AutoML.

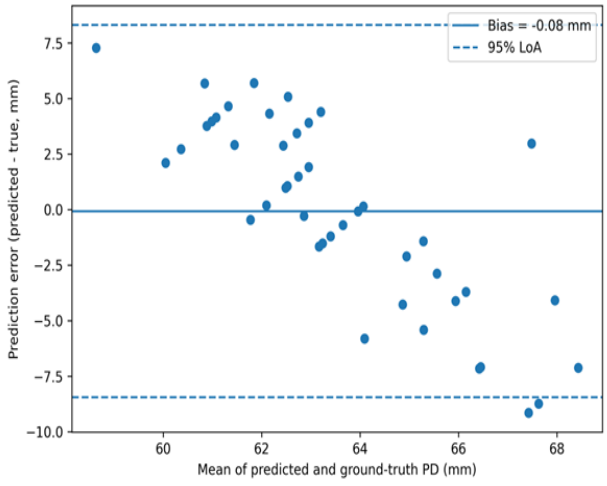

Fig. 3. Bland-Altman analysis of nested AutoML PD estimation.

The predicted-versus-ground-truth plot in Fig. 2 shows that the AutoML predictions were concentrated within a narrower prediction range than the ground-truth PD values. This indicates that the model learned a calibrated central tendency but had limited ability to fully represent the extremes of the PD distribution. This behaviour is common in small-sample regression models, where predictions may shrink toward the mean when the training set contains limited variation [18].

D. Agreement Analysis

The Bland–Altman plot in Fig. 3 provides a clinically interpretable view of agreement between the nested AutoML predictions and the ground-truth PD values. The average bias was close to zero, which indicates that the learned calibration model corrected the systematic overestimation observed in the direct proxy baseline. However, the limits of agreement remained relatively wide, showing that individual-level prediction error may still be meaningful.

Bland–Altman analysis is useful in this context because correlation or fitted trend alone does not fully describe agreement between a new measurement method and a reference method [19]. Therefore, Fig. 3 is more informative than relying only on the scatter plot in Fig. 2, especially for evaluating whether the method is suitable for measurement-related applications.

E. AutoML Model-Selection Behaviour

The AutoML selection results are summarized in Table III. The selection frequencies indicate that simpler eye-region and geometry-based models were selected more often than the full all-landmark model. The eye-centroid estimator was selected in 46.0% of outer folds, whereas the all-landmarks PLS model was selected in only 6.0% of outer folds. This suggests that the nested selection procedure did not consistently favor the highest-dimensional representation. Nevertheless, the small sample size means that feature selection may still be unstable, and the selected model frequencies should be interpreted as exploratory rather than definitive evidence of feature importance.

This selection pattern suggests that interpretable eye-region features were more stable than the full high-dimensional landmark representation. The relatively low selection frequency of the all-landmark PLS model indicates that using all 468 landmarks did not consistently improve model selection. This is methodologically important because the sample size was small relative to the number of potential landmark predictors. Prediction modelling literature recommends caution when the number of predictors is large compared with the number of observations, since overfitting and unstable selection may occur [17], [18].

F. Comparison Between Direct Proxy and Learned Calibration

The comparison between the direct proxy and nested AutoML model shows that the main benefit of AutoML was not simply reducing error magnitude but reducing systematic bias. The direct canthus-scaled proxy was easy to compute but overestimated PD. In contrast, the nested AutoML model used the same canthus-scaled landmark information while learning a calibration relationship between proxy geometry and reference PD. This explains why the AutoML model achieved a much smaller bias while also improving MAE and RMSE.

For the manuscript, the direct proxy should therefore be presented as a baseline rather than the final proposed estimator. The final model should be described as a calibrated 468-

landmark canthus-scaled PD estimation framework. This framing is important because the 468-landmark FaceMesh model does not provide anatomical pupil centres; consequently, the model estimates PD from canthus-derived ocular proxies rather than measuring pupil-centre distance directly.

G. Practical Implications and Limitations

The proposed approach is useful when only standard 468 FaceMesh landmarks are available, and iris landmarks cannot be extracted. It provides a practical method for converting landmark geometry into estimated millimetres using Malay young-adult normative palpebral fissure width and then calibrating the estimate using nested AutoML.

However, several limitations should be considered. First, the method relies on normative palpebral fissure width rather than subject-specific physical calibration. Second, the eye-centre proxy is derived from canthus landmarks and is not equivalent to a true pupil centre. Third, the dataset contained only 44 subjects, so the reported performance should be interpreted as preliminary. Finally, the model should be externally validated on a larger independent Malay undergraduate cohort before being used as a general-purpose PD estimation method.

Overall, the results support the feasibility of canthus-scaled 468-landmark PD estimation, but they also show that direct geometric conversion is not sufficient on its own. Learned calibration through nested AutoML improved bias and overall error while preserving an interpretable landmark-based measurement framework.

H. Clinical Suitability and Agreement Limits

Although the nested AutoML model reduced the average bias to −0.08 mm, the Bland–Altman limits of agreement remained wide, ranging from −8.45 to +8.30 mm. The total agreement width was therefore approximately 16.75 mm. This level of variability is not acceptable for clinical spectacle dispensing, where centration errors of only a few millimetres can be practically important, particularly for progressive or high-powered lenses. Therefore, the proposed framework should not be interpreted as a replacement for clinical PD measurement using a pupillometer, autorefractor, or trained manual measurement. Its current value is as an approximate proxy-estimation method for research use, preliminary screening, or retrospective datasets where only 468 FaceMesh landmarks are available.

I. Interpretation of the Canthus-Derived Eye-Centre Proxy

The central assumption of the geometric stage is that the midpoint between the medial and lateral canthus provides a stable periocular reference. This assumption is useful computationally because the landmarks are available in the standard 468-landmark FaceMesh representation. However, the canthus midpoint is not equivalent to the anatomical pupil centre. Differences between the proxy and the true pupil centre may arise from gaze direction, eyelid shape, palpebral fissure morphology, subject pose, and landmark localization error. The systematic overestimation observed in the direct proxy baseline is consistent with this limitation. The calibrated AutoML model partially reduced this bias, but it cannot fully recover true pupil-centre geometry when the relevant anatomical landmarks are absent.

J. Effect of Periocular Anatomical Variability

The use of Malay young-adult normative palpebral fissure width provides a population-relevant scale reference, but it does not eliminate inter-subject anatomical variability. Subjects with wider or narrower palpebral fissures than the normative value, asymmetric canthal landmarks, or atypical eyelid morphology may receive inaccurate scale factors. The observed scale-factor variability and palpebral fissure asymmetry support this concern. Future work should investigate sex-specific scaling, subject-specific calibration objects, iris-based scaling when available, and exclusion or quality-control criteria for cases with high canthal asymmetry.

K. Small Sample Size and Model Generalizability

The dataset contained 44 subjects, which is small relative to the number of possible landmark-derived predictors. Although repeated nested cross-validation was used to reduce optimism from model selection, it cannot fully remove the risk of unstable feature selection or limited external generalizability. The results should therefore be interpreted as preliminary internal-validation findings. External validation on a larger independent cohort is required before the method can be recommended for broader deployment.

L. Robustness to Real-World Image Variation

The present dataset contained processed landmark coordinates rather than raw images. Therefore, the study could not directly evaluate the effect of illumination, image blur, camera resolution, head pose, gaze direction, eyelid occlusion, or partial landmark failure. These factors may substantially affect practical performance because the proposed framework depends on accurate detection of canthus and eye-contour landmarks. The perturbation analysis provides an initial landmark-level sensitivity assessment, but image-level robustness must be evaluated in future studies using controlled acquisition protocols and raw image data.

V. Conclusion

This study developed and evaluated a canthus-scaled 468-landmark framework for estimating pupillary distance from MediaPipe FaceMesh landmarks. Since the standard 468-landmark representation does not provide true pupil centres or iris-boundary landmarks, the proposed approach used medial and lateral canthus landmarks to construct eye-centre proxies and to derive a subject-specific metric scale based on normative Malay young-adult palpebral fissure width. This enabled pupillary distance estimation from landmark geometry without requiring iris-refined landmarks or a physical calibration object.

The direct canthus-scaled proxy provided a simple and interpretable baseline; however, it was limited by systematic overestimation because canthus-derived midpoints are not equivalent to anatomical pupil centres. The nested AutoML calibration model improved the estimation by learning the relationship between canthus-scaled landmark features and ground-truth pupillary distance. The use of repeated nested cross-validation reduced optimism from model selection and

provided a more reliable estimate of predictive performance for the small dataset.

Overall, the findings indicate that standard 468 FaceMesh landmarks can support approximate proxy-based PD estimation when combined with population-specific canthus scaling and calibration. However, the approach should not be regarded as a direct clinical measurement of pupil-centre distance. The wide limits of agreement, small sample size, reliance on normative scaling, and absence of true iris or pupil landmarks limit immediate clinical applicability. Future work should validate the framework on a larger independent cohort, compare simple calibration models with AutoML models, evaluate robustness to landmark and image-level errors, incorporate sex-specific or subject-specific scaling, and benchmark the 468-landmark approach against iris-refined landmark models and standard clinical PD measurement instruments.

Declaration on Generative AI

ChatGPT 5.5 was used to assist with language editing and to improve clarity during manuscript preparation. The authors take full responsibility for the content, confirm that all scientific interpretations and conclusions are their own, and approve the final version of the manuscript.

Acknowledgment

The authors would like to express their sincere appreciation to Ms. Nor Alya Batrisyia Nor Izani for her valuable assistance during the data collection process. Her support contributed to the successful preparation of the dataset used in this study.

References

[1] Y. R. Jung and B. S. Chu, “A comparative analysis of interpupillary distance measurement techniques evaluation in modern times: From rulers to apps,” Clinical Optometry, vol. 16, pp. 309–316, 2024.

[2] K. D. Han, M. Jaafar, I. M. Stoakes, P. C. Hoopes, and M. Moshirfar, “Comparing the effectiveness of smartphone applications in the measurement of interpupillary distance,” Cureus, vol. 15, no. 7, Art. no. e42744, 2023.

[3] Z. Zhang, H. Xiang, D. Li, and C. Leng, “Measurement method of interpupillary distance and pupil height based on ensemble of regression trees and the BlendMask algorithm,” Applied Sciences, vol. 13, no. 15, Art. no. 8628, 2023.

[4] A. Larumbe-Bergera, G. Garde, S. Porta, R. Cabeza, and A. Villanueva, “Accurate pupil center detection in off-the-shelf eye tracking systems using convolutional neural networks,” Sensors, vol. 21, no. 20, Art. no. 6847, 2021.

[5] A. Ablavatski, I. Grishchenko, Y. Kartynnik, K. Raveendran, and M. Grundmann, “Attention Mesh: High-fidelity face mesh prediction in real-time,” arXiv preprint arXiv:2006.10962, 2020.

[6] Google MediaPipe, “MediaPipe Face Mesh,” Google, 2024.

[7] Google Research, “MediaPipe Iris: Real-time iris tracking & depth estimation,” Google Research Blog, 2020.

[8] W. C. Ngeow and S. T. Aljunid, “Craniofacial anthropometric norms of Malays,” Singapore Medical Journal, vol. 50, no. 5, pp. 525–528, 2009.

[9] S. A. Othman, L. P. Majawit, W. N. W. Hassan, M. C. Wey, and R. M. Razi, “Anthropometric study of three-dimensional facial morphology in Malay adults,” PLOS ONE, vol. 11, no. 10, Art. no. e0164180, 2016.

[10] T. Y. Lu, K. Kadir, W. C. Ngeow, and S. A. Othman, “The prevalence of double eyelid and the 3D measurement of orbital soft tissue in Malays and Chinese,” Scientific Reports, vol. 7, Art. no. 14829, 2017.

[11] V. Packiriswamy, P. Kumar, and K. G. M. Rao, “Photogrammetric analysis of palpebral fissure dimensions and its position in Malaysian South Indian ethnic adults by gender,” North American Journal of Medical Sciences, vol. 4, no. 10, pp. 458–462, 2012.

[12] W. C. Ngeow and S. T. Aljunid, “Craniofacial anthropometric norms of Malaysian Indians,” Indian Journal of Dental Research, vol. 20, no. 3, pp. 313–319, 2009.

[13] N. P. Murray, M. Hunfalvay, and T. Bolte, “The reliability, validity, and normative data of interpupillary distance and pupil diameter using eye-tracking technology,” Translational Vision Science & Technology, vol. 6, no. 4, Art. no. 2, 2017.

[14] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” BMC Bioinformatics, vol. 7, Art. no. 91, 2006.

[15] G. S. Collins, K. G. M. Moons, P. Dhiman, and et al., “TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, Art. no. e078378, 2024.

[16] G. Varoquaux, “Cross-validation failure: Small sample sizes lead to large error bars,” NeuroImage, vol. 180, pp. 68–77, 2018.

[17] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. Cham, Switzerland: Springer, 2019.

[18] F. E. Harrell Jr., Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd ed. Cham, Switzerland: Springer, 2015.

[19] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” The Lancet, vol. 327, no. 8476, pp. 307–310, 1986.