Machine Learning–Based Prediction of Postoperative Refraction in Cataract Surgery: A Stacking Ensemble Approach

M. Gabar Zeyadi[1] and Selcan Ipek-Ugay[1]

Affiliation

[1]Department of Computer Science and Media, Medical Informatics Section (FSI), Berlin University of Applied Sciences (BHT), Germany

Correspondence: selcan.ipek-ugay@bht-berlin.de

ABSTRACT

Background: Achieving precise postoperative refractive outcomes remains a significant challenge in cataract surgery. While advanced intraocular lens (IOL) power calculation formulas exist, they are constrained by their singular algorithmic structures. This study investigated whether a stacking ensemble machine learning approach could overcome these limitations.

Methods: A dataset of 1,710 eyes from patients who underwent cataract surgery with monofocal IOL implantation (Vivinex or SA60AT) was utilized. Following rigorous preprocessing and feature engineering, a stacking ensemble architecture was developed comprising three diverse base learners (Multi-Layer Perceptron, Support Vector Regressor with RBF kernel, and SplineTransformer with Linear Regression) and a Ridge Regressor meta-learner. The model was trained on 80% of the data using 5-fold cross-validation and evaluated on an independent 20% test set (n=341). Performance was compared against six standard IOL formulas.

Results: The stacking ensemble model demonstrated excellent predictive accuracy, achieving a Mean Absolute Error (MAE) of 0.272 D on the independent test set (n=341). The model achieved lower MAE compared to all six standard IOL formulas, including Kane (MAE 0.295 D) and Barrett Universal II (MAE 0.318 D). Clinically, 85.1% of eyes achieved predictions within ±0.50 D, compared to 82.5% for Kane formula and 81.8% for Barrett Universal II.

Conclusion: The stacking ensemble machine learning model significantly enhances postoperative refraction prediction accuracy compared to established IOL calculation formulas. By leveraging algorithmic diversity and data-driven learning, this approach represents a promising advancement toward reducing refractive surprises and improving patient satisfaction in cataract surgery. External validation on independent datasets is required to confirm generalizability.

INTRODUCTION

Cataract surgery is among the most frequently performed surgical procedures globally, with modern objectives extending beyond visual restoration to achieving precise postoperative refractive outcomes [1]. Beyond removing the opacified crystalline lens, the key objective of modern surgery is to achieve a postoperative refractive outcome close to emmetropia, near 0.00 D.

The accurate calculation of intraocular lens (IOL) power directly determines the postoperative spherical equivalent (SEQ) and patient visual quality of life [2]. However, achieving this precision remains challenging, with “refractive surprises” clinically significant deviations from intended outcomes causing patient dissatisfaction and potentially requiring additional corrective procedures [3].

The fundamental challenge in IOL power calculation is predicting the postoperative Effective Lens Position (ELP), the final stabilized position of the IOL within the eye’s capsular bag [4,5]. This parameter cannot be measured preoperatively and must be estimated based on biometric data. While modern optical biometers provide highly accurate measurements, ELP prediction remains the primary source of refractive prediction error [4].

IOL calculation formulas have evolved from simple regression-based methods to sophisticated theoretical models [6]. Third-generation formulas (SRK/T, Hoffer Q, Holladay 1) incorporated optical principles but relied on limited biometric variables [7]. Fourth-generation formulas (Haigis, Barrett Universal II) integrated additional parameters to refine ELP predictions [8,9]. Recently, artificial intelligence-enhanced formulas like Kane have emerged as clinical benchmarks, combining theoretical optics with machine learning to optimize predictions [10,11]. Hybrid and AI-based formulas such as Kane, Hill-RBF, PEARL-DGS, Nallasamy and others have leveraged large datasets and machine learning to further improve accuracy [11].

Despite these advances, significant challenges persist. Even state-of-the-art formulas achieve ±0.50 D accuracy in only 75-90% of cases,[12] and performance degrades in eyes with atypical biometry. Each formula, regardless of sophistication, represents a single algorithmic hypothesis about the biometric-refractive relationship, creating an inherent performance ceiling.

Machine learning offers a potential solution through ensemble methods, which combine multiple diverse algorithms to achieve more robust predictions than any single model [13,14]. The Nallasamy formula recently demonstrated proof-of-concept for stacking ensemble approaches in IOL calculation, achieving very good accuracy across diverse eye types [15]. However, the potential of advanced stacking ensembles remains largely unexplored in this domain.

A key limitation of most existing approaches is their reliance on a single, fixed algorithmic structure. Ensemble learning offers a principled way to overcome this limitation. By combining multiple diverse models, ensemble methods can exploit complementary strengths, reduce individual biases, and improve robustness.

Stacking is a particularly powerful ensemble strategy, in which a second-level “meta-learner” is trained to optimally combine predictions from multiple base learners. In IOL power calculation, the Nallasamy formula has provided an important

proof-of-concept for stacking-based ensemble models, demonstrating performance gains over several standard formulas. However, the optimal stacking architecture choice and diversity of base learners, selection of meta-learner, and their relative contributions remains incompletely explored.

This study aimed to develop and validate a stacking ensemble machine learning model for predicting postoperative SEQ in cataract surgery and to compare its performance against established IOL calculation formulas, including the current clinical benchmarks.

We hypothesized that: (1) a stacking ensemble with diverse base learners would outperform individual models and traditional formulas; (2) the ensemble would achieve superior accuracy compared to current benchmarks (Kane, Barrett Universal II); and (3) algorithmic diversity would provide robust performance across different biometric ranges.

METHODS

Study Design and Dataset

This study utilized a dataset of 1,710 eyes from 1,710 patients who underwent cataract surgery at the Augen- und Laserklinik Castrop Rauxel, Germany.

The dataset consisted of two subsets based on the implanted monofocal intraocular lens (IOL), 888 eyes received the Vivinex (Hoya Surgical Optics) [16] and 822 eyes received the SA60AT (Alcon) [17].

Patient Eligibility Criteria

The inclusion criteria required patients to have undergone successful cataract surgery, with no history of prior ophthalmic surgical interventions, no perioperative or postoperative complications, and no ocular pathologies that could influence biometric measurements, including corneal or retinal pathologies. The refractive power of the inserted lens and the postoperative refraction (sphere and cylinder) were measured by an experienced optometrist at least 4 weeks after surgery. The datasets included only data with a postoperative Snellen decimal visual acuity of 0.8 (20/25 Snellen lines) or higher to ensure reliable refraction measurements.

Biometric Measurements

Preoperative biometry was performed using the IOLMaster 700 (Carl Zeiss Meditec, Jena, Germany). Initial parameters included axial length (AL), central corneal thickness (CCT), anterior chamber depth (ACD), aqueous depth (AQD), lens thickness (LT), corneal radii (R1, R2), keratometry (K1, K2), and white-to-white distance (WTW). Additionally, the refractive power of the inserted intraocular lens (IOL) and postoperative refraction measurements (SEQ) were recorded.

(SEQ: spherical equivalent power of postoperative refraction)

Data Preprocessing and Feature Engineering

Multicollinearity analysis using Variance Inflation Factor (VIF < 5 threshold):

- High correlations were observed between ACD and AQD. ACD was retained over AQD.
- K1 and K2 (keratometry values) were retained over R1 and R2 (corneal radii). Keratometry readings, measured in diopters, directly represent the refractive power of the cornea and are therefore more directly relevant to the task of predicting a refractive outcome.
- CCT was excluded due to low correlation with SEQ (Pearson r = 0.08, p = 0.31) and minimal contribution in recursive feature elimination.

The final feature set comprised eight variables: AL, ACD, LT, K1, K2, WTW, IOL power, and IOL type. The outcome variable (target) was SEQ.

The clinical data are summarized in Table 1.

Parameters	**Mean ± SD**
Mean average K (D)	43.70 ± 1.54
AL (mm)	23.64 ± 1.53
ACD (mm)	3.12 ± 0.41
LT (mm)	4.62 ± 0.45
IOL power (D)	21.64 ± 4.30
SEQ (D)	-0.52 ± 0.83

Table 1 *: Summary of clinical data in the entire dataset (n = 1710)*

Categorical Feature Encoding: Machine learning algorithms operate on numerical data, so the categorical feature IOLtype must be converted to numerical values. The two text values, 'Vivinex' and 'SA60AT', were converted to numerical representations using the LabelEncoder from the Scikit-learn library, which assigns integer values (0 and 1) to the two categories. For a binary categorical variable, this method is computationally efficient and functionally equivalent to one-hot encoding.

Numerical Feature Scaling: The numerical features in the dataset possess different scales and units (e.g., AL in mm, K1 in D). Many machine learning algorithms, including the Support Vector Regressor (SVR) and Multi-layer Perceptron (MLP) used in this study, are sensitive to the scale of input features. To ensure that all features contribute equitably to the model's learning and to accelerate gradient-based optimization, feature scaling was applied. The StandardScaler from Scikit-learn was used to transform each numerical feature. This process standardizes the features by

subtracting the mean and scaling to unit variance, resulting in a distribution with mean 0 and standard deviation 1. A critical protocol was strictly followed to prevent data leakage from the test set into the training process. The StandardScaler object was fit exclusively to the training data, and the resulting scaler was subsequently used to transform both the training and the test data. This rigorous approach ensures that the model evaluation is unbiased and provides a true estimate of the model's generalization performance on new data.

Data Partitioning Strategy: To ensure a robust, unbiased evaluation of the final model's performance, the dataset was split into a training set and an independent test set. An 80:20 split was employed, allocating 80% of the data (n=1368) for model development and the remaining 20% (n=341) as a hold-out test set. A stratified random sampling approach was used for this partition. Stratification was performed by IOL type to ensure proportional representation of the two lens types in both the training and test sets. This prevents distributional bias and ensures the test set provides a fair, representative sample for final performance assessment. The training set was used for all model development activities, including the training of base learners, hyperparameter optimization, and cross-validation. The test set was strictly held out and was used only once at the conclusion of the research for the final, unbiased performance evaluation of the chosen model and for its comparison against established IOL calculation formulas. This strict separation is a fundamental principle of machine learning methodology, critical for avoiding overfitting and accurately assessing the model's true generalization capability on unseen data.

Rationale for Ensemble Learning

To achieve the highest possible predictive accuracy, this study adopted a stacking ensemble machine learning framework. Ensemble methods combine the predictions of multiple models to produce a final prediction that is more accurate and robust than any of the individual constituent models. The core principle behind the stacking approach is that by combining the outputs of several different, well-performing models (known as base learners or Level-1 models), a second-level model (the meta-learner) can learn to optimally weigh these predictions, effectively correcting for the errors and biases of the individual models. This framework leverages the principle of model diversity; different algorithms, such as kernel methods, neural networks, and linear models, learn various aspects of the complex, non-linear relationships between the input features and the target variable. Stacking provides a principled, data-driven methodology for integrating diverse perspectives into a single, more powerful predictive system.

Algorithm Selection

The selection of constituent models for the stacking ensemble was performed through a systematic, multi-stage process designed to identify a diverse set of high-performing algorithms. First, a comprehensive screening of a wide array of supervised learning regression algorithms was conducted. These candidate models were trained and evaluated individually to establish a performance baseline. The algorithms tested included tree-based methods (XGBoost, LightGBM, CatBoost, Random Forest) [18, 19], kernel methods (Kernel Ridge, SVR, Gaussian Process Regressor) [20], linear models (Ridge, ElasticNet, Huber, SplineTransformer + Linear Regression) [21,22,23], and neural networks (Multi-Layer Perceptron Regressor) [24].

Second, following individual performance evaluations, the best-performing models are characterized by the lowest Mean Absolute Error (MAE). During this selection phase, overfitting was carefully checked. The coefficients of the meta learner were analysed to gain insights into the relative weights and contributions assigned to each base model's predictions, providing a deeper understanding of the ensemble's decision-making process.

Stacking Ensemble Architecture

A stacking ensemble regression framework was developed to predict the postoperative spherical equivalent (SEQ) (Figure 1). The architecture consisted of two levels:

- Level-0 (Base Learners): Three diverse algorithms were selected to capture different data patterns: a Multi-Layer Perceptron (MLP) regressor for non-linear interactions [24], a Support Vector Regressor (SVR) with a radial basis function (RBF) kernel [20], and a SplineTransformer with Linear Regression [22,23].
- Level-1 (Meta-Learner): A Ridge Regressor [21] was employed to optimally weight the out-of-fold predictions from the base learners.

Hyperparameters for all models were optimized using Bayesian optimization (Optuna) [25] with 5-fold cross-validation on the training set, minimizing the Mean Absolute Error (MAE), preventing leakage and overfitting.

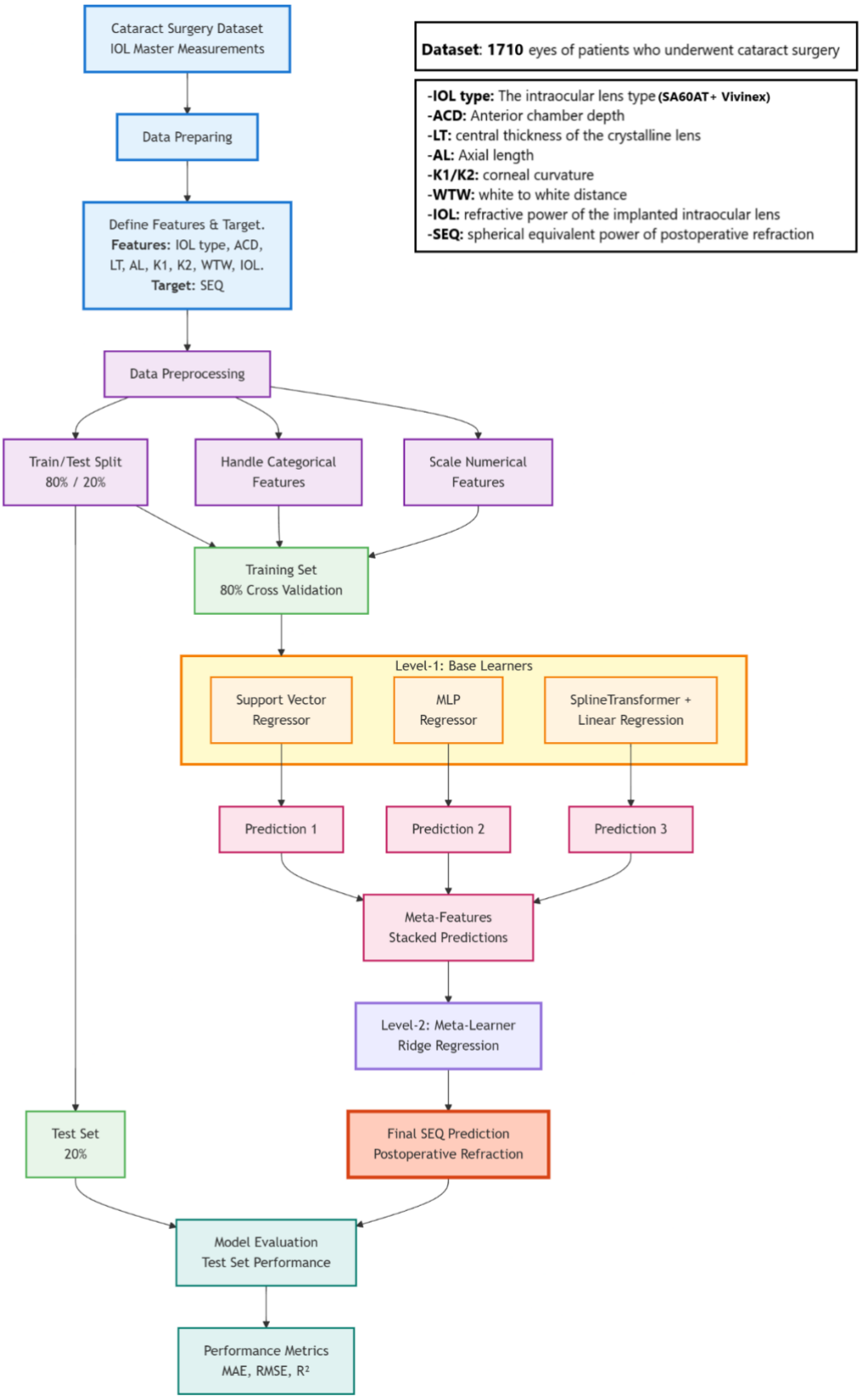


Figure 1: *Stacking Ensemble Pipeline for SEQ Prediction*

Performance Evaluation and Statistical Analysis

The fundamental unit of evaluation is the prediction error, which was defined for each eye as:

Error = actual postoperative refraction − predicted postoperative refraction

The model's predictive performance was evaluated on the independent test set using Mean Absolute Error (MAE), Median Absolute Error (MedAE), and Root Mean Squared Absolute Error (RMSAE), Standard Deviation (SD), Mean Error (ME), R-squared (R^2).

Clinical accuracy was quantified as the percentage of eyes with absolute prediction error within ±0.25 D, ±0.50 D, ±0.75 D, and ±1.00 D. A subgroup performance analysis was conducted based on axial length (AL).

Residuals were analysed for bias and distribution, normality was tested using the Shapiro–Wilk test. The Agreement between predicted and actual SEQ was further assessed with Bland–Altman analysis (mean bias and 95% limits of agreement).

Comparative Benchmark

The stacking model was benchmarked against six standard IOL calculation formulas: Barrett Universal II, Kane, Haigis, Hoffer Q, Holladay 1, and SRK/T. The statistical significance of the difference between the performance in the testing set was assessed using a paired Wilcoxon test with Bonferroni correction. Statistical significance was set at $p < 0.05$. Bootstrap resampling with 1,000 iterations was used to compute 95% confidence intervals for all performance metrics.

Barrett Universal II and Kane predictions were obtained via their respective online calculators[10], [26]. Holladay 1 [27], Hoffer Q [28,29], SRK/T [7,30], and Haigis [8], [31] were implemented in Python based on the published equations. For each IOL model, formula-specific lens constants were optimized using the training dataset. The optimal constant was chosen by zeroing the mean prediction error. The optimized constant then applied when evaluating performance in the test set.

SHAP (SHapley Additive exPlanations):

Employed for model interpretability. SHAP analysis quantified feature contributions to predictions, providing transparency and validating clinical plausibility [32].

Software and Tools Used

The implementation of the machine learning models and the associated data preprocessing steps relied heavily on Python and its rich scientific computing ecosystem. Key software and tools utilized include:

- **Python**: The primary programming language for all computational tasks [33].
- **Scikit-learn (sklearn):** A comprehensive machine learning library providing efficient tools for data preprocessing [34].
- **Pandas:** An essential library for data manipulation and analysis, particularly for handling tabular data structures (DataFrames) [35].
- **NumPy:** The fundamental package for numerical computing in Python, providing support for large, multi-dimensional arrays and matrices[36].
- **SciPy**: Used for advanced statistical analysis, including the specific statistical tests required for the comparative framework [37].
- **Optuna:** An optimization framework used for automated hyperparameter tuning via Bayesian optimization, enabling efficient search of the hyperparameter space to identify optimal model configurations [25].
- **Tabulate**: Utilized to generate clean, publication-ready tables for presenting results in the console and output documents [38].
- **Matplotlib and Seaborn**: Used for data visualization during the exploratory data analysis phase and for generating plots and figures to present the final results [39].
- **Google Colaboratory**: Computing environment [40].

RESULTS

The stacking ensemble model achieved high predictive performance on the independent test set of 341 eyes. The stacking ensemble model achieved a Mean Absolute Error (MAE) of 0.272 D, a Mean Squared Error (MSE) of 0.121 D^2, and a Standard Deviation (SD) of 0.349 D in the prediction errors. Furthermore, the model achieved a Coefficient of Determination (R^2) of 0.808, indicating that it successfully explains approximately 80.8% of the variance in postoperative SEQ outcomes in the test set. The Mean Error (ME) of 0.013 D, close to zero, indicates a minimal systematic bias in the predictions. These performance metrics are summarized in **Table 2**, and **Figure 2** shows a scatter plot of Predicted vs. Actual Spherical Equivalent (SEQ).

Metric	Value (D)
Mean Absolute Error (MAE)	0.272
Mean Squared Error (MSE)	0.121
R-squared (R^2)	0.808
Mean Error (ME)	0.013
Standard Deviation of Prediction Error (SD)	0.349

Table 2: *Overall Performance Metrics of the Stacking Ensemble Model on the Independent Test Set (n=341 eyes).*

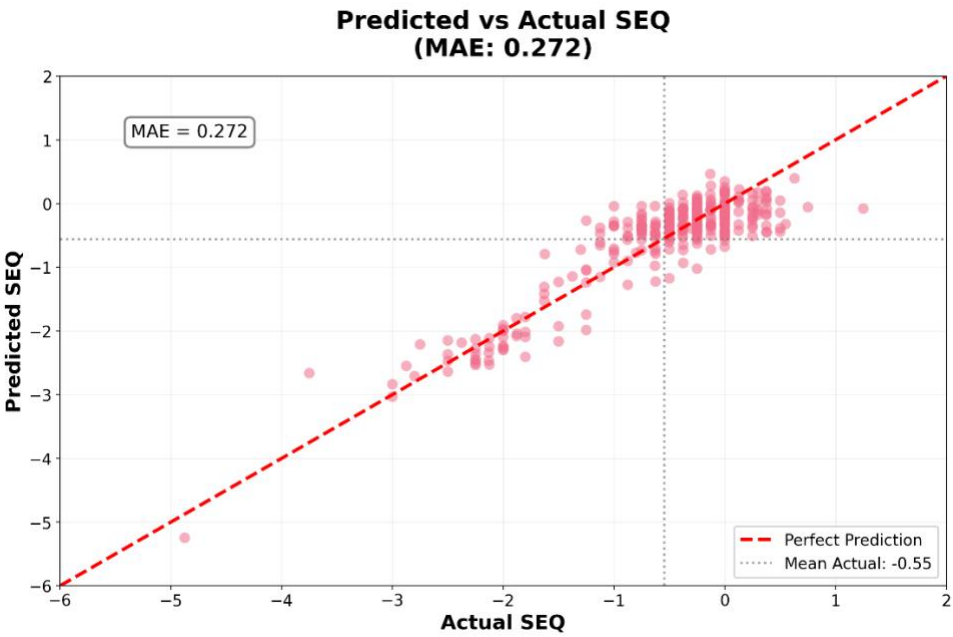


Figure 2: *Predicted vs. Actual Spherical Equivalent (SEQ).*
This scatter plot illustrates the agreement between the predicted and actual postoperative SEQ values. The solid diagonal line represents the line of perfect prediction.

Training versus test performance showed excellent consistency (training MAE 0.266 D vs. test MAE 0.272 D), with a test-to-training MSE ratio of 1.039, indicating no overfitting. A summary of this comparative analysis is presented in **Table 3.**

Metric	Training Set	Test Set	Difference	Ratio
MAE	0.266	0.272	0.0060	1.0226
MSE	0.116	0.121	0.0046	1.0394
R^2	0.835	0.808	0.0270	0.9676
SD	0.340	0.349	0.0090	1.0265

Table 3*: Training vs Test Performance Comparison.*

MSE: Mean Squared Error, MAE: Mean Absolute Error, R^2: R-squared, SD: Standard Deviation of the Prediction Error.

Residual Analysis

An analysis of the prediction errors, or residuals (Error = Actual SEQ – Predicted SEQ), was performed on the test set to assess systematic bias and the distribution of errors. The mean of the residuals was 0.013 D, a value very close to zero, indicating the absence of any significant systematic bias. The model does not consistently over- or under-predict the postoperative SEQ. The standard deviation of the residuals was 0.349 D, quantifying the typical spread of the prediction errors around the mean. A summary of this analysis is provided in **Table 4**. **Figure 3** shows a scatter plot of residuals vs. Actual Spherical Equivalent (SEQ).

Statistic	Value
Mean	0.013491
Standard Deviation	0.3490
Min	-1.087473
Max	1.330069
Range	2.417542

***Table 4**: Residual Analysis*

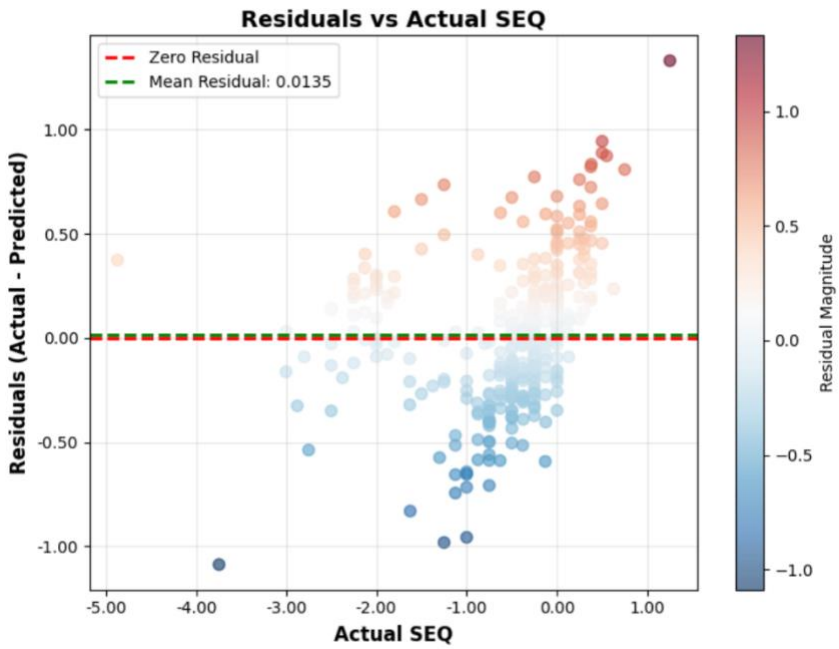
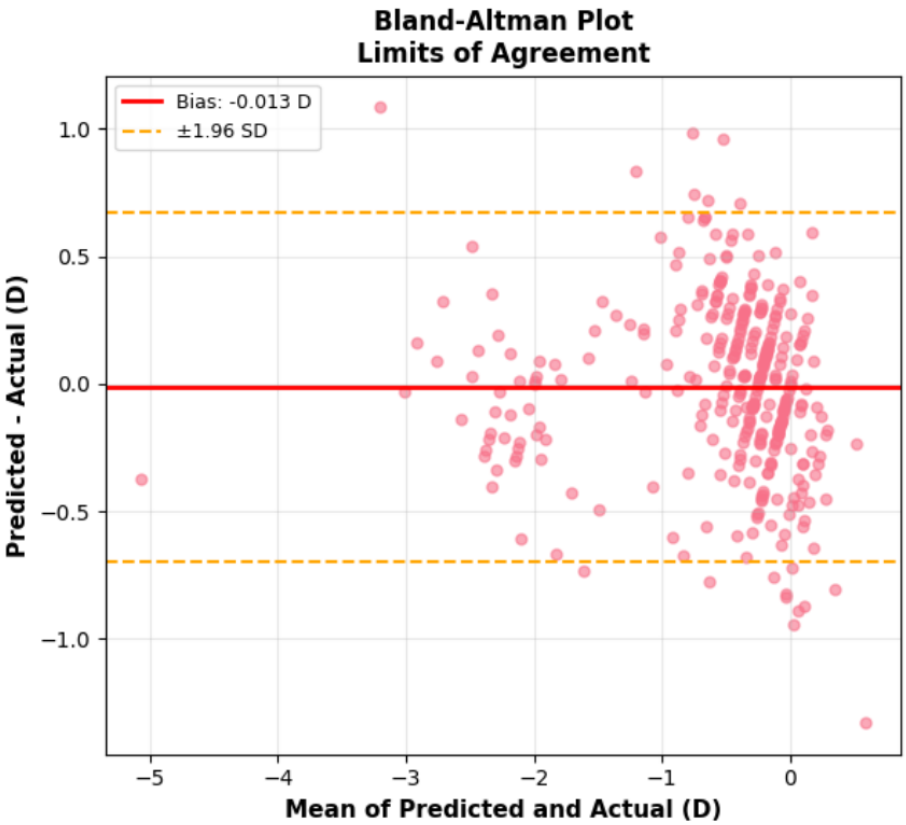


***Figure 3:** Residuals vs. Actual Spherical Equivalent (SEQ)*

Precision and Limits of Agreement: The standard deviation (SD) of the prediction error was 0.349 D. The Bland-Altman analysis established the 95% Limits of Agreement (LOA) between the predicted and actual SEQ at [-0.696 D, 0.669 D]. The 95% Confidence Interval (CI) for the mean signed error was narrow, ranging from -0.051 D to 0.024 D, confirming that the systematic bias is close to zero. **Figure 4** shows a scatter plot of the Bland-Altman analysis.



Mean Signed Error (Bias): -0.013 D [95% CI -0.051 to 0.024]
Bland-Altman Bias: -0.013 D, SD: 0.349 D
Limits of Agreement: [-0.696, 0.669] D

***Figure 4**: Bland-Altman analysis*

Comparative Benchmark Against Standard IOL Formulas

The stacking ensemble model was benchmarked against six standard IOL calculation formulas on the independent test set of 341 eyes. The statistical prediction error metrics (MAE, MedAE, RMSAE and SD) are presented in Table 5 and visualized in Figure 5**.** The stacking model demonstrated the lowest MAE (0.272 D) among all evaluated methods, followed by Kane formula (MAE 0.295 D), Haigis (0.316 D), and Barret Universal II (0.318 D).

Model/Formula	MAE [95% CI]	MedAE	RMSAE	SD	P-Value
Stacking Model	0.272 [0.249-0.295]	0.215	0.348	0.349	--
Kane	0.295 [0.270–0.320]	0.230	0.377	0.365	0.162
Haigis	0.316 [0.290–0.342]	0.253	0.402	0.402	< 0.001
Barrett Universal II	0.318 [0.288–0.350]	0.250	0.432	0.423	0.004
Holladay1	0.341 [0.312–0.372]	0.273	0.444	0.444	< 0.001
SRK/T	0.348 [0.317–0.380]	0.271	0.458	0.458	< 0.001
HofferQ	0.357 [0.326–0.388]	0.289	0.461	0.461	< 0.001

Table 5: *Prediction Performance summary in the testing set (n=341 eyes).*
MAE: Mean Absolute Error with 95% confidence intervals computed via bootstrap resampling (1,000 iterations), MedAE: Median Absolute Error, RMSAE: Root Mean Squared Absolute Error, SD: Standard Deviation of the Prediction Error. P-values from paired Wilcoxon signed-rank test comparing each formula to the Stacking Model.

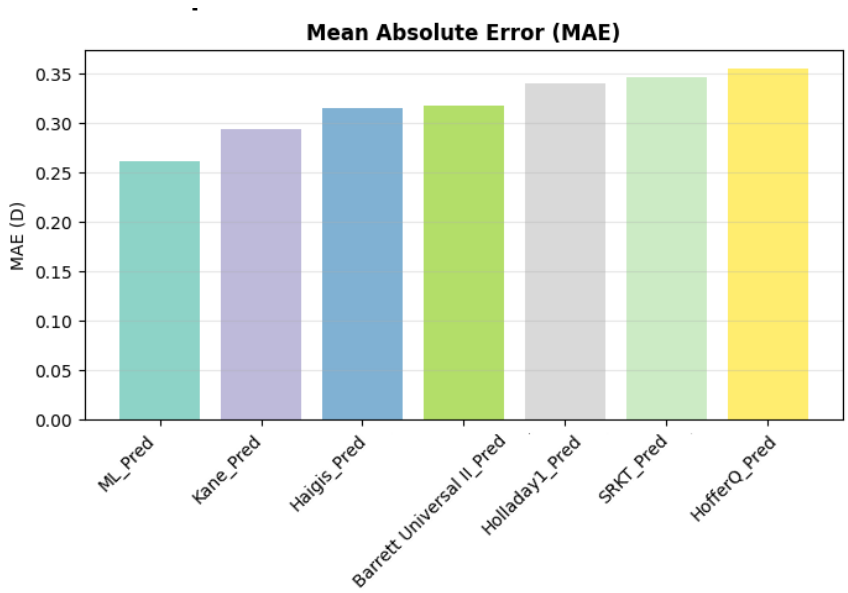


***Figure 5**: Prediction Performance for the stacking model and the six standard formulas*

Comparison of Clinical Accuracy Thresholds

The clinical efficacy of each method was compared by calculating the percentage of eyes achieving an absolute prediction error within the four standard clinical thresholds. As shown in **Table 6** and visualized in **Figure 6**, the stacking model consistently outperformed all standard formulas.

Within ±0.50 D, the stacking model correctly predicted 85.1% of cases, compared to 82.5% for Kane and 81.8% for Barrett Universal II. At the ±0.25 D threshold, the stacking model achieved 54.7% , versus 53.8% for Kane and 51.3% for Barrett Universal II.

Model/ Formula	**±0.25 D**	**±0.50 D**	**±0.75 D**	**±1.00 D**
Stacking Model	54.7 %	85.1 %	96.2 %	99.4 %
Kane Formula	53.8 %	82.5 %	95.6 %	98.2 %
Barrett Universal II	51.3 %	81.8 %	92.1 %	97.1 %
Haigis Formula	49.4 %	81.3 %	94.7 %	98.0 %
Holladay 1 Formula	45.6 %	74.9 %	90.4 %	97.1 %
SRK/T Formula	45.6 %	74.6 %	89.8 %	96.8 %
Hoffer Q Formula	45.0 %	71.3 %	91.8 %	96.2 %

***Table 6**: The percentage of Absolute Prediction Errors within Clinically Significant Thresholds for each formula, calculated based on the results in the testing dataset*

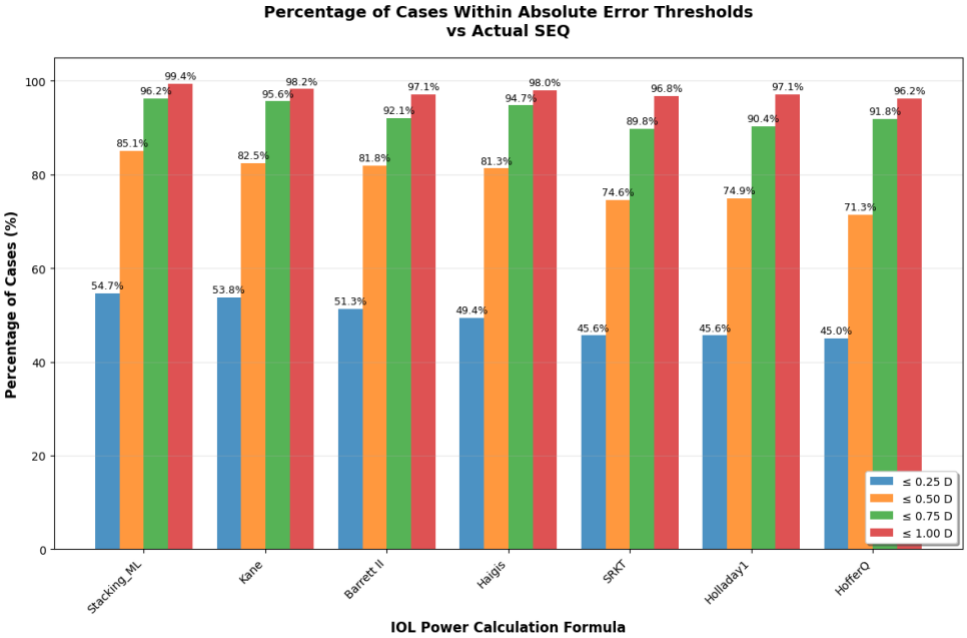


***Figure 6**: The percentage of Absolute Prediction Errors within Clinically Significant Thresholds for each formula, calculated based on the results in the testing dataset*

Subgroup Analysis by Axial Length

Performance varied across eye sizes. As shown in **Table 7** and visualized in **Figures 7 and 8**, the stacking model demonstrated the lowest MAE in the medium AL group and the short AL group.

In the short AL group, the stacking model's MAE of (0.386 D) was lower than Kane's (0.397 D). In the long AL group, the Kane achieved an MAE of 0.205 D, compared to 0.231 D for the stacking model.

Axial Length Group	Model / Formula	N	MAE	MedAE	RMSAE	SD
Short AL (<22.0 mm)	**Stacking Model**	40	**0.386**	**0.312**	**0.467**	**0.466**
	Kane Formula	40	0.397	0.350	0.485	0.478
	Haigis	40	0.410	0.352	0.526	0.532
	Holladay 1	40	0.453	0.381	0.573	0.566
	SRK/T	40	0.474	0.415	0.618	0.625
	Barrett Universal II	40	0.508	0.398	0.687	0.695
	Hoffer Q Formula	40	0.543	0.500	0.674	0.587
Medium AL (22.0–26.0 mm)	**Stacking Model**	273	**0.259**	**0.199**	**0.334**	**0.335**
	Kane	273	0.289	0.225	0.370	0.354
	Barrett Universal II	272	0.298	0.232	0.393	0.380
	Holladay 1	273	0.308	0.252	0.403	0.400
	Haigis	273	0.308	0.250	0.390	0.389
	Hoffer Q	273	0.324	0.257	0.420	0.420
	SRK/T	273	0.333	0.261	0.437	0.434
Long AL (>26.0 mm)	**Kane**	28	**0.205**	**0.160**	**0.253**	**0.257**
	Stacking Model	28	0.231	0.216	0.283	0.274
	Barrett Universal II	28	0.238	0.240	0.302	0.272
	Haigis	28	0.249	0.212	0.303	0.293
	SRK/T	28	0.309	0.276	0.389	0.310
	Hoffer Q	28	0.411	0.435	0.474	0.304
	Holladay 1	28	0.504	0.527	0.586	0.320

***Table 7**: Prediction Performance Analysis by Axial Length.*
MAE: Mean Absolute Error, MedAE: Median Absolute Error, RMSAE: Root Mean Squared Absolute Error, SD: Standard Deviation of the Prediction Error, N: Number of cases.

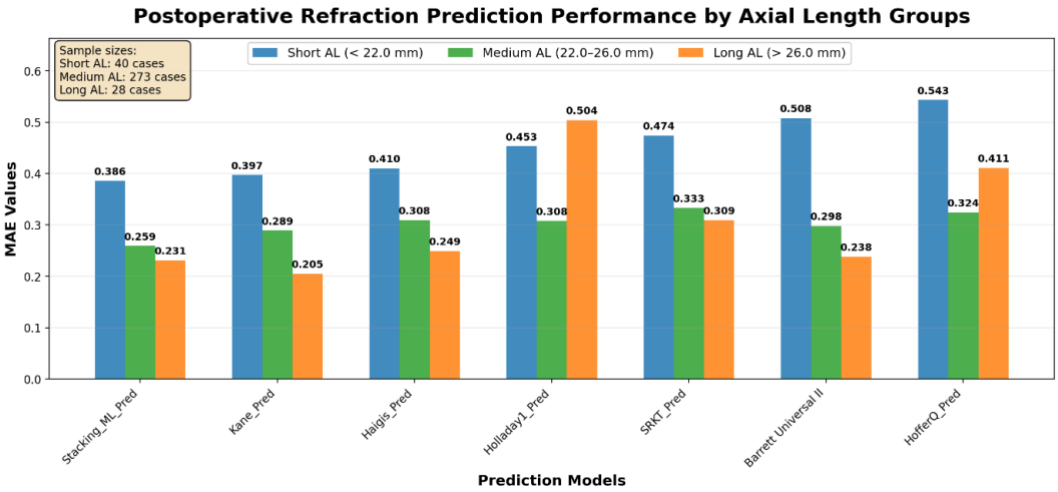


***Figure 7**: Prediction Performance Analysis by Axial Length*
MAE: Mean Absolute Error. AL: Axial length.

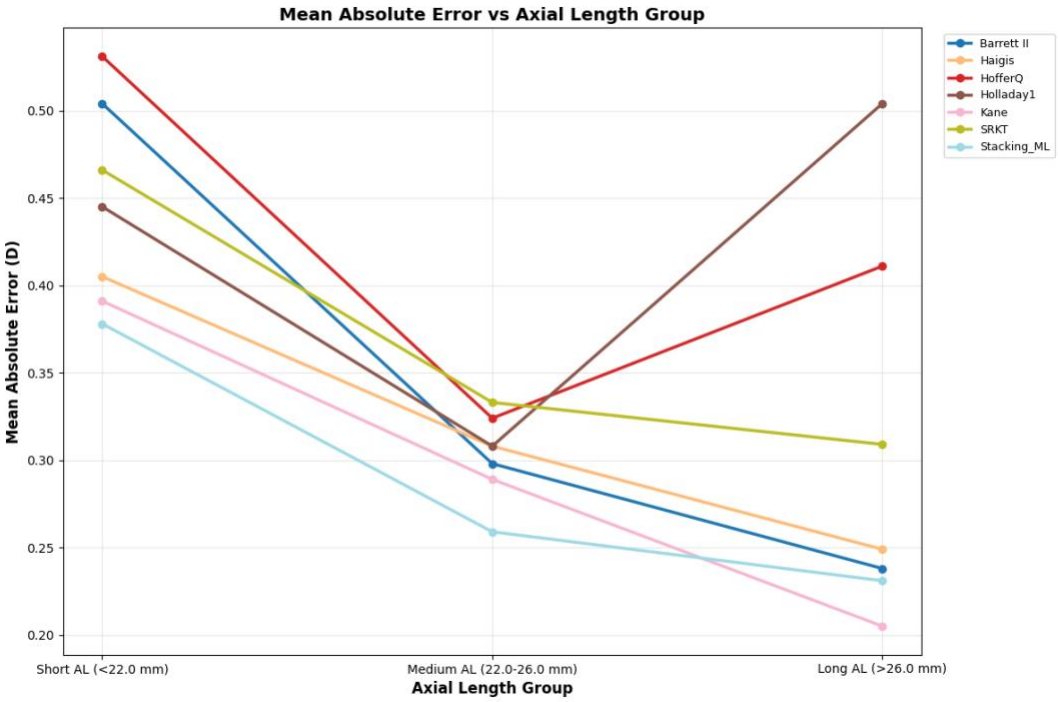


***Figure 8**: Mean Absolute Error vs Axial Length*

Model Interpretability SHAP (SHapley Additive exPlanations) analysis identified Axial Length and IOL Power as the most influential features driving the model's predictions, followed by keratometry (K1, K2) . This feature prioritization aligns with established optical principles, confirming the clinical plausibility of the model's decision-making process. The features influencing the stacking ensemble model's predictions are listed in **Figure 9**.

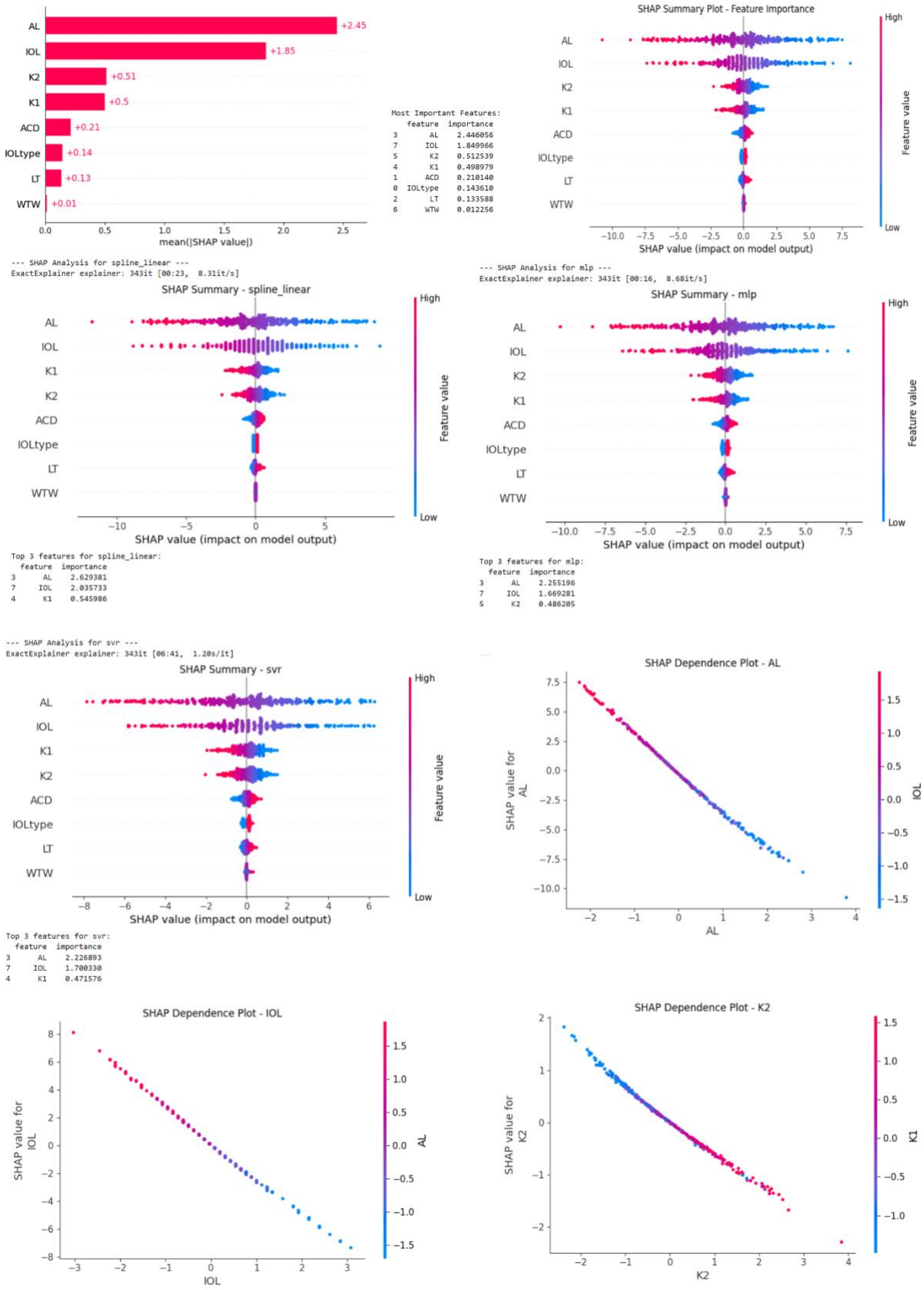


***Figure 1**: SHAP Analysis. The features influencing the stacking model's predictions.*

DISCUSSION

Principal Findings

Interpretation of Findings This study demonstrates that a stacking ensemble machine learning model can significantly outperform established theoretical and hybrid IOL calculation formulas. Our stacking model significantly outperforms Barrett Universal II, Haigis, Holladay1, SRK/T, and HofferQ. It shows comparable performance to the state-of-the-art Kane formula.

The stacking model achieved a Mean Absolute Error (MAE) of 0.272 D (**Table 5**) and predicted 85.1% of eyes within ±0.50 D of the target refraction (**Table 6**). This exceeds the performance of the current clinical gold standards, the Kane formula (MAE 0.295 D; 82.5% within ±0.50 D) and Barrett Universal II (MAE 0.318 D; 81.8% within ±0.50 D). These improvements translate directly to reduced refractive surprises and enhanced patient outcomes.

Model Architecture and Performance

The performance of the ensemble architecture is attributable to algorithmic diversity (**Figure 1**). By combining a neural network (MLP), a kernel-based method (SVR), and a spline-based linear model, the meta-learner could leverage the unique strengths of each. The MLP captures complex non-linear feature interactions, the SVR provides robustness to outliers with its epsilon-insensitive loss function, and the SplineTransformer offers flexible piecewise polynomial modelling. The Ridge meta-learner intelligently weights the predictions demonstrating data-driven synthesis that mitigates individual model weaknesses.

The model's generalization capability was rigorously validated. Minimal training-test performance discrepancy (MAE 0.266 D vs. 0.272 D) (**Table 3**), close cross-validation alignment, near-zero mean error, and tight Bland-Altman (**Figure 4**) limits of agreement collectively confirm the model learned true underlying patterns rather than overfitting training data.

Clinical Significance and Comparison with Current Standards

The 85.1% success rate at ±0.50 D represents clinically meaningful improvement. Recent literature confirms that 80-85% accuracy at this threshold represents excellent performance. [12]

Model Interpretability and Clinical Trust

SHAP analysis (**Figure 9**) addressed the critical “black box” concern inherent in complex machine learning models. The identification of AL and IOL power as dominant features, followed by keratometry values, precisely mirrors the theoretical foundation of IOL calculation. This alignment with established ophthalmological principles provides evidence that the model learns clinically relevant, causal relationships rather than spurious correlations. Such interpretability is essential for clinical adoption and regulatory acceptance.

Study Strengths

This study employed methodological rigor across multiple dimensions. The large, carefully curated dataset (1,710 eyes) with high-quality biometry provided substantial statistical power. Algorithmic diversity in base learner selection maximized ensemble capability. The benchmarking strategy ensured fair comparison by evaluating all methods on identical test data. Finally, proactive SHAP analysis addressed interpretability, a sophisticated and vital component of modern clinical AI research [34].

Limitations

Several limitations merit consideration. First, single-center data origin may limit external generalizability. Model performance may differ with other biometry devices, surgical techniques, or patient demographics, necessitating multi-center validation. Second, training exclusively on two monofocal IOL models (Vivinex and SA60AT) means generalizability to other IOL designs, particularly multifocal and Toric lenses, remains untested. Third, strict exclusion criteria removed patients with prior corneal refractive surgery (e.g., post-LASIK), significant ocular pathologies, or perioperative complications. While appropriate for foundational model development, performance in these complex cases which represent significant clinical challenges remains unproven. The model is currently for uncomplicated eyes only.

Clinical Implications and Implementation

The demonstrated outperformance suggests strong potential for clinical integration as a decision-support tool. When existing formulas provide divergent predictions, the model’s output could guide final IOL power selection. The ultimate goal is consistently reducing refractive surprises unexpected hyperopic or myopic outcomes that compromise visual quality and necessitate secondary procedures.

Future Research Directions

Multi-center external validation on large, heterogeneous datasets from multiple surgeons and biometry devices is essential. Model expansion should include diverse IOL designs and types (Multifocal, Toric). A dedicated research track should develop specialized models for post-refractive surgery eyes and other challenging scenarios.

CONCLUSION

This study demonstrates that stacking ensemble machine learning significantly improves postoperative refraction prediction accuracy in cataract surgery compared to established IOL calculation formulas. By leveraging algorithmic diversity and data-driven learning, this approach overcomes limitations of single-model methods and represents a promising advancement toward precision ophthalmology. The results support the integration of ensemble ML models as powerful clinical decision-support tools to reduce refractive surprises and improve patient satisfaction.

ACKNOWLEDGMENTS

The authors acknowledge Prof. Dr. Achim Langenbucher (Institute of Experimental Ophthalmology, Saarland University, Germany) for making the dataset publicly available, and the Augen- und Laserklinik Castrop-Rauxel, Germany, for the original data collection.

COMPETING INTERESTS

The authors declare no competing interests.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY

The dataset used in this study is publicly available through the Figshare repository: https://figshare.com/authors/Achim_Langenbucher/12446667. Preprocessed data are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

Python implementation code for the stacking ensemble model, hyperparameter optimization, and traditional IOL formula calculations is available upon request from corresponding authors. All analyses were performed using Python 3.x with scikit-learn, NumPy, pandas, and Optuna libraries.

ETHICS STATEMENT

The study used publicly available, de-identified data released by the data provider. The dataset contains no directly identifying patient information. No new data were collected for this study. According to institutional guidelines for secondary analysis of anonymized data, no additional ethical approval was required.

Distribution: CC BY-NC-ND (Attribution - NonCommercial - NoDerivatives)

REFERENCES

[1] J. Xu, X. Chen, and K. Yao, "Cataract: Advances in surgery and whether surgery remains the only treatment in future," Adv. Ophthalmol. Pract. Res., vol. 1, no. 1, p. 100008, Nov. 2021, doi: 10.1016/j.aopr.2021.100008.

[2] R. Khoramnia, G. Auffarth, G. Łabuz, G. Pettit, and R. Suryakumar, "Refractive Outcomes after Cataract Surgery," Diagnostics, vol. 12, no. 2, p. 243, Jan. 2022, doi: 10.3390/diagnostics12020243.

[3] J. Z. Kieval et al., "Prevention and management of refractive prediction errors following cataract surgery," J. Cataract Refract. Surg., vol. 46, no. 8, pp. 1189–1197, Aug. 2020, doi: 10.1097/j.jcrs.0000000000000269.

[4] "IOL Power Calculation Formulas Explained." .Accessed: Oct, 2025 Available: https://www.zeiss.com/meditec/en/c/optical-biometry/iol-power-calculation-formulas-explained.html

[5] J. G. Ladas and S. R. Lin, "Artificial Intelligence in Calculating the IOL Power," in Artificial Intelligence in Ophthalmology, A. Grzybowski, Ed., Cham: Springer International Publishing, 2021, pp. 257–262. doi: 10.1007/978-3-030-78601-4_23.

[6] W. Stopyra, A. Langenbucher, and A. Grzybowski, "Intraocular Lens Power Calculation Formulas—A Systematic Review," Ophthalmol. Ther., vol. 12, no. 6, pp. 2881–2902, Dec. 2023, doi: 10.1007/s40123-023-00799-6.

[7] J. A. Retzlaff, D. R. Sanders, and M. C. Kraff, "Development of the SRK/T intraocular lens implant power calculation formula," J. Cataract Refract. Surg., vol. 16, no. 3, p. 333, May 1990, doi: 10.1016/S0886-3350(13)80705-5.

[8] W. Haigis, B. Lege, N. Miller, and B. Schneider, "Comparison of immersion ultrasound biometry and partial coherence interferometry for intraocular lens calculation according to Haigis," Graefes Arch. Clin. Exp. Ophthalmol. Albrecht Von Graefes Arch. Klin. Exp. Ophthalmol., vol. 238, no. 9, pp. 765–773, Sep. 2000, doi: 10.1007/s004170000188.

[9] G. D. Barrett, “An improved universal theoretical formula for intraocular lens power prediction,” J. Cataract Refract. Surg., vol. 19, no. 6, p. 713, Nov. 1993, doi: 10.1016/S0886 3350(13)80339-2.

[10] Kane Formula. Accessed Oct, 2025. Available: https://www.iolformula.com

[11] W. Stopyra, D. L. Cooke, and A. Grzybowski, “A Review of Intraocular Lens Power Calculation Formulas Based on Artificial Intelligence,” J. Clin. Med., vol. 13, no. 2, p. 498, Jan. 2024, doi: 10.3390/jcm13020498.

[12] “Accuracy of IOL Power Calculation Formulas,” CRST Global. Accessed: Oct, 2025. Available: https://crstodayeurope.com/articles/2020-oct/accuracy-of-iol-power-calculation-formulas/

[13] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “Ensemble Learning for Disease Prediction: A Review,” Healthcare, vol. 11, no. 12, p. 1808, Jun. 2023, doi: 10.3390/healthcare11121808.

[14] B. Naderalvojoud and T. Hernandez-Boussard, “Improving machine learning with ensemble learning on observational healthcare data,” AMIA. Annu. Symp. Proc., vol. 2023, pp. 521–529, Jan. 2024.

[15] T. Li, J. Stein, and N. Nallasamy, “Evaluation of the Nallasamy formula: a stacking ensemble machine learning method for refraction prediction in cataract surgery,” Br. J. Ophthalmol., vol. 107, no. 8, pp. 1066–1071, Aug. 2023, doi: 10.1136/bjophthalmol-2021-320599.

[16] “VivinexTM IOLs - Hoya Surgical Optics Clarity & Support.” Accessed: Oct, 2025. Available: https://www.hoyasurgicaloptics.com/vivinex/iol

[17] “SA60AT.” Accessed: Oct, 2025. Available: https://www.alconnordicsurgicalproducts.com/node/1081

[18] “Ensembles: Gradient boosting, random forests, bagging, voting, stacking,” scikit-learn. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/modules/ensemble.html

[19] “What Is Random Forest? | IBM.” Accessed: Oct, 2025. Available: https://www.ibm.com/think/topics/random-forest

[20] “Support Vector Machines,” scikit-learn. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/modules/svm.html

[21] “Kernel ridge regression, scikit-learn”. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/modules/kernel_ridge.html

[22] “SplineTransformer, scikit-learn”. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.SplineTransformer.html

[23] “Polynomial and Spline interpolation, scikit-learn”. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_polynomial_interpolation.html

[24] “MLPRegressor,scikit-learn”. Accessed: Oct, 2025. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

[25] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. "Optuna: A Next-generation Hyperparameter Optimization Framework." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631 (2019).

[26] “Barrett Universal II Formula V1.05.” Accessed: Oct, 2025. Available: https://calc.apacrs.org/barrett_universal2105/

[27] J. T. Holladay, K. H. Musgrove, T. C. Prager, J. W. Lewis, T. Y. Chandler, and R. S. Ruiz, “A three-part system for refining intraocular lens power calculations,” J. Cataract Refract. Surg., vol. 14, no. 1, p. 17, Jan. 1988, doi: 10.1016/S0886-3350(88)80059-2.

[28] K. J. Hoffer, “The Hoffer Q formula: A comparison of theoretic and regression formulas,” J. Cataract Refract. Surg., vol. 19, no. 6, p. 700, Nov. 1993, doi: 10.1016/S0886-3350(13)80338-0.

[29] B. Zuberbuhler and A. J. Morrell, “Errata in printed Hoffer Q formula,” J. Cataract Refract. Surg., vol. 33, no. 1, p. 2, Jan. 2007, doi: 10.1016/j.jcrs.2006.08.054.

[30] “Erratum,” J. Cataract Refract. Surg., vol. 16, no. 4, p. 528, Jul. 1990, doi: 10.1016/S0886-3350(13)80820-6.

[31] W. Haigis, “Matrix-optical representation of currently used intraocular lens power formulas,” J. Refract. Surg. Thorofare NJ 1995, vol. 25, no. 2, pp. 229–234, Feb. 2009, doi: 10.3928/1081597X-20090201-09.

[32] Lundberg, S. M., & Lee, S. I. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems 30 (2017).

[33] Python Software Foundation. "Python Language Reference, version 3.x." Available at http://www.python.org.

[34] Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[35] W. McKinney, “Data Structures for Statistical Computing in Python,” in Proc. 9th Python Sci. Conf. (SciPy), 2010.

[36] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. "Array programming with NumPy." Nature 585, 357–362 (2020).

[37] Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” Nat. Methods, vol. 17, pp. 261–272, 2020.

[38] Astanin, S. "python-tabulate: Pretty-print tabular data." Accessed: Oct, 2025. Available at: https://pypi.org/project/tabulate/

[39] Waskom, M. L. "Seaborn: statistical data visualization." Journal of Open Source Software 6, no. 60 (2021): 3021.

[40] “Google Colab.” Accessed: Oct, 2025. Available: https://colab.research.google.com/