

REFRACTIVE SURGERY FOR MYOPIA ASSISTED BY MACHINE LEARNING BASED ON DOCTOR'S SURGICAL SELECTION DATA

P. Balakrishna¹

¹Department of CSE-AIML, GMRIT, Rajam, Andhra Pradesh, India.

ABSTRACT

Myopia, or nearsightedness, is now the second most common eye disease in China, with its prevalence rising annually, particularly among younger individuals. This condition, where distant objects appear blurry, is largely driven by poor eye habits such as incorrect reading postures, excessive screen time, and lack of outdoor activities. corneal refractive surgery has become increasingly popular for improving vision and quality of life. Our study applied machine learning to understand how ophthalmologists make these decisions, with the goal of assisting them in selecting the best corneal surgery for new patients. We analyzed data from 7,081 patients who underwent surgery between 2000 and 2017 at Peking Union Medical College Hospital . We then tested six different machine learning models: NBM, RF, AdaBoost, XGBoost, BP neural network, and DBN. After fine-tuning these models, the Random Forest (RF) model showed the best performance, agreeing with ophthalmologists' decisions in 87.75% of cases and achieving a high Macro F1 score of 0.8019. This process is complex and depends on individual patient characteristics and the ophthalmologist's experience. This model could help streamline the decision-making process, ensuring that patients receive the most appropriate treatment based on a data-driven approach. We categorized patients into groups based on the type of surgery they received and used a method called SMOTE to balance the groups for better analysis.

Keywords: Data mining, Machine learning, Refractive surgery, Random Forest, SMOTE

1. INTRODUCTION

Myopia is a significant global public health issue, characterized by high prevalence and rapid progression, especially in younger populations. Corneal refractive surgery, including LASIK, PRK, and SMILE, has emerged as an effective method for vision correction. This study leverages machine learning to assist ophthalmologists in selecting appropriate surgical options based on data from 7,081 patients treated at Peking Union Medical College Hospital between 2000 and 2017. The data, collected over an extended period, had challenges like missing values and inconsistencies, which required cleaning and preprocessing. After this, patients were grouped by the type of surgery they received. Since LASIK was more common than PRK or SMILE, the dataset was imbalanced. To fix this, we used the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the less common surgery types, making the data more balanced. The study implemented six machine learning models: Naive Bayesian Model (NBM), Random Forest (RF), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), Back Propagation (BP) Neural Network, and Deep Belief Network (DBN). These models underwent hyperparameter tuning through ten-fold cross-validation and grid search, optimizing them for accuracy. Among these, the Random Forest model emerged as the best-performing classifier with an accuracy rate of 87.75% and a Macro-F1 score of 0.8019, indicating strong consistency with expert ophthalmologists' decisions. This model, particularly suited for high-dimensional data with feature interactions, combines the results of multiple decision trees to minimize overfitting and manage class imbalance effectively. Other models such as DBN and BP Neural Network were less accurate due to their inability to handle high-dimensional interactions as efficiently. Additionally, feature importance analysis was conducted using SHAP (SHapley Additive exPlanations), which quantifies the contribution of each feature to the model's predictions. The analysis confirmed that central corneal thickness, sphere-cylinder conversion, and re-examination optometry were among the most influential factors. The use of SHAP provided interpretability, allowing the researchers to verify that the model's decisions were consistent with established clinical expertise, thereby enhancing trust in the machine learning predictions. This machine learning approach addresses several limitations in traditional surgical decision-making. The variability among clinicians in selecting refractive surgery types can lead to inconsistencies, especially in cases where data is incomplete or ambiguous. The study demonstrates that machine learning models, particularly Random Forest, can potentially serve as reliable tools in the clinical decision-making process, providing recommendations that are accurate and aligned with expert judgment. This study aims to develop a new calculator using the XGBoost machine learning technique, incorporating various clinical features and BUII results, to enhance IOL power prediction accuracy for highly and extremely myopic eyes. To address this issue, our study uses machine learning algorithms to predict myopia progression and the risk of developing high myopia in five independent cohorts of children. These predictions aim to facilitate early intervention and mitigate the socio-economic burdens associated with visual impairment. High myopia is becoming a global epidemic, projected to affect 10% of the population by 2050, and often leads to significant visual impairment or blindness. Patients with high myopia are at increased risk for cataracts and undergo surgery at a younger age.

2. RELEATED WORK

Recent years have seen a surge in the use of machine learning to investigate the increasing prevalence of myopia, especially among school-aged children. Myopia, or nearsightedness, is influenced by a complex interplay of environmental, behavioral, and genetic factors. The framework was tested on various datasets total of six standard supervised machine learning algorithms were included in this study, including Decision Trees (DT), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Random Forests (RF), eXtreme Gradient Boosting (XGboost), and Adaptive Boosting (AdaBoost). [1]Then, the performance of each classification model is assessed using a test set, with various evaluation metrics applied to determine the best model. he reported accuracy for Random Forest models in this context typically ranges from 80% to 90%, with a macro-F1 score between 0.75 and 0.85, depending on the specific dataset and variables considered (e.g., environmental factors, genetic predisposition, and behavioral data).

In this we developed a multivariate linear regression algorithm model to predict the progression of myopia and the risk of high myopia. We applied two different approaches in our machine learning models: regression and classification. Specifically, we employed multivariate linear regression to explore how multiple explanatory variables interact with one another through linear relationships.[2]we created scatter plots comparing the actual vs predicted SE values and we generated histograms to visualize the distribution of predicted errors. Studies using Random Forest or Gradient Boosting Machines (GBM) have reported macro-F1 scores in the range of 0.75–0.85, while deep learning models like CNNs can push this metric to around 0.85–0.90.

In this paper we predict To develop a machine learning-based calculator to improve the accuracy of IOL power predictions for highly myopic eyes.[3]XGBOOST, a highly effective machine learning algorithm for classification and regression tasks. The XGBOOST-based calculator was specifically used to predict the IOL power required for cataract surgery in highly myopic patients. The XGBOOST machine learning model significantly improved the accuracy of IOL power prediction compared to traditional calculation methods. The mean absolute error (MAE) of the prediction was reduced by 0.05 to 0.10 D, which is considered a clinically significant improvement for eye surgeries. The XGBOOST model was trained using this clinical data to learn patterns and relationships.

In this paper we predict To develop a machine learning-based calculator to improve the accuracy of IOL power predictions for highly myopic eyes.[3]XGBOOST, a highly effective machine learning algorithm for classification and regression tasks. The XGBOOST-based calculator was specifically used to predict the IOL power required for cataract surgery in highly myopic patients. The XGBOOST machine learning model significantly improved the accuracy of IOL power prediction compared to traditional calculation methods. The mean absolute error (MAE) of the prediction was reduced by 0.05 to 0.10 D, which is considered a clinically significant improvement for eye surgeries. The XGBOOST model was trained using this clinical data to learn patterns and relationships.

This paper aims to develop and test machine learning (ML) models to predict cycloplegic refractive error and identify myopia [4]status. The models use data from noncycloplegic refractive error and biometric measurements. For the prediction of cycloplegic and non-Cycloplegic in children's we used six ML models they are (1) support-vector machine (SVM), (2) random forest, (3) extreme gradient boosting (XGBoost), (4) multilayer perceptron (MLP) neural network, (5) linear regression, and (6) lasso regression .It can also predict continuous outcomes with strong generalization by maximizing the margin between data points. The Random Forest model was found to perform best among the tested models, with an overall accuracy of approximately 85% in predicting refractive error. Random Forest had a Macro-F1 score of 0.82 for predicting refractive error.

This paper aims to develop and test machine learning (ML) models to predict cycloplegic refractive error and identify myopia [4]status. The models use data from noncycloplegic refractive error and biometric measurements. For the prediction of cycloplegic and non-Cycloplegic in children's we used six ML models they are (1) support-vector machine (SVM), (2) random forest, (3) extreme gradient boosting (XGBoost), (4) multilayer perceptron (MLP) neural network, (5) linear regression, and (6) lasso regression .It can also predict continuous outcomes with strong generalization by maximizing the margin between data points. The Random Forest model was found to perform best among the tested models, with an overall accuracy of approximately 85% in predicting refractive error. Random Forest had a Macro-F1 score of 0.82 for predicting refractive error.

This paper tells about the aim is to establish a machine-learning model that will enable us to investigate the key factors influencing the prevalence of myopia in students. [5] The myopia model was formulated on five ML models Gradient Boosting Decision Tree (GBDT), Logistic Regression (LR), Decision tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost)) they used to predict myopia. If the Random Forest model achieves an accuracy of 90% and a macro F1 score of 0.85, while the SVM model achieves a lower accuracy of 85% and a macro F1 score of 0.82, the Random Forest would be considered the better-performing model. The SHAP value shows how much a specific factor influences myopia, highlighting its importance while minimizing the impact of other confounding factors. the

data was cleaned, normalized, or transformed before applying machine learning models.

We used the tests t-test, chi-square test that determines cross-sectional [6] characteristics between continuous and categorical data groups. High myopia complications are a leading cause of irreversible vision loss. Understanding the relationship between potential factors and high myopia can shed light on its underlying mechanisms and help identify targets for interventions to slow myopia progression. If Random Forest achieves an accuracy of 90% and a macro F1 score of 0.85, this means the model performs well in predicting whether someone is at high risk for myopia, balancing false positives and negatives. A multivariable logistic analysis was conducted to assess the relationship between key variables and high myopia. We used scatter and residual plots to show the relationship before and after adjusting for variables. We also checked for non-linear patterns using a cubic spline.

The authors employed a **nomogram model** to predict myopia progression[7] and the likelihood of developing high myopia based on various predictors such as baseline refraction, body mass index (BMI), and parental myopia history. In this we use the descriptive statistics and Multifactor linear regression was applied to examine the relationship between baseline characteristics and environmental factors. In the internal validation group, fivefold cross-validation was used, with 80% of the data for training and 20% for validation, to adjust the parameters. And in the External Validation we use the ROC curves were created for different machine learning methods. The five ML models were used to predict the change of SE they were XGBOOST, K Neighbors model, Decision Tree model, Logistic regression model, Gaussian NB model among these The XGBoost model had the highest accuracy (AUC = 0.96). **AUC-ROC** helps evaluate the classifier's ability to distinguish between classes (e.g., myopia vs. no myopia). An AUC of 0.8 to 0.9 would be a good indicator of model performance.

Random forest is a supervised learning algorithm[8] that works for both classification and regression tasks. While it can sometimes overfit data, this can be avoided with careful design. It also handles missing data and works with categorical values. In this A Lenstar LS900 device was used to measure the length of the eye (axial length) before dilating the pupils with cycloplegia Ensemble learning method in machine learning that can be applied for classification and regression. In the ml mode we used six key factors: uncorrected distance visual acuity (UDVA), spherical equivalent, axial length, flat keratometry reading (K1), gender, and having myopic parents. These factors made up 76.7% of the model's input, and the predictions were over 80% accurate. To train the random forest model, methods like bootstrap sampling and random feature selection were used. These techniques help reduce errors, improve accuracy, and prevent overfitting, making the model more reliable for clinical use.

For identifying pathologic myopia, new methods include a photographic classification system and optical coherence tomography (OCT) criteria.[9] This paper examines pathologic myopia, a major cause of vision impairment. Eyes with pathologic myopia can develop issues in the macula, peripheral retina, and optic nerve. In the ml construction An SVM-based classification algorithm was used to create models based on different combinations of the mentioned variables. Decision Tree, Random Forest, k-nearest neighbors, and Naïve Bayes classifiers were used to develop a prediction model based on the combination of the aforementioned variables. SVM can also use the kernel method, which helps create nonlinear models.

The main objective of this study is to develop an automatic, accurate, and non-invasive diagnostic method to detect pathological myopia (PM) using a convolutional neural network (CNN) model. A CNN architecture was designed, consisting of convolutional layers, pooling layers, fully connected layers, and activation functions like ReLU and sigmoid. [10] Training a machine learning model, such as a convolutional neural network (CNN), to recognize features associated with pathological myopia. Improve the accuracy and speed of diagnosis, reduce the reliance on manual expertise and subjectivity. AUC-ROC Curve: The Area Under the Receiver Operating Characteristic Curve, which measures the model's ability to distinguish between positive and negative classes. The paper likely concludes with a discussion of the model's potential benefits in clinical practice, such as increased diagnostic efficiency and reduced error rates.

In this paper the study aim is to test and analyze the results of different ML models that determine the characteristics and detect and classify the presence of ophthalmic disease (Glaucoma, high myopia, low myopia).[11] The pipeline is categorised in three phases data collection, data processing, data transformation, model building and performance evaluation. The models which were trained and tested to detect the presence of eye disease. Feature selection is the process of reducing the number of input variables when developing a predictive model its goal is to remove the uninformative or redundant predictors from the model. We compared ten classifier performance those are RF, SGD, GBC and KNN models .among those GBC model has the better accuracy performance with 0.92. RF model obtained the best performance result with high myopia ,SGDC model accuracy is 0.85, and precision is 0.76. KNN model obtained with a recall of 0.88.

In this paper However, missing data and measurement error (ME) are common challenges in risk prediction modelling that can introduce bias in myopia prediction. To predict the onset of myopia we use various statistical models, such as discrete-time survival analysis,[12] Cox proportional hazard models , ordinary logistic regression, multilevel ordinal logistic regression .Researchers have explored various imputation techniques, such as mean, median, mode imputation, multiple imputation (MI), and deep learning-based methods to handle missing data in risk prediction studies. In this paper, we assume that the ME of a continuous predictor, i.e., non-cycloplegic SE, is systematic. There are ml methods were explored in this paper Cart Decision Tree(tree), Naive Bayes(nb), Random Forest(rf), and Xgboost(xgb).

The AC parameters and TP parameters of the Ortho-K lens worn by the patients were used as target predictors. In this paper we[13] are aiming at the ML Method to predict the Alignment Curve (AC) and Target Power (TP) of orthokeratology in vision shaping treatment (VST) to assist in fitting. We use four machine learning algorithms (including Random Forest, Gaussian Process, Support Vector Machine, and Linear Regression) were used to predict AC and TP. The AC and TP optimize the hyper-parameters of the machine learning model to achieve the best predictive performance of the machine learning model. Gaussian Process and Random Forest performed best in predicting AC and TP, with R2 of 0.84 and 0.88, respectively.

Therefore, there is a critical need to develop an AI-based formula with significantly improved accuracy exclusively designed for highly myopic eyes. The eXtreme Gradient Boosting (XGBoost) and support vector[14] regression (SVR) are two popular and widely used machine learning models that provide applications in data classification and regression. This paper tells about to develop a novel machine learning-based intraocular lens (IOL) power calculation formula for highly myopic eyes. The study likely involved collecting a large dataset of pre-operative and post-operative eye measurements from patients with high myopia. The Zhu-Lu formula potentially offers improved accuracy in predicting the correct IOL power for highly myopic eyes compared to conventional methods. Machine learning models can adapt to individual patient characteristics, leading to better tailored and more effective treatments.

The paper explores how a machine learning algorithm can enhance the accuracy of fitting orthokeratology (ortho-K) lenses used in vision shaping treatments. Enhancing the precision of lens fitting to optimize treatment outcomes and reduce complications.

machine learning model aims to provide more precise fitting compared to traditional methods, which could lead to better visual outcomes[15] and comfort. Developing and training a machine learning algorithm (such as regression models, neural networks, or support vector machines) using the collected data to predict optimal lens parameters. Testing the model's predictions against actual fitting outcomes to evaluate its performance and accuracy. The machine learning algorithm significantly enhances the accuracy of ortho-K lens fitting, leading to better patient outcomes. The study shows that using advanced computer techniques can improve the fitting of ortho-K lenses.

3. METHODOLOGY

3.1 Data Collection:

The dataset used in this study was obtained from clinical records

of 7,081 patients who underwent corneal refractive surgery at the Peking Union Medical College Hospital between January 2000 and October 2017. The data include patients who had one of three types of surgeries: LASIK, PRK, or SMILE. Data from the left and right eyes were treated separately, resulting in 13,723 data points.

3.2 Data Preprocessing:

Data Cleaning: Samples with key data missing were removed to ensure data integrity.

Standardization and Normalization: Z-score normalization and Min-Max scaling were applied to adjust data magnitudes, improving the performance and convergence of the machine learning models.

Z-score normalization equation:

$$z = \frac{x - \mu}{\sigma}$$

Min-Max scaling equation:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Handling Imbalanced Data: The Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalance, particularly as LASIK had the most cases and PRK had the fewest.

3.3 Model Selection:

Six machine learning models were chosen based on their suitability for multi-class classification:

- Naive Bayesian Model (NBM)
- Random Forest (RF)
- AdaBoost
- XGBoost
- Back Propagation (BP) Neural Network
- Deep Belief Network (DBN)

3.4 Architecture Design:

The chosen models were trained using a 10-fold cross-validation approach with grid search to identify the optimal hyperparameters. Feature selection involved identifying the top 12 features contributing to the prediction, based on SHAP (SHapley Additive exPlanations) values.

3.5 Training Process:

During model training, SMOTE was applied in each cross-validation fold to balance the data. The training was conducted on 80% of the dataset, while 20% was set aside for testing. Models were evaluated for both left and right eye data, with different hyperparameters such as max_depth tuned to improve performance.

3.6 Evaluation Metrics:

Two primary metrics were used to evaluate model performance:

Accuracy: The proportion of correctly classified instances out of all instances

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}$$

Macro F1: A weighted average of precision and recall, useful for imbalanced datasets.

$$F1 = \frac{1}{C} \sum_{i=1}^C \frac{2 \times P_i \times R_i}{P_i + R_i}$$

where C is the total number of classes.

3.6 Performance Comparison:

Among the six models, the Random Forest (RF) model demonstrated the best performance, achieving:

- **Accuracy:** 87.75% for the left eye and 82.29% for the right eye.
- **Macro F1:** 0.8019 for the left eye and 0.8080 for the right eye.

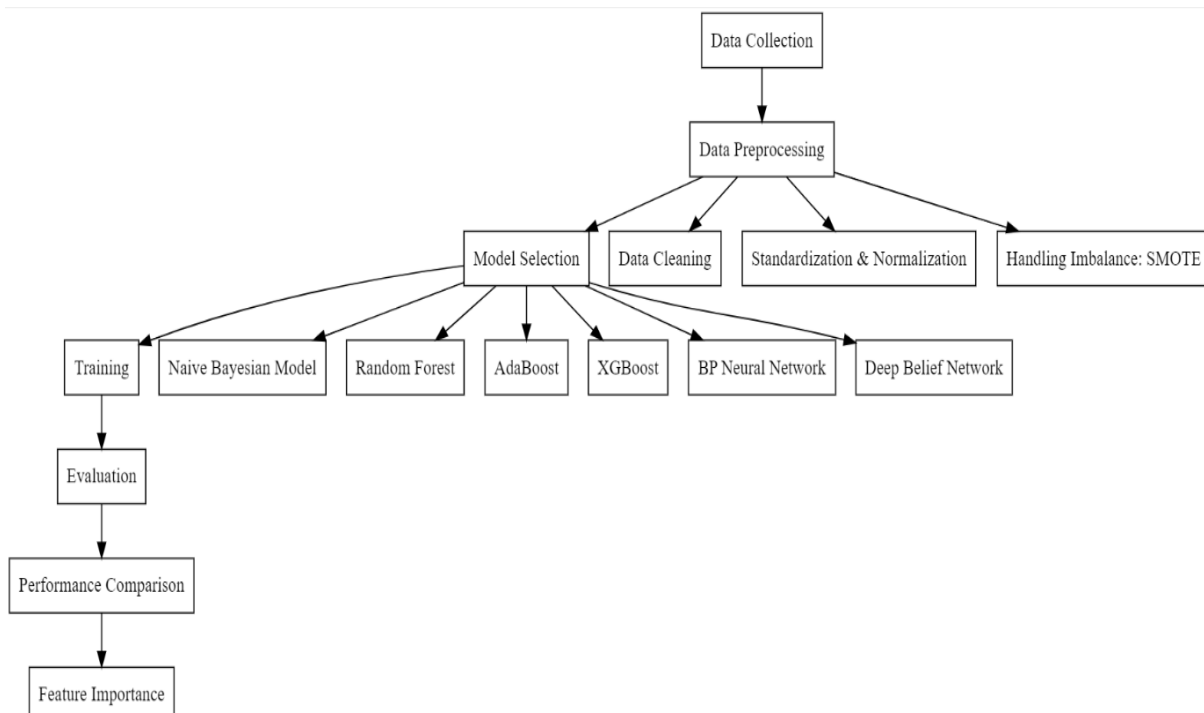


Figure 1

4. RESULTS

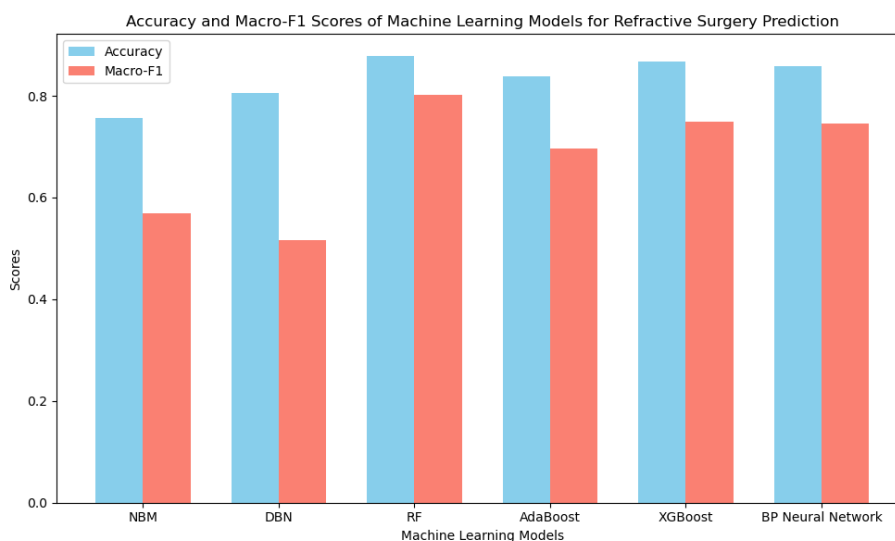


Figure-2

4.1 Comparison Table:

MODEL	Accuracy	Macro-F1
Random FOREST(RF)	0.8775	0.8019
XGBoost	0.8677	0.7484
Back Propagation(BP)	0.8578	0.7462
AdaBoost	0.8378	0.6970
Naïve Bayesian	0.7566	0.5996
Deep Belief Network(DBN)	0.8054	0.5165

5. CONCLUSION

The study finds that the Random Forest (RF) model is the best machine learning method for predicting the type of refractive surgery (LASIK, PRK, or SMILE) a patient should receive. Using the top 12 most important features and setting the model's depth to 10, the RF model achieved the highest accuracy (87.75%) and Macro-F1 score (80.19%), closely matching decisions made by ophthalmologists. This suggests RF could be a reliable tool to help doctors choose the right surgery for each patient. The study also highlights that key features like Central Corneal Thickness, Sphere-Column Conversion, and DS Spherical Power significantly influenced the predictions. These findings align with expert knowledge, making the model more transparent and useful in clinical settings. Future plans include testing this model with data from other hospitals to ensure it works well across different patient groups, overcoming the study's single-center data limitation.

6. REFERENCES

- Tong, H., Huang, Z., Li, Y., Chen, Y., Tian, B., Ding, L., & Zhu, L. (2023). Machine learning to analyze the factors influencing myopia in students of different school periods. *Frontiers in Public Health*, 11. <https://doi.org/10.3389/fpubh.2023.1169128>
- Li J, Zeng S, Li Z, Xu J, Sun Z, Zhao J, Li M, Zou Z, Guan T, Zeng J, Liu Z, Xiao W, Wei R, Miao H, Ziyar I, Huang J, Gao Y, Zeng Y, Zhou XT, Zhang K. Accurate prediction of myopic progression and high myopia by machine learning. *Precis Clin Med*. 2024 Mar 4;7(1):pbac005. doi: 10.1093/pcmedi/pbac005. PMID: 38558949; PMCID: PMC10981449. Analysis of Machine Learning in Myopic Progression Prediction
- Wei, L., Song, Y., He, W., Chen, X., Ma, B., Lu, Y., & Zhu, X. (2020). Accuracy Improvement of IOL power prediction for highly myopic eyes with an XGBOOST Machine Learning-Based calculator. *Frontiers in Medicine*, 7. <https://doi.org/10.3389/fmed.2020.592663>
- Ying, B., Chandra, R. S., Wang, J., Cui, H., & Oatts, J. T. (2024). Machine learning models for predicting cycloplegic refractive error and myopia status based on Non-Cycloplegic data in Chinese students. *Translational Vision Science & Technology*, 13(8), 16. <https://doi.org/10.1167/tvst.13.8.16>

5. Tu, Y., Hu, X., Zeng, C., Ye, M., Zhang, P., Jin, X., Zhang, J., & Zhou, L. (2022). A machine-learning approach to discerning prevalence and causes of myopia among elementary students in Hubei. *International Ophthalmology*, 42(9), 2889–2902. <https://doi.org/10.1007/s10792-022-02279-5>
6. Zhang, R., Dong, L., Yang, Q., Zhou, W., Wu, H., Li, Y., Li, H., & Wei, W. (2022). Screening for novel risk factors related to high myopia using machine learning. *BMC Ophthalmology*, 22(1). <https://doi.org/10.1186/s12886-022-02627-0>
7. Li, W., Tu, Y., Zhou, L., Ma, R., Li, Y., Hu, D., Zhang, C., & Lu, Y. (2024). Study of myopia progression and risk factors in Hubei children aged 7–10 years using machine learning: a longitudinal cohort. *BMC Ophthalmology*, 24(1). <https://doi.org/10.1186/s12886-024-03331-x>
8. Li, S., Ren, M., Gan, J., Zhang, S., Kang, M., Li, H., Atchison, D. A., Rozema, J., Grzybowski, A., & Wang, N. (2022). Machine learning to determine risk factors for myopia progression in primary school children: the Anyang Childhood Eye Study. *Ophthalmology and Therapy*, 11(2), 573–585. <https://doi.org/10.1007/s40123-021-00450-2>
9. Kim, Y. C., Chang, D. J., Park, S. J., Choi, I. Y., Gong, Y. S., Kim, H., Hwang, H. B., Jung, K. I., Park, H. L., Park, C. K., & Kang, K. D. (2021). Machine learning prediction of pathologic myopia using tomographic elevation of the posterior sclera. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-85699-0>
10. Rauf, N., Gilani, S. O., & Waris, A. (2021). Automatic detection of pathological myopia using machine learning. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-95205-1>
11. Leite, D., Campelos, M., Fernandes, A., Batista, P., Beirão, J., Menéres, P., & Cunha, A. (2022). Machine Learning automatic assessment for glaucoma and myopia based on Corvis ST data. *Procedia Computer Science*, 196, 454–460. <https://doi.org/10.1016/j.procs.2021.12.036>
12. Lai, H., Gao, K., Li, M., Li, T., Zhou, X., Zhou, X., Guo, H., & Fu, B. (2024). Handling missing data and measurement error for early-onset myopia risk prediction models. *BMC Medical Research Methodology*, 24(1). <https://doi.org/10.1186/s12874-024-02319-x>
13. Gong, J., Li, K., Hu, J., Chen, C., Chen, H., Cao, Q., & Wu, G. (2022). Application of machine learning method in clinical fitting of Ortho-K lens for myopia correction. *Research Square (Research Square)*. <https://doi.org/10.21203/rs.3.rs-1815078/v1>
14. Guo, D., He, W., Wei, L., Song, Y., Qi, J., Yao, Y., Chen, X., Huang, J., Lu, Y., & Zhu, X. (2023). The Zhu-Lu formula: a machine learning-based intraocular lens power calculation formula for highly myopic eyes. *Eye And Vision*, 10(1). <https://doi.org/10.1186/s40662-023-00342-5>
15. Fan, Y., Yu, Z., Tang, T., Liu, X., Xu, Q., Peng, Z., Li, Y., Wang, K., Qu, J., & Zhao, M. (2022). Machine learning algorithm improves accuracy of ortho-K lens fitting in vision shaping treatment. *Contact Lens and Anterior Eye*, 45(3), 101474. <https://doi.org/10.1016/j.clae.2021.101474>