tvst

Special Issue

Explainable Machine Learning Approach as a Tool to Understand Factors Used to Select the Refractive Surgery Technique on the Expert Level

Tae Keun Yoo[1], Ik Hee Ryu[2], Hannuy Choi[2], Jin Kuk Kim[2], In Sik Lee[2], Jung Sub Kim[2], Geunyoung Lee[3], and Tyler Hyungtaek Rim[4]

[1] Department of Ophthalmology, Aerospace Medical Center, Republic of Korea Air Force, Cheongju, South Korea
[2] B&VIIT Eye Center, Seoul, South Korea
[3] MediWhale, Seoul, South Korea
[4] Singapore Eye Research Institute, Singapore National Eye Centre, Duke-NUS Medical School, Singapore, Singapore

Correspondence: Tae Keun Yoo, Department of Ophthalmology, Aerospace Medical Center, Republic of Korea Air Force, 635 Danjae-ro, Sangdang-gu, Cheongju, South Korea. e-mail: eyetaekeunyoo@gmail.com
Ik Hee Ryu, B&VIIT Data Research Institute, B&VIIT Eye Center, GT Tower, 1317-23 Seocho-Dong, Seocho-Gu, Seoul, South Korea. e-mail: ikheeryu@gwbnviit.com

Received: July 25, 2019
Accepted: November 18, 2019
Published: February 12, 2020

Keywords: explainable machine learning; multiclass classification; refractive surgery; corneal laser surgery

Citation: Yoo TK, Ryu IH, Choi H, Kim JK, Lee IS, Kim JS, Lee G, Rim TH. Explainable machine learning approach as a tool to understand factors used to select the refractive surgery technique on the expert level. Trans Vis Sci Tech. 2020;9(2):8, https://doi.org/10.1167/tvst.9.2.8

Purpose: Recently, laser refractive surgery options, including laser epithelial keratomileusis, laser in situ keratomileusis, and small incision lenticule extraction, successfully improved patients' quality of life. Evidence-based recommendation for an optimal surgery technique is valuable in increasing patient satisfaction. We developed an interpretable multiclass machine learning model that selects the laser surgery option on the expert level.

Methods: A multiclass XGBoost model was constructed to classify patients into four categories including laser epithelial keratomileusis, laser in situ keratomileusis, small incision lenticule extraction, and contraindication groups. The analysis included 18,480 subjects who intended to undergo refractive surgery at the B&VIIT Eye center. Training (n = 10,561) and internal validation (n = 2640) were performed using subjects who visited between 2016 and 2017. The model was trained based on clinical decisions of highly experienced experts and ophthalmic measurements. External validation (n = 5279) was conducted using subjects who visited in 2018. The SHapley Additive ex-Planations technique was adopted to explain the output of the XGBoost model.

Results: The multiclass XGBoost model exhibited an accuracy of 81.0% and 78.9% when tested on the internal and external validation datasets, respectively. The SHapley Additive ex-Planations explanations for the results were consistent with prior knowledge from ophthalmologists. The explanation from one-versus-one and one-versus-rest XGBoost classifiers was effective for easily understanding users in the multicategorical classification problem.

Conclusions: This study suggests an expert-level multiclass machine learning model for selecting the refractive surgery for patients. It also provided a clinical understanding in a multiclass problem based on an explainable artificial intelligence technique.

Translational Relevance: Explainable machine learning exhibits a promising future for increasing the practical use of artificial intelligence in ophthalmic clinics.

Introduction

Refractive surgery techniques were developed during the past decade and successfully improved patients' quality of life. Laser refractive surgery procedures including laser epithelial keratomileusis (LASEK), laser in situ keratomileusis (LASIK), and small incision lenticule extraction (SMILE) produced excellent visual outcomes for patients with refractive error.[1] Currently, a selection of refractive surgery options are available in most eye clinics to treat refractive error by considering each patient's

ophthalmologic information. Each surgical option exhibits advantages and disadvantages, and thus a surgeon should recommend an optimal option after carefully reviewing patient data.[2]

Recently, machine learning, which is an area of artificial intelligence research, is increasingly popular in clinical medicine due to its ability to handle large data with high accuracy. It constructs statistical prediction models from datasets and estimates a new data instance. Support vector machines (SVMs), random forests (RFs), artificial neural networks (ANNs), and least absolute shrinkage and selection operator (LASSO) constitute widely used approaches in machine learning.[3,4] A previous study indicated that the machine learning technique can evaluate medical information to identify candidates for corneal refractive surgery.[5] However, previous machine learning models are considered as a black box and lack an explicit knowledge representation.[6] They are unable to provide reasoning and explanations on a decision in a manner similar to human experts.

Currently, the concept of explainable artificial intelligence is introduced in the field of medicine.[7] The explainable model allows users to focus on a rational decision and to verify if the model operates properly. The SHapley Additive ex-Planations (SHAP) is a promising solution to construct an explainable system.[8] This technique is used in several tasks in data mining research while selecting informative variables and predicting clinical values with higher interpretability. With advances in the visualization method for SHAP values, the technique is widely used to analyze data.[9] However, previous methods were limited in explaining the result of a single instance in a multicategorical problem because it is impossible for a single SHAP value to indicate 3 or more classes.[10]

To determine the optimal surgical technique based on medical evidence and patient's expectation for surgery and recovery, surgeons should consider several ocular measurements and patient factors such as dry eye, lifestyle, and budget. In the study, we constructed an expert-level decision support system to recommend the surgical option based on large clinical data and machine learning. An explainable machine learning method was adopted to demonstrate as to why the machine learning model should decide the surgery technique in each case. Specifically, we construct a multicategorical prediction model because there are multiple surgical options, including LASEK, LASIK, SMILE, and contraindication to corneal laser surgery. The machine learning model was constructed based on clinical decisions of highly experienced experts and was validated in a Korean population.

Methods

Study Population and Dataset

Study subjects included 18,480 healthy Korean patients who intended to undergo refractive surgery at the B&VIIT Eye Center from January 2016 to June 2018.[5] The retrospective study adhered to the tenets of the Declaration of Helsinki. The study protocol was approved by the Institutional Review Board of the Korean National Institute for Bioethics Policy (KoNIBP, 2018-2734-001). Written consent from subjects was waived because of the retrospective design of the present study. Protected personal health information was removed for the purpose of the study.

All patients underwent preoperative measurements of corrected distance visual acuity, manifest refraction, slit-lamp examination, and dilated fundus examination. A Pentacam scheimpflug device (Oculus GmbH, Wetzlar, Germany) was used to measure corneal topography. Pachymetry (NT-530P; Nidek Co., Ltd., Aichi, Japan) was used to evaluate the central corneal thickness. Pupil size was measured via Keratograph 4 (Oculus GmbH, Wetzlar, Germany) and noninvasive tear breakup time (NIBUT) was determined via Keratograph 5M (Oculus GmbH, Wetzlar, Germany). Each patient was interviewed and asked to complete a questionnaire survey on his or her occupation, anticipated surgery option, anticipated recovery period after surgery, budget concerns, and medical history (Table S1). The patients determined the anticipated surgery options after consulting an expert advisor. Eighty features from the corneal tomography on both eyes were automatically extracted from the 4 Maps Refractive Display via a custom-built optical character recognition (OCR) algorithm that simply converted digits in a Pentacam image into text data. We manually reviewed the digitized Pentacam data within the top and bottom 1% of all collected values, and all digitalized values were transformed properly via OCR. A total of 142 variables including the demographics data, ophthalmic measurements, and interview questionnaires are listed in Table S2. The same surgery technique was conducted in both eyes simultaneously. Moreover, keratoconus is the most important status for refractive surgery, present bilateral, but asymmetrically progressive thinning of the cornea.[11] Therefore, measurements of both eyes should be included in the analysis.

All patients were categorized into 4 groups based on the type of surgery conducted—LASEK, LASIK, SMILE, and contraindication to corneal laser surgery. A reference standard categorization was assigned

based on the clinical decision obtained from a full evaluation by 9 experts. Before surgery, the expert made a decision on the surgery option based on a patient's condition. Essentially, a surgeon was involved in a surgery option decision for each patient. All experts were board-certified ophthalmologists with an average experience of 10 years in refractive surgery. General criteria for consideration in surgery (which may vary in terms of several items from criteria used in other refractive practices) included the following parameters: age 18 years or older; myopia spherical equivalent >-10.0 diopters (D); hyperopia spherical equivalent $<+4.50$ D; central corneal thickness, measured with pachymetry, $>500\,\mu m$ for LASIK and $>480\,\mu m$ for LASEK and SMILE; residual corneal thickness >380 mm after surgery, NIBUT >5 s for LASIK; and absence of corneal abnormalities suggestive of keratoconus or other corneal ectatic diseases. Photorefractive keratectomy (PRK) procedures have evolved, and LASEK may combine several advantages presented by both PRK and LASIK.[12] Therefore, PRK has currently been replaced by LASEK in the B&VIIT Eye Center, and PRK was not considered as a surgical option. In this center, corneal refractive surgery has been considered as the primary correction method for refractive error. If a patient is not a candidate for corneal ablation as determined in the preoperative examination, phakic intraocular lens (ICL) is considered as an alternative treatment. The aforementioned are not absolute criteria, and expert ophthalmologists can recommend corneal refractive surgery based on their own clinical experience. An ophthalmologic examination was performed postoperatively at 1 and 6 months to screen for postoperative ectasia.

The study protocol is defined in our previous study.[5] Recent studies have shown that there has been a lack of external validation for novel prediction models.[13] Therefore, both internal and external validation procedures have been recommended and splitting the data by calendar time was determined as a good option for the development of prediction models.[14] We built a machine learning model and it was validated prospectively according to the design of a previous study.[3,15] Training and internal validation were performed using subjects who visited between 2016 and 2017. These subjects were considered as retrospective cohorts for model construction and calibration. The dataset was randomly separated into training (80%, n = 10,561) and validation sets (20%, n = 2640). In the training dataset, we designed a 10-fold cross-validation, which currently corresponds to the preferred technique in data mining to assess performance and to optimize the prediction models. To obtain the optimal result, we adopted a grid search (Cartesian method) in which a range of parameter values were tested via the 10-fold cross-validation strategy. This method trains machine learning models with each combination of possible hyperparameter values, and selects a hyperparameter to maximize accuracy.[16] The model was trained based on clinical decisions of highly experienced experts and large multi-instrument measurements. External validation (n = 5279) was conducted using subjects who visited in 2018. These subjects were considered as independent prospective cohorts to validate the machine learning model prospectively.

Machine Learning Technique

An architecture of our proposed machine learning model is shown in Figure 1. The study focuses on XGBoost (a recently developed meta-algorithm) due to its reliable and superior performance compared with other classic machine learning methods.[17] Additionally, XGBoost is derived from the extreme gradient boosting, which falls under larger parallel tree boosting. The technique optimizes both the training loss and regularization of the model for the ensemble of the trees generated. In the study, we used XGBoost to predict the surgery option class y_i for the given input feature vectors $X_i = \{x_1, x_2, \cdots, x_N\}$ including preoperative measurements and questionnaires. The training procedure was conducted via an additive strategy. Given a residue i with X_i, a tree ensemble model uses K additive functions to predict the output value $\hat{y}_i$ as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in F$$

where $f_k(X_i)$ denotes an independent tree structure with leaf scores of X_i and F denotes the space of functions containing all regression trees. The XGBoost algorithm introduces the regular function to control overfitting. The target function of XGBoost is expressed as follows:

$$\min\{Obj\} = \min\left\{\sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t)\right\}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2$$

where $l(y_i, \hat{y}_i)$ denotes a function of the difference between the ground truth value and predicted value of the ith observation in the iteration. The $\Omega(f_t)$ term penalizes the complexity of the model where T denotes the number of leaves and ω_j denotes the leaf node output in each subdecision tree model. The variables

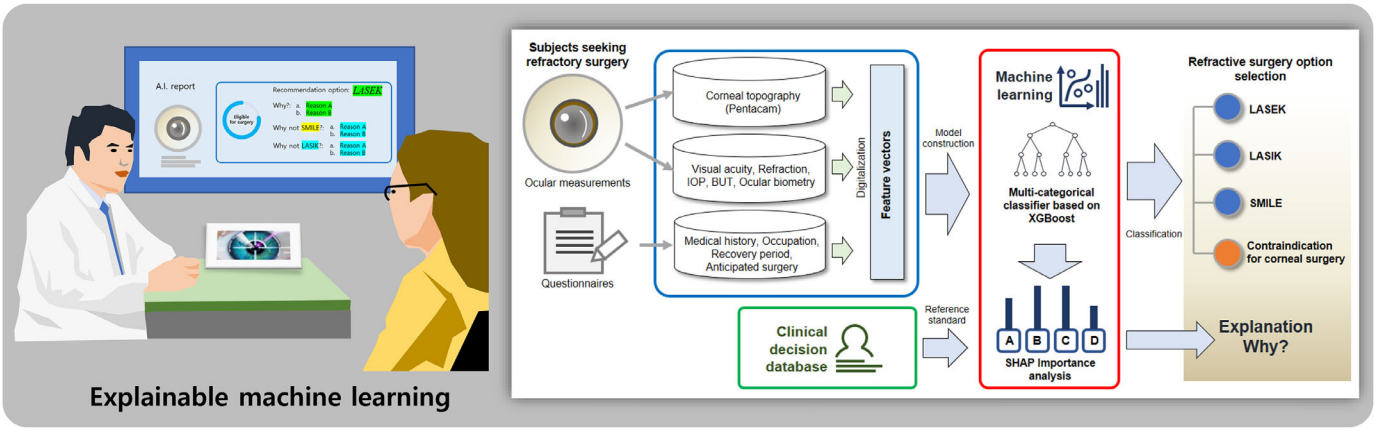

Figure 1. Architecture of the proposed machine learning models for corneal laser refractive surgery recommendation.

γ and λ are constants that control the degree of regularization. In the iterative learning process of XGBoost, the loss function is expanded into the Taylor second order series to quickly optimize the objective function while $L1$ and $L2$ regularizations are introduced to solve in a manner similar to penalty functions of LASSO and ridge regression.[18] Hence, the XGBoost model offers a more accurate prediction model and efficiently prevents overfitting.[19]

Although a multiclass XGBoost model analyzes the multiple surgery option dataset simultaneously, we also adopted additional two strategies including one-versus-rest (OVR) and one-versus-one (OVO) for explainable classification.[20] The OVR method corresponds to the simplest multiclass classification as shown in Figure S1. The final decision function selects the class that corresponds to the maximum value of binary decision function among all OVR binary classifiers. The OVO method constructs binary classifiers for all pairs of classes. The decision function selects the class that exhibits the largest number of votes by all binary classifiers as shown in Figure S2. In our study using the 4-class dataset, the OVR XGBoost classifier requires four binary classifiers and the OVO XGBoost classifier uses 6 pairwise binary classifiers for voting.

Other representative machine learning methods were also used for comparison purposes. The SVM is based on mapping data to a higher dimensional space via a kernel function and selects the maximum-margin hyperplane that separates training data.[21] The hyperplane is located based on a set of boundary training instances known as support vectors. Thus, the goal of the SVM is to improve accuracy by the optimization of space separation and it is well fitted for binary classification problems. The SVM was the most important development before the introduction of deep learning and tree-based techniques. The optimization process is formulated in a way that allows for nonseparable data by penalizing misclassifications. The OVR, OVO, and directed acyclic graph (DAG) SVM were adopted for multicategorical classification.[20] When the SVM models were built, feature selection using information gain technique was also performed to increase accuracy.[22] RF is an ensemble learning method for classification and consists of a collection of decision trees.[23] Specifically, RF can withstand high dimensional data in training faster than other methods with extremely robust performance in a multicategorical classification problem. An ANN model based on multilayer perceptron uses mathematical systems that mimic biological neural networks. We used a multilayer perceptron neural network with backpropagation for nonlinear pattern classification.

SHAP Technique

SHAP is a recently developed technique that aims to interpret black box machine learning models.[8] It provides a post-hoc interpreting method that is more aligned with human intuition.[24] Most previous machine learning algorithms provide predictors with global feature importance, and it is difficult to interpret each prediction case. However, the SHAP technique calculates the contribution of each input variable in each decision of a machine learning model. The SHAP value corresponds to the measure of additive feature attributions. The calculation formula for the SHAP value $\emptyset_i$ is defined as follows:

$$\emptyset_i = \sum_{S \in I_f \setminus \{i\}} \frac{|S|! \times (M - |S| - 1)!}{M!} \times \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_s \right) \right]$$

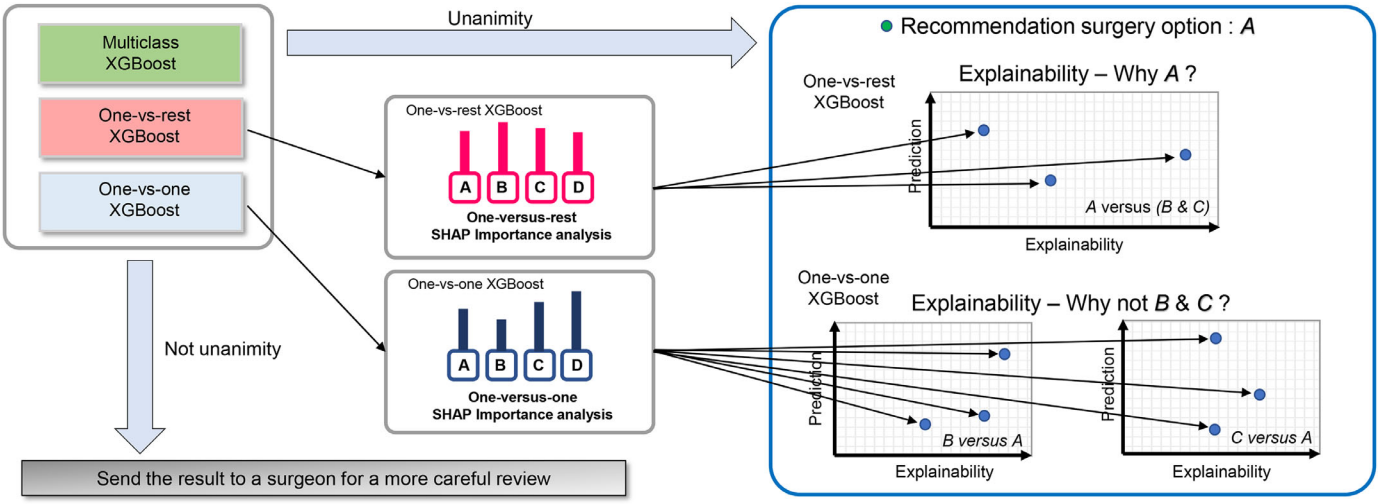

Figure 2. Schematic diagram of our proposed interpretable model for corneal laser refractive surgery recommendation.

where I_f denotes the set of input features, S denotes a subset of input features, and M denotes the number of input features. The term $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ corresponds to the machine learning model output when the ith feature is present. The term $f_S(x_s)$ denotes the output when the ith feature is withheld. A novel SHAP package developed by Lundberg provided a sampling-based and decision tree–based estimation of SHAP value.[25] The package computes the SHAP values of each prediction case with local interpretability, and this is shown via a force plot. A force plot is a new method for visualizing individual model predictions using the SHAP technique. In this plot, red arrows show the feature effects that drive the XGBoost results higher while blue arrows present the feature effects that drive these outcomes lower. Global interpretability was calculated by aggregating the SHAP values across the instances. The SHAP value can provide almost a full explanation that is crucial for a machine learning decision considering the effects of all variables. It was difficult for the current form of SHAP to show an intuitive diagram for multiclass classification, and thus we adopted the OVR and OVO strategies. We extracted information necessary to explain the results from listing the OVR and OVO XGBoost-based classification (Fig. 2). In our study, we noted only two most influential factors for a better view of the result. The detailed method to compare the primary factors between the explainable XGBoost model and clinician's decision is presented in Figure S3. We sampled subsets from all classes for manual chart review due to the large number of patients in this study.

Training Process

Imbalanced data is a challenging problem in multiclass machine learning because it significantly decreases the classification performance.[26] We adopted the synthetic minority oversampling technique (SMOTE) to overcome our data imbalance. The SMOTE method is an oversampling method that randomly generates new instances of minority class to balance the number of classes.[27] The method is the most popular and effective method to balance the dataset during a training process. In the study, the SMOTE technique was applied to generate a completely balanced dataset including the same number of instances based on the major group. When a binary variable (gender or yes/no questionnaires) was generated by SMOTE, a round off function was applied after the SMOTE process to recover the binary variable property. In the SMOTE process, the orthogonality of the questionnaire features for anticipated recovery and dry eye symptoms was preserved by selecting the feature with the highest value.

We faced a multiclass classification problem, and thus the measurement of multicategorical classifier performance was based on accuracy, relative classifier information (RCI), and Cohen's κ metric.[28] Accuracy is a standard metric to evaluate a classifier. Specifically, RCI denotes the performance with unbalanced classes capable of distinguishing among different misclassification distributions. Kappa is a standard meter for a multicategorical problem that is generally applied in several fields.[29] The detailed calculation methods are introduced in Table S3.

The parameter optimization under 10-fold cross-validation was operated to maximize accuracy under the SMOTE process. In the 10-fold cross-validation, the training dataset was randomly partitioned into 10 equal sized subgroups. One subgroup was retained as the validation set for testing the machine learning model, and the remaining 9 subgroups were used as training data. This cross-validation process was then repeated 10 times as each of the subgroups were used once for validation. The optimal model of multiclass XGBoost corroborated that the eta corresponded to 0.3, maximum tree depth for base learners corresponded to 5, and γ corresponded to 0.01, while maintaining default values of the other parameters. The OVR and OVO XGBoost models shared the parameters of the multiclass model. In the DAG SVM model, the optimal model was obtained via a Gaussian kernel function with a penalty parameter C corresponding to 10.0 and a scaling factor γ corresponding to 0.1. The OVR and OVO SVM models shared the parameters of the DAG model for the purpose of convenience. In RF, the optimal number of trees corresponded to 1000, and the number of predictors for each node corresponded to 3. The optimal multilayer perceptrons for the ANN model were set with two hidden layers (5 and 2 nodes). The neurons that exhibited rectified linear unit (ReLU) activation were used, and other training parameters were set to the default values of scikit-learn. ReLUs have been reported to be easier to optimize and are more easily generalized than Sigmoid or Tanh functions.[30]

The scikit-learn Python library and R version 3.5.1 (The Comprehensive R Archive Network; http://cran.r-project.org) were adopted to perform XGBoost and SHAP algorithms. We used the SHAP and XGBoost packages that are available on GitHub repository (https://github.com/slundberg/shap and https://github.com/pablo14/shap-values).

Data Availability

Data are not easily redistributable to researchers other than those engaged in the Institutional Review Board–approved research collaborations with the B&VIIT Eye Center, South Korea. The datasets utilized during the study are not publicly available due to reasonable privacy and security concerns.

Results

The characteristics of the subjects are listed in Table 1. Among a total of 18,480 subjects, 4893 subjects underwent LASEK, 6123 underwent LASIK, and 5834 underwent SMILE surgery as recommended by the surgeon. After a comprehensive examination, the remaining 1630 subjects were considered to exhibit a contraindication to corneal refractive surgery. The comparison between 4 groups shows significant differences in all variables as listed in Table 1. Supplementary Table S4 lists the characteristics of the subjects in terms of the training, internal validation, and external validation datasets. During the study, post-LASIK ectasia was developed in one patient among the development dataset including follow-up data. All patients were followed with for 6 months and there were no perioperative complications.

Table 2 shows the performance of final multiclass classifiers via 10-fold cross-validation in the training dataset with the SMOTE process. We obtained the accuracy of the multiclass, OVR, and OVO XGBoost models corresponding to 82.1%, 81.7%, and 81.9%, respectively. The average accuracy of random forest, OVR SVM, OVO SVM, DAG SVM, and ANN models corresponded to 81.5%, 75.3%, 75.7%, 75.5%, and 76.0% respectively. The multiclass XGBoost model performed statistically better than the SVM and ANN models ($P < 0.001$). Similarly, consistent results were obtained by using other metrics including average RCI, and Cohen's κ. Without the SMOTE process, a multiclass XGBoost accuracy of 80.2% was obtained via 10-fold cross-validation. When the variables of the anticipated option were excluded, the average accuracy of the multiclass XGBoost was 70.2%. Additionally, SHAP importance results from the 10-fold cross-validation procedure are shown in Figure 3. The results indicated that the multiclass XGBoost did not explain each decision for a specific surgery option in a multicategorical setting. Each OVR model showed the global interpretable form as to why the surgery option is selected. The SHAP importance from 4 OVR XGBoost classifiers revealed that different factors affected each classifier. Specifically, the anticipated option of patients exhibited the greatest effect on the decision in each OVR model for LASEK, LASIK, and SMILE. The SHAP clustering force plots are presented in Figure S4, and they also demonstrated that the anticipated option of patients was a major factor in surgery selection.

As shown in Figure 4, the XGBoost-based methods are successfully performed in the internal and external validation dataset. The multiclass XGBoost indicated that the performance in the training set and exhibited an accuracy of 81.0% in the internal validation set and 78.9% in the external validation set. Performances varied for the multiclass, OVR, and OVO XGBoost models in the internal and external validation

Table 1. Comparison Between LASEK, LASIK, SMILE, and Contraindication Cases

	LASEK	LASIK	SMILE	Contraindication	*P* Value[a]
Number	4893	6123	5834	1630	
Age (years)	26.9 ± 5.6	27.3 ± 6.1	27.3 ± 6.0	33.8 ± 8.0	<0.001
Sex, female (%)	2723 (55.7)	3202 (52.3)	3069 (52.6)	868 (53.3)	<0.001
Spherical equivalent (Diopter)	−5.38 ± 2.23	−3.94 ± 1.84	−4.39 ± 1.69	−7.82 ± 5.02	<0.001
CDVA (logMAR)	−0.012 ± 0.039	−0.012 ± 0.038	−0.011 ± 0.038	0.017 ± 0.121	<0.001
IOP (mm Hg)	15.1 ± 3.5	15.5 ± 3.4	15.4 ± 3.0	15.2 ± 2.4	<0.001
Pupil diameter (mm)	2.93 ± 0.62	2.88 ± 0.55	2.88 ± 0.57	2.86 ± 0.52	<0.001
Central corneal thickness (μm)	530.5 ± 33.0	549.3 ± 27.4	545.9 ± 33.1	503.8 ± 42.9	<0.001
NIBUT (seconds)	6.67 ± 6.40	7.04 ± 6.70	7.06 ± 6.67	5.20 ± 3.61	<0.001
Anticipated surgery option					
LASEK (%)	2402 (49.1)	1586 (25.9)	1488 (25.5)	448 (27.5)	<0.001
LASIK (%)	983 (20.1)	2951 (48.2)	915 (15.7)	463 (28.4)	<0.001
SMILE (%)	1052 (21.5)	1193 (19.5)	3296 (56.5)	523 (32.1)	<0.001
ICL or none	456 (9.3)	392 (6.4)	134 (2.3)	196 (12.0)	<0.001
Occupation					
Sports (%)	680 (13.9)	380 (6.2)	583 (10.0)	155 (9.5)	<0.001
Driver (%)	298 (6.1)	416 (6.8)	543 (9.3)	1277 (7.8)	<0.001
Computer or smartphone (%)	2906 (59.4)	3472 (56.7)	3506 (60.1)	950 (58.3)	<0.001
Anticipated recovery time					
One day (%)	274 (5.6)	1702 (27.8)	1762 (30.2)	414 (25.4)	<0.001
Three days (%)	2251 (46.0)	3594 (58.7)	2818 (48.3)	843 (51.7)	<0.001
One week (%)	1649 (33.7)	1163 (19.0)	1202 (20.6)	328 (20.1)	<0.001
Concern about budget (%)	3303 (67.5)	4298 (70.2)	4230 (72.5)	1154 (70.8)	<0.001

CDVA, corrected distance visual acuity; IOP, intraocular pressure; NIBUT, noninvasive breakup time.
[a]Comparison using the 1-way ANOVA test and χ^2 test.

Table 2. Classification Performance of Machine Learning Models to Predict Laser Corneal Refractive Surgery Option Via 10-Fold Cross-Validation

	Accuracy (%) (95% CI)	RCI (95% CI)	κ (95% CI)	*P* Value[a]
Trained with SMOTE				
Multiclass XGBoost	82.1 (81.1–83.0)	0.537 (0.525–0.549)	0.758 (0.747–0.769)	Reference
One-versus-rest XGBoost	81.7 (80.7–82.6)	0.531 (0.519–0.543)	0.753 (0.742–0.764)	0.578
One-versus-one XGBoost	81.9 (80.9–82.8)	0.534 (0.522–0.546)	0.756 (0.745–0.767)	0.780
Random forest	81.5 (80.5–82.4)	0.527 (0.515–0.539)	0.750 (0.739–0.761)	0.407
One-versus-rest SVM	75.3 (74.2–76.3)	0.422 (0.410–0.434)	0.668 (0.656–0.680)	<0.001
One-versus-one SVM	75.7 (74.7–76.7)	0.428 (0.415–0.441)	0.674 (0.662–0.686)	<0.001
DAG SVM	75.5 (74.5–76.5)	0.425 (0.412–0.438)	0.671 (0.659–0.683)	<0.001
Artificial neural network	76.0 (74.9–77.0)	0.432 (0.419–0.445)	0.677 (0.665–0.689)	<0.001
Trained without SMOTE				
Multiclass XGBoost	80.2 (79.2–81.2)	0.514 (0.502–0.526)	0.730 (0.719–0.741)	0.011
One-versus-rest XGBoost	80.1 (79.1–81.1)	0.513 (0.501–0.525)	0.727 (0.715–0.738)	0.015
One-versus-one XGBoost	78.5 (77.4–79.5)	0.505 (0.493–0.517)	0.721 (0.709–0.732)	0.001
Without anticipated surgery option				
Multiclass XGBoost	70.3 (69.2–71.4)	0.407 (0.394–0.420)	0.593 (0.581–0.605)	<0.001
One-versus-rest XGBoost	68.8 (67.7–69.9)	0.385 (0.372–0.398)	0.571 (0.559–0.583)	<0.001
One-versus-one XGBoost	68.3 (67.2–69.4)	0.380 (0.366–0.393)	0.565 (0.552–0.568)	<0.001

CI, confidence interval; RCI, relative classifier information; SVM, support vector machine.
[a]Comparison of accuracy with the best machine learning technique (multiclass XGBoost with SMOTE).

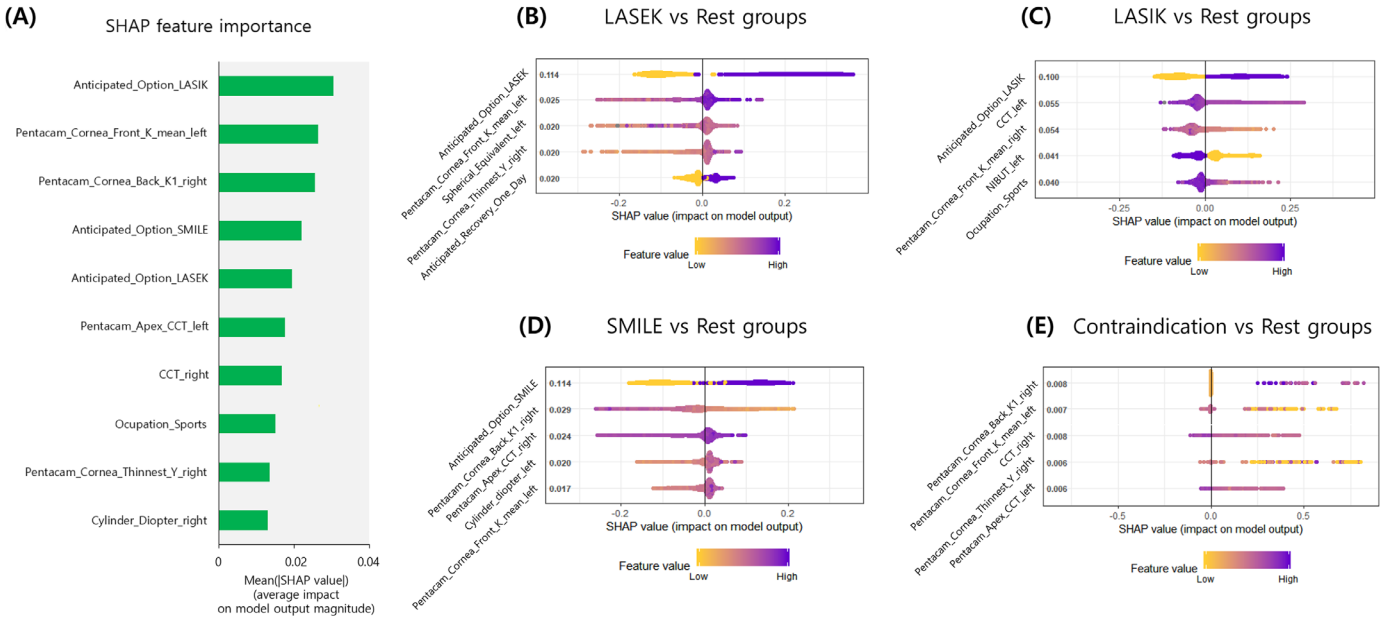

Figure 3. Global feature importance estimates selected by the XGBoost-based SHAP technique. (A) Multiclass (4 classes) classification problem. (B) Binary classification with LASEK versus rest groups. (C) Binary classification with LASIK versus rest groups. (D) Binary classification with SMILE versus rest groups. (E) Binary classification with Contraindication versus rest groups.

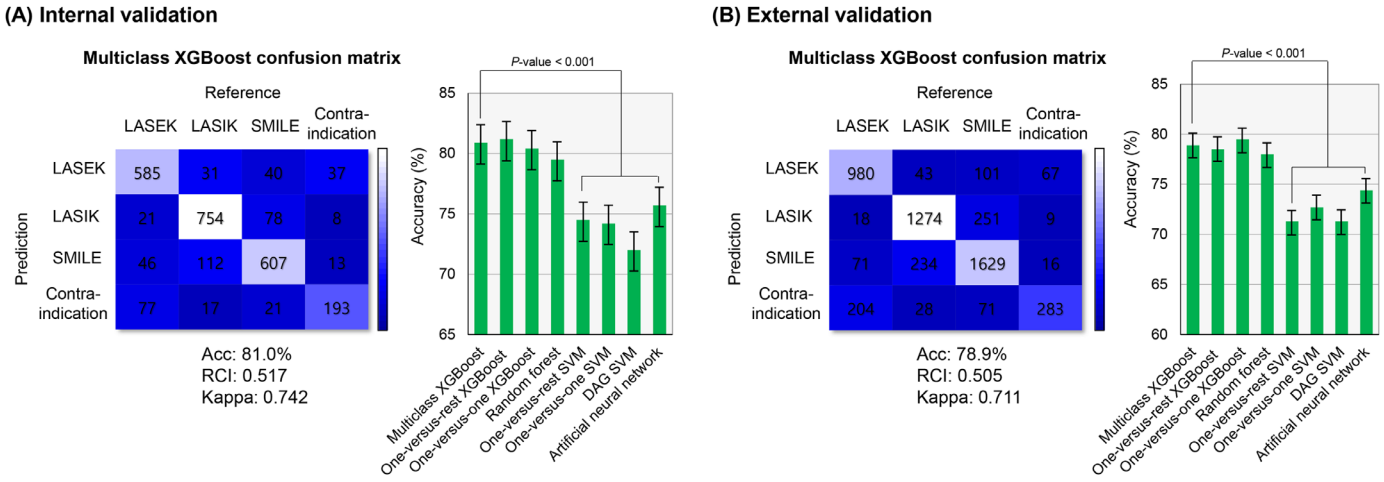

Figure 4. Confusion matrix of multiclass XGBoost and performance for multiclass XGBoost and other algorithms. (A) Performance in the internal validation. (B) Performance in the external validation. The error bars demonstrate the 95% confidence intervals.

sets. The multiclass XGBoost models exhibited better accuracy than the SVM and ANN models ($P < 0.001$). The RCI and Cohen's κ metrics indicated a performance similar to the accuracies of machine learning models. Figure 5 presents a comparison of the primary factors between the explainable XGBoost model and the clinician's decision among the external validation dataset. Most electronic charts did not present the factors used for the clinical decision. In the surgical cases, the major factor for making a decision was the patient's anticipated option. The agreement rate of the primary decisional factor was 92.7% in the subjects with identified decisional factors in their medical records.

A classification result from randomly selected SMILE case is demonstrated in Figure 6. The force plots include the prediction explanation bar that exhibits pink blocks that push the prediction value higher and blue blocks that push the value lower. The blocks labelled by variables were sorted based on their impact on the result. The results of the OVR XGBoost model explained that SMILE was the choice because the patient's keratometric power of corneal back surface was inside a safe range and the patient

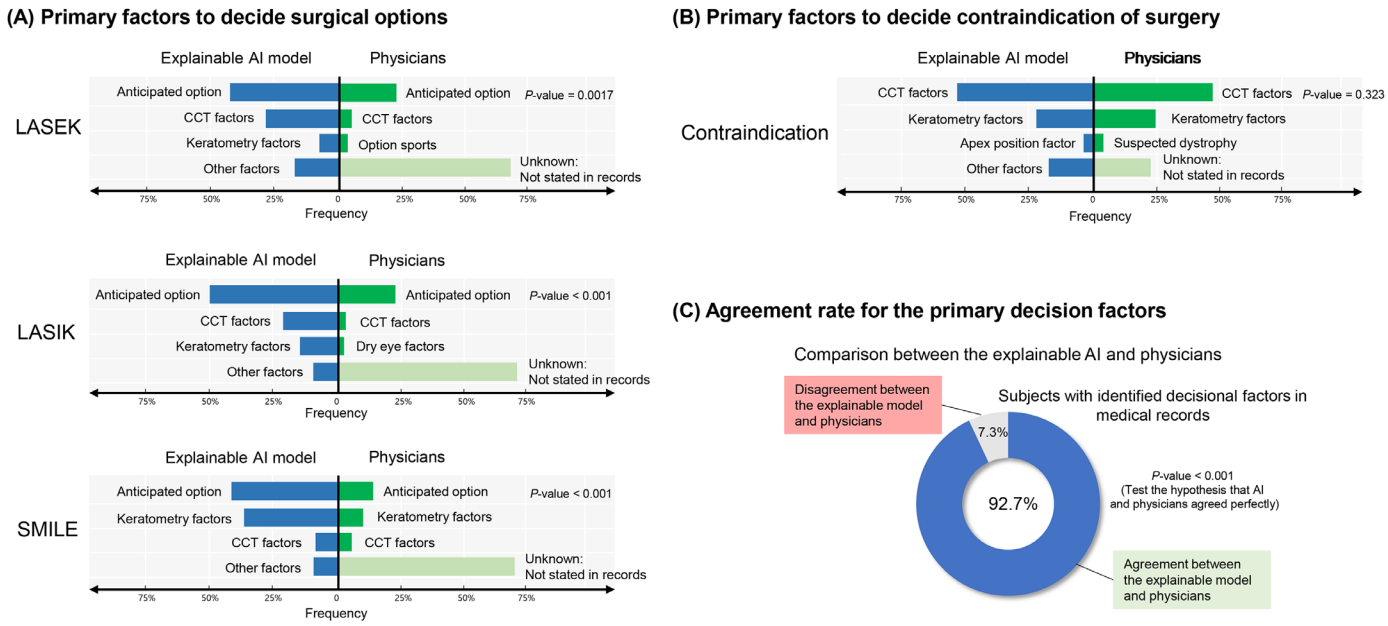

Figure 5. Comparison of the primary factors between the explainable XGBoost model and clinician's decision. The surgical decision was based on a review of electronic health records. One hundred samples for each group were extracted randomly from the external validation dataset for comparison.

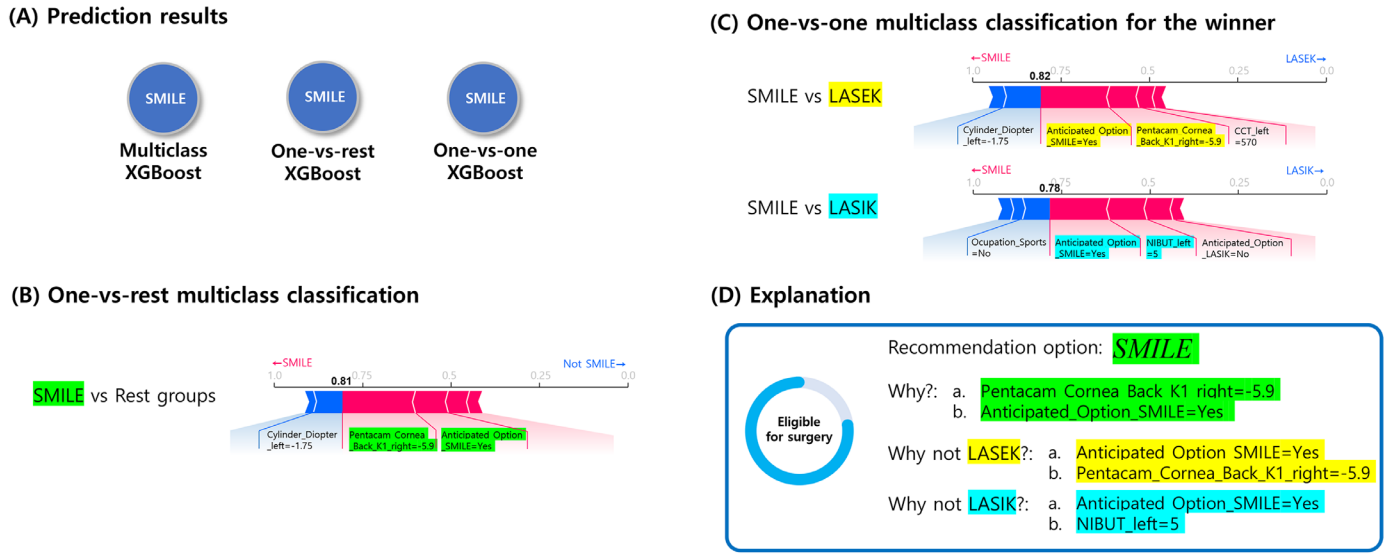

Figure 6. SMILE case example showing the machine learning prediction result with local interpretation via force plots.

anticipated SMILE surgery. The OVO model explained as to why SMILE corresponded to a better option than LASEK and LASIK. In this case, the patient's anticipation for SMILE was the most influential factor in the decision. It also revealed that the patient was a candidate for surgery due to the safe keratometric power of corneal back surface and normal corneal thickness. The SHAP explanations for the results were consistent with prior knowledge from ophthalmologists. Figure 7 shows an example contraindication case for corneal refractive surgery. The results indicated that thin corneal thickness and high myopia of the patient contributed to the machine learning decision. The informed machine learning prediction can enable surgeons in more effectively and objectively reviewing patients' data to determine the surgery option on the expert level.

We analyzed the training time to build an explainable XGBoost model using the whole training dataset. The training process was completed within 45 minutes to obtain the final multiclass model. During the validation process, the execution time required for one case

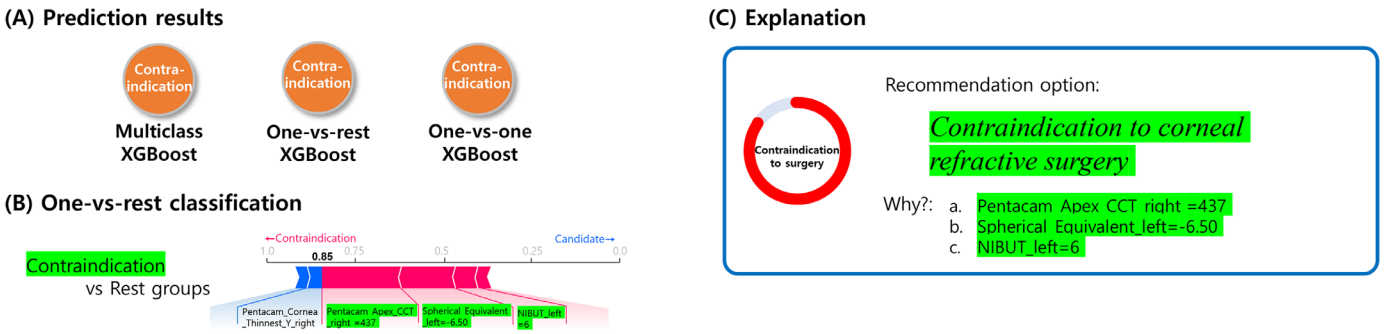

Figure 7. Contraindication case example showing the machine learning prediction result with local interpretation via force plots.

was 30 ms using our platform without a graphic processing unit.

Discussion

The study presents a multicategorical explainable machine learning framework for laser refractive surgery selection. In the study, the explainable XGBoost technique was successfully extended to a multiclass classification problem. An appropriate laser surgery strategy for refractive correction is an important issue to decrease postoperative complications. Prior to surgery, an expert decides on the surgery option based on a patient's condition. The proposed method involves comprehensively combining ophthalmic data and patients' interview data to select the surgery option on the expert level based on the clinical decision database. We built the machine learning model based on the retrospective cohorts (training and internal validation), and it was prospectively validated (external validation). The final model predicted the surgery option with an accuracy of 78.9% in external validation, and this corresponded to the multicategorical classification problem of the 4 classes. The framework adopts intuitive interpretability and is expected to assist in making user friendly and less risky clinical decision. To the best of the authors' knowledge, this is the first study to select the corneal refractive surgery option on the expert level using artificial intelligence.

As indicated by a previous study, machine learning has not yet been perfected to predict keratoconus.[31] No complications were noted in the patients during the postoperative period except for the diagnosis of post-LASIK ectasia in one patient. Because there is no definitive diagnosis method, the decision of the expert was used for reference. Although the multiclass machine learning model was trained to imitate the expert clinicians, the preoperative selection of forme furste keratoconus is key for selecting the ideal surgical option. Recently, several techniques have been developed to deal with a large amount of clinical information to screen for keratoconus.[32] Our previous study demonstrated that artificial intelligence can integrate medical information to identify candidates for corneal refractive surgery.[5] The XGBoost model in this study extended the previous binary model to a multiclass machine learning model for more comprehensive support of the final decision. In the future, measurements of the biomechanical properties of the corneal tissue will be combined with the machine learning models to improve the keratoconus screening and surgery selection.

We contemplated the interpretability of the multiclass prediction result to understand machine learning users. The explanation involved listing the force plots of the OVR and OVO XGBoost models and was effective in terms of easily understanding the users in the study. The previous SHAP technique is suitable for regression and binary classification problems.[8] Zhang et al.[10] indicated that it is more difficult to understand interpretability in a multiclass setting. His research team developed additive postprocessing for interpretability technique, and this transformed a machine learning model to a simplified form for global interpretability. The approach involved a complex computational procedure and did not exhibit local interpretability for each instance. We offered a simple solution to overcome the problem by using the OVR and OVO classification. The computational load increases exponentially when the number of classes increases in an OVR classification strategy and this is disadvantageous. However, the proposed approach exhibited an intuitive and task-based interface that a patient can understand. The explainable model answers why the surgery is selected and why other options are not recommended, and it does not break the interpretability of the SHAP technique.

The agreement rate between the explainable XGBoost model and the clinician's decision was 92.7% although the anticipated option of the patients was the most influential factor in most surgery cases. The results also revealed that the anticipation of a specific option for the patients affected the decisions, and this is potentially because most patients could consider all surgery options. Recent technical advances developed various surgical options available to patients who seek refractive surgery. In order to satisfy patients, it is vital for surgeons to consider the pros and cons of each refractive surgery option.[1] The proposed model provided an surgical option on the expert level based on a large clinical decision database. Generally, LASEK is a good choice for patients with thin cornea and high risk to expose to ocular trauma. However, it involves several disadvantages including more painful postoperative periods, longer visual recovery, and risk of corneal opacity.[2] Specifically, LASIK exhibits advantages of fast and painless recovery although it can cause dry eye syndrome due to corneal nerve damage.[33] Additionally, SMILE is a flapless and minimally invasive procedure that shows the benefits of fast recovery of vision and reduced symptom of dry eyes.[34] However, SMILE is a relatively new technique, and thus it is generally more expensive than other methods in South Korea. Our SHAP importance result indicates that the explainable machine learning model reflects the advantages and disadvantages of each option.

In our study, the interpretability of our explainable machine learning model suffered from the lack of a validation scheme and the presence of multicollinearity. A method for achieving an objective validation of the interpretability has not yet been found. The interpretability is able to elucidate the logic of the decisions made by the model, but it may sometimes provide a local neighborhood of the input and fail to obtain insight into the underlying mechanism.[35] The explanations provided by the interpretability may occasionally be unreliable and misleading because the explainable machine learning technique generally uses important factors that allow for safe applications.[36] Our results might be considered as a failure in obtaining insight into the underlying mechanism because the machine learning model demonstrated that patients' anticipated option was frequently the most influential factor in determining surgery. However, we believe that it is another insight explaining the machine learning model. The explainable system has a potential to discover deep patterns for personalized medicine, which are not accessible to clinicians, and it could provide a tool to understand factors used to select the surgery technique for patients.[37] Additionally, the multicollinearity of the variables affected the interpretability of our model. Because both eyes were included in the analysis and the variables were highly correlated to each other (Fig. S5), the XGBoost models selected important factors from both eyes with an inconsistent logic. Future studies attempting to achieve explainable machine learning should be carried out to solve these issues.

Moreover, there were significant differences of the decisional factors between surgeons (Fig. S6), and this shows the interclinician variability in selecting surgical options. Clinical data accumulation and machine learning techniques could potentially eliminate the possibility of interclinician variability with evidence-based decision making. In this study, only 9 clinicians were represented from the same location, and they attempted to follow the same criteria when considering surgical options. As there are significant variations in defining surgical criteria between the modalities used in different eye clinics, it would be difficult to use our machine learning model to predict the recommendations of other clinics. In particular, the data we used was set for the specific measurement equipment and consulting system for surgery selection present at the chosen eye clinic. If the measurement process is changed or the surgical criteria are updated, the machine learning system would be unusable and a new model should be built from the newly collected data.

In the study, machine learning models using XGBoost performed better than other methods based on various performance metrics. Previous studies indicated that tree algorithms performed well in predicting the multiclass disease classification problem.[5,38] Decision tree–based XGBoost exhibited advantages including flexibility, regularization, parallel processing, and feature selection.[39] A previous study revealed that the multiclass OVO XGBoost model performed well in various datasets.[19] The XGBoost model adjusted several scalable variables and can be highly adopted to specific datasets because overfitting is avoided by regularization processing. Although the XGBoost is limited by combining weak learners, it can select important features and can render a complicated problem by constructing sparse classification rules.[17] Parallel computing of XGBoost can offer better optimization via generating a number of examples. As shown in previous studies, XGBoost constitutes an extremely powerful technique and is applied with immense success in several regression and pattern classification problems.[40] However, there is a paucity of studies on the usefulness of XGBoost in ophthalmology. Given its efficiency, it is expected that XGBoost will be increasingly used by medical researchers and clinicians in the future.

In training the OVR classifiers, a severely imbalanced distribution of the training dataset occurred. To overcome this imbalance, the minor training datasets were augmented with SMOTE to obtain balanced training, and the slight improvements in the predictions were experimentally demonstrated. Because the XGBoost and SVM tend to fit large datasets to increase accuracy, SMOTE generally improves the classification performances of these methods.[41] Previous studies have demonstrated successful applications of SMOTE in various datasets.[27,42] However, this technique can sometimes disregard the dominating training set and increase overfitting because the algorithm generates new instances close to an existing cluster of the minor datasets.[43] In SMOTE, the synthesized samples were created within a data space and feature space, and there was no data transformation in this study. A previous study found that SMOTE did not change the data properties by calculating the average, kurtosis, and skewness compared to propensity score matching.[44] This means that the SMOTE did not disturb the data distribution of the contraindication group that included patients with pathologic high myopia, thin cornea, or keratoconus. Because Euclidean distances were used to determine the nearest neighbors in SMOTE, unimportant factors might affect the generation of samples. Although we equally balanced the classes using SMOTE, an optimal imbalanced ratio could exist. Therefore, future studies need to include a more elaborate evaluation of SMOTE to maximize the accuracy of the constructed models.

There are several limitations in the study. First, the study was performed in a single center. Specifically, there was no absolute criteria for corneal refractive surgery selection. Given that SMILE is a relatively novel technique, many eye clinics have not established equipment for SMILE. Therefore, it is not possible to apply our proposed model to other eye clinics. Second, the study did not analyze the cost of surgery in a quantitative method. Although medical indication is important, the cost can significantly affect surgery selection for the patients. Ophthalmic clinics for refractive surgery compete over price and service in South Korea. If there are changes in surgery costs, our machine learning model should be also modified after reviewing new data. Third, the study did not include subjects who underwent customized treatments such as corneal topography-guided laser ablation and collagen cross-linking with laser ablation. The aforementioned methods can affect surgery indication and represent a significant cost increase.[45] Future studies should focus on automatic identification of candidates for customized treatment with a combination of the proposed surgery recommendation framework. Fourth, we did not consider the surgeon factor. In our study, the surgeons possessed significant experience and were skilled in all surgery options. However, it is possible that incorrectly classified cases can be associated with a surgeon's preference for a specific option. Fifth, whether the result of the machine learning algorithm is optimal cannot be decided because there is no ground truth. Interclinician variability in this study also undermines the practical usefulness of this machine learning model from a clinical point of view. Additional postoperative data are required to overcome this fundamental problem.

Overall, medical society is moving toward artificial intelligence, and there is a demand for a better understanding of the operation of machine learning models to promote application of the decision support system. The study shows the potential of a multiclass explainable machine learning method in predicting corneal refractive surgery option on the expert level. Although the study was limited to corneal refractive surgery, the interpretability extension based on the OVR and OVO framework provides more interpretable decision systems in various multiclass medical problems. Explainable machine learning exhibits a promising future to increase the practical use of artificial intelligence in ophthalmic clinics.

Acknowledgments

TKY and IHR contributed equally to this study.

Disclosure: **T.K. Yoo,** None; **I.H. Ryu,** None; **H. Choi,** None; **J.K. Kim,** None; **I.S. Lee,** None; **J.S. Kim**, None; **G. Lee,** Medi-whale, Inc. (E); **T.H. Rim,** Medi-whale, Inc. (R)

References

1. Kim T-I, Del Barrio JLA, Wilkins M, Cochener B, Ang M. Refractive surgery. *Lancet Lond Engl*. 2019;393:2085–2098. doi:10.1016/S0140-6736(18)33209-4
2. Ambrósio R, Wilson S. LASIK vs LASEK vs PRK: advantages and indications. *Semin Ophthalmol*. 2003;18:2–10.
3. Oh E, Yoo TK, Park E-C. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. *BMC Med Inform Decis Mak*. 2013;13:106. doi:10.1186/1472-6947-13-106
4. Caixinha M, Nunes S. Machine learning techniques in clinical vision sciences. *Curr Eye*

Res. 2017;42:1–15. doi:10.1080/02713683.2016.1175019
5. Yoo TK, Ryu IH, Lee G, et al. Adopting machine learning to automatically identify candidate patients for corneal refractive surgery. *Npj Digit Med*. 2019;2:59. doi:10.1038/s41746-019-0135-8
6. Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. Visible machine learning for biomedicine. *Cell*. 2018;173:1562–1565. doi:10.1016/j.cell.2018.05.056
7. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160. doi:10.1109/ACCESS.2018.2870052
8. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749. doi:10.1038/s41551-018-0304-0
9. Lundberg SM, Erion G, Chen H, et al. Explainable AI for trees: from local explanations to global understanding. *ArXiv190504610 Cs Stat*. May 2019. http://arxiv.org/abs/1905.04610. Accessed July 16, 2019.
10. Zhang X, Tan S, Koch P, Lou Y, Chajewska U, Caruana R. Interpretability is harder in the multiclass setting: axiomatic interpretability for multiclass additive models. *ArXiv Prepr ArXiv181009092*. 2018.
11. Salomão M, Hoffling-Lima AL, Lopes B, et al. Recent developments in keratoconus diagnosis. *Expert Rev Ophthalmol*. 2018;13:329–341.
12. Lee JB, Seong GJ, Lee JH, Seo KY, Lee YG, Kim EK. Comparison of laser epithelial keratomileusis and photorefractive keratectomy for low to moderate myopia. *J Cataract Refract Surg*. 2001;27:565–570. doi:10.1016/s0886-3350(00)00880-4
13. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25–34. doi:10.1016/j.jclinepi.2014.09.007
14. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247. doi:10.1016/j.jclinepi.2015.04.005
15. Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res*. 2018;20:e22. doi:10.2196/jmir.9268
16. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H. Hyperparameter optimization for machine learning models based on bayesian optimization. *J Electron Sci Technol*. 2019;17:26–40.
17. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:785–794.
18. Pesantez-Narvaez J, Guillen M, Alcañiz M. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*. 2019;7:70. doi:10.3390/risks7020070
19. Pang L, Wang J, Zhao L, Wang C, Zhan H. A novel protein subcellular localization method with CNN-XGBoost model for Alzheimer's disease. *Front Genet*. 2018;9:751. doi:10.3389/fgene.2018.00751
20. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinforma Oxf Engl*. 2005;21:631–643. doi:10.1093/bioinformatics/bti033
21. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. *PLoS Comput Biol*. 2008;4(10):e1000173.
22. Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans Knowl Data Eng*. 2003;15:1437–1447.
23. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
24. Stojić A, Stanić N, Vuković G, et al. Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. *Sci Total Environ*. 2019;653:140147. doi:10.1016/j.scitotenv.2018.10.368
25. Lundberg SM, Lee S-I. Consistent feature attribution for tree ensembles. *ArXiv Prepr ArXiv170606060*. 2017.
26. Wang S, Yao X. Multiclass imbalance problems: analysis and potential solutions. *IEEE Trans Syst Man Cybern Part B Cybern*. 2012;42:1119–1130. doi:10.1109/TSMCB.2012.2187280
27. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford ExercIse Testing (FIT) project. *PloS One*. 2017;12:e0179805. doi:10.1371/journal.pone.0179805
28. Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. *PloS One*. 2017;12(11):e0187336. doi:10.1371/journal.pone.0187336

29. Jiang J, Zhou Z, Yin E, Yu Y, Liu Y, Hu D. A novel Morse code-inspired method for multiclass motor imagery brain–computer interface (BCI) design. *Comput Biol Med*. 2015;66:11–19. doi:10.1016/j.compbiomed.2015.08.011
30. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J Cheminformatics*. 2017;9:42. doi:10.1186/s13321-017-0226-y
31. Lin SR, Ladas JG, Bahadur GG, Al-Hashimi S, Pineda R. A review of machine learning techniques for keratoconus detection and refractive surgery screening. *Semin Ophthalmol*. 2019;34:317–326. doi:10.1080/08820538.2019.1620812
32. Lopes BT, Eliasy A, Ambrosio R. Artificial intelligence in corneal diagnosis: where are we? *Curr Ophthalmol Rep*. 2019;7:204–211. doi:10.1007/s40135-019-00218-9
33. Tuisku IS, Lindbohm N, Wilson SE, Tervo TM. Dry eye and corneal sensitivity after high myopic LASIK. *J Refract Surg Thorofare NJ* 1995. 2007;23:338–342.
34. Ivarsen A, Asp S, Hjortdal J. Safety and complications of more than 1500 small-incision lenticule extraction procedures. *Ophthalmology*. 2014;121:822–828. doi:10.1016/j.ophtha.2013.11.006
35. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *ArXiv181110154 Cs Stat*. November 2018. http://arxiv.org/abs/1811.10154. Accessed September 27, 2019.
36. Kraus M, Feuerriegel S. Forecasting remaining useful life: interpretable deep learning approach via variational Bayesian inferences. *Decis Support Syst*. 2019;125:113100. doi:10.1016/j.dss.2019.113100
37. Samek W, Müller K-R. Towards Explainable Artificial Intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Muller K-R, Eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2019:5–22. doi:10.1007/978-3-030-28954-6_1
38. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Med Biol Eng Comput*. 2019;57:677–687. doi:10.1007/s11517-018-1915-z
39. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. *Brain Inform*. 2017;4:159–169. doi:10.1007/s40708-017-0065-7
40. Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput Sci*. 2017;3:e127. doi:10.7717/peerj-cs.127
41. Mahmud SMH, Chen W, Jahan H, Liu Y, Sujan NI, Ahmed S. iDTi-CSsmoteB: identification of drug–target interaction based on drug chemical structure and protein sequence using XGBoost with over-sampling technique SMOTE. *IEEE Access*. 2019;7:48699–48714. doi:10.1109/ACCESS.2019.2910277
42. Wang Q, Luo Z, Huang J, Feng Y, Liu Z. a novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM. Computational Intelligence and Neuroscience. doi:10.1155/2017/1827016
43. Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? *ArXiv160908764 Cs*. September 2016. http://arxiv.org/abs/1609.08764. Accessed September 27, 2019.
44. Rivera WA, Goel A, Kincaid JP. OUPS: a combined approach using SMOTE and propensity score matching. In: *2014 13th International Conference on Machine Learning and Applications*. 2014:424–427. doi:10.1109/ICMLA.2014.106
45. Kymionis GD, Portaliou DM, Kounis GA, Limnopoulou AN, Kontadakis GA, Grentzelos MA. Simultaneous topography-guided photorefractive keratectomy followed by corneal collagen cross-linking for keratoconus. *Am J Ophthalmol*. 2011;152:748–755. doi:10.1016/j.ajo.2011.04.033