

RESEARCH

Open Access



Retinal vein occlusion risk prediction without fundus examination using a no-code machine learning tool for tabular data: a nationwide cross-sectional study from South Korea

Na Hyeon Yu¹, Daeun Shin¹, Ik Hee Ryu^{2,3}, Tae Keun Yoo^{4*}  and Kyungmin Koh^{1,5*}

Abstract

Background Retinal vein occlusion (RVO) is a leading cause of vision loss globally. Routine health check-up data—including demographic information, medical history, and laboratory test results—are commonly utilized in clinical settings for disease risk assessment. This study aimed to develop a machine learning model to predict RVO risk in the general population using such tabular health data, without requiring coding expertise or retinal imaging.

Methods We utilized data from the Korea National Health and Nutrition Examination Surveys (KNHANES) collected between 2017 and 2020 to develop the RVO prediction model, with external validation performed using independent data from KNHANES 2021. Model construction was conducted using Orange Data Mining, an open-source, code-free, component-based tool with a user-friendly interface, and Google Vertex AI. An easy-to-use oversampling function was employed to address class imbalance, enhancing the usability of the workflow. Various machine learning algorithms were trained by incorporating all features from the health check-up data in the development set. The primary outcome was the area under the receiver operating characteristic curve (AUC) for identifying RVO.

Results All machine learning training was completed without the need for coding experience. An artificial neural network (ANN) with a ReLU activation function, developed using Orange Data Mining, demonstrated superior performance, achieving an AUC of 0.856 (95% confidence interval [CI], 0.835–0.875) in internal validation and 0.784 (95% CI, 0.763–0.803) in external validation. The ANN outperformed logistic regression and Google Vertex AI models, though differences were not statistically significant in internal validation. In external validation, the ANN showed a marginally significant improvement over logistic regression ($P=0.044$), with no significant difference compared to Google Vertex AI. Key predictive variables included age, household income, and blood pressure-related factors.

Conclusion This study demonstrates the feasibility of developing an accessible, cost-effective RVO risk prediction tool using health check-up data and no-code machine learning platforms. Such a tool has the potential to enhance early detection and preventive strategies in general healthcare settings, thereby improving patient outcomes.

*Correspondence:

Tae Keun Yoo
eyetaekeunyoo@gmail.com; fawoo2@yonsei.ac.kr
Kyungmin Koh
kmkoh@kimeye.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Keywords No-code machine learning, Retinal vein occlusion, Risk factors, Code-free tool

Background

Retinal vein occlusion (RVO) is a major critical health-care issue involving vision loss [1]. RVO is a disease in which vision complications occur due to blockage of the central or branch retinal veins by a thrombus or significant narrowing of retinal veins at the arteriovenous crossing site. Ischemia following RVO can induce macular edema and neovascular complications such as hemorrhage, retinal fibrosis, and secondary glaucoma. The prevalence of RVO has been around 0.5–0.7% in adult populations [2, 3]. Because RVO is a vascular complication, it is closely associated with chronic systematic diseases, including hypertension, diabetes, atherosclerosis, and other cardiovascular diseases [4]. Therefore, by combining information related to multiple vascular diseases, it will be possible to predict and prevent the risk of RVO.

Currently, retinal vein occlusion (RVO) is typically detected through fundus photography or other retinal imaging techniques. However, routine health check-ups often include laboratory tests that assess risk factors associated with RVO, such as blood pressure and cholesterol levels. These tests are advantageous as cardiovascular disease biomarkers due to their low cost and consistency [5]. Although regular health check-ups with medical history questionnaires and laboratory examinations can potentially be RVO biomarkers, a single calculation or test has not been established to predict RVO. Recent advancements in machine learning have demonstrated the capability to integrate diverse data sources for improved diagnostic performance without fundus photography [6]. For instance, this technology has been applied to algorithms predicting diabetic retinopathy without taking fundus photographs [7].

Although evidence has been reported on various risk factors for RVO, previous studies primarily focused on image-based approaches such as using optical coherence tomography (OCT) images to predict RVO [8, 9]. In contrast, our study uniquely integrates demographic, medical history, and laboratory test data in a tabular format to develop a risk prediction model. This approach enables a broader application of machine learning techniques in clinical settings without requiring imaging data, making it more accessible and cost-effective. Many medical researchers face challenges in learning and implementing coding skills to develop machine learning models independently [10]. While data availability has been a barrier to entry, risk prediction remains particularly challenging for diseases with low prevalence, such as RVO. Recently, no-code tools for machine learning development, such as Google Vertex AI and Orange Data Mining, have been introduced, enabling researchers to create predictive

models without requiring programming expertise [11]. These platforms empower medical researchers to utilize readily available data to develop models for various diseases with minimal technical barriers [12]. Furthermore, biases in data stemming from differences in country, center, and race highlight the importance of designing individualized and customized machine learning models [13]. To address these challenges, creating an environment where machine learning models can be developed easily and without coding is essential for advancing predictive healthcare and enabling widespread adoption of these tools.

The research gap for machine learning studies to predict RVO can be summarized as follows: Prior studies predominantly focus on imaging data, such as fundus photographs or OCT scans, which require specialized equipment and expertise, limiting their applicability in settings lacking access to such tools [14]. Additionally, while health check-up data—encompassing demographics, medical history, and laboratory results—are widely available [15], their potential use in machine learning models for RVO prediction remains underexplored. Another significant barrier is the reliance on programming skills for developing machine learning models, which restricts accessibility for many medical researchers and clinicians [10, 12], highlighting the need for user-friendly, no-code platforms. Furthermore, the low prevalence of RVO presents challenges in addressing imbalanced datasets, which many existing studies fail to overcome, leading to suboptimal predictive performance. Lastly, most current models are based on small, homogeneous datasets, raising concerns about their generalizability to larger and more diverse populations.

Our study aimed to create a machine learning model that can identify the risk of RVO in a general population using health check-up data. This data includes demographic data, medical history, and laboratory tests. To select high-risk patients to receive fundus examination or who should control their systemic risk factors, we used traditional risk factors and clinical laboratory examinations without fundus examination to predict RVO. We attempted to develop RVO risk prediction models using general healthcare data in a large population-based dataset with more than 14,000 participants. The key contributions of this study are as follows:

- Development of a Machine Learning Model for RVO: We developed a machine learning model using user-friendly, code-free platforms (Orange Data Mining and Google Vertex AI), enabling medical researchers without coding experience to create accurate

predictive models. Notably, this is the first study to predict the risk of RVO using routine screening data without the need for imaging. Additionally, we introduced an easily accessible oversampling function that does not require additional coding, enhancing the usability of the workflow.

- Application to a Large, Nationwide Dataset: The study utilized health check-up data, including demographics, medical history, and laboratory results, demonstrating the feasibility of predicting RVO risk without imaging data. The study employed data from the Korea National Health and Nutrition Examination Surveys (KNHANES), with a development dataset of over 12,000 participants and an independent external validation set, ensuring robustness and generalizability.
- Comparison of No-Code Platforms: This study compared the performance of two no-code platforms, highlighting their strengths and limitations for clinical applications. By eliminating the need for coding or specialized imaging equipment, this approach provides a low-cost, accessible tool for identifying high-risk individuals in general healthcare settings.

Methods

Study design and participants

This study developed machine learning prediction models for RVO risk using medical history and laboratory examination data (Fig. 1). This study used data from the

Korea National Health and Nutrition Examination Surveys (KNHANES) conducted between 2017 and 2021. The KNHANES is a nationwide cross-sectional study conducted by the Korea Disease Control and Prevention Agency (KDCA). The data collection protocol was approved by the Institutional Review Board (IRB) of the KDCA, and the dataset is publicly available for research purposes (link: <https://knhanes.kdca.go.kr/knhanes/ng/main.do>). Informed consent was obtained from all subjects in the study in the data collection stage. Ethical approval for this study was waived by the institutional review board of the Korean National Institute for Bioethics Policy. The study adhered to the guidelines of the Declaration of Helsinki. All participants in KNHANES were selected using stratified random sampling in which the following factors were considered: sex, age, and residential area [16]. KNHANES comprised health records based on health interviews, examinations, and nutrition surveys. Each participant provided health and socioeconomic information regarding age, household income, alcohol use, smoking status, presence of hypertension, diabetes, dyslipidemia, stroke, heart disease, osteoarthritis, and osteoporosis by completing a questionnaire. The health examinations included body mass index (BMI), routine blood tests, and biochemistry tests in a general check-up. All participants underwent laboratory tests after overnight fasting.

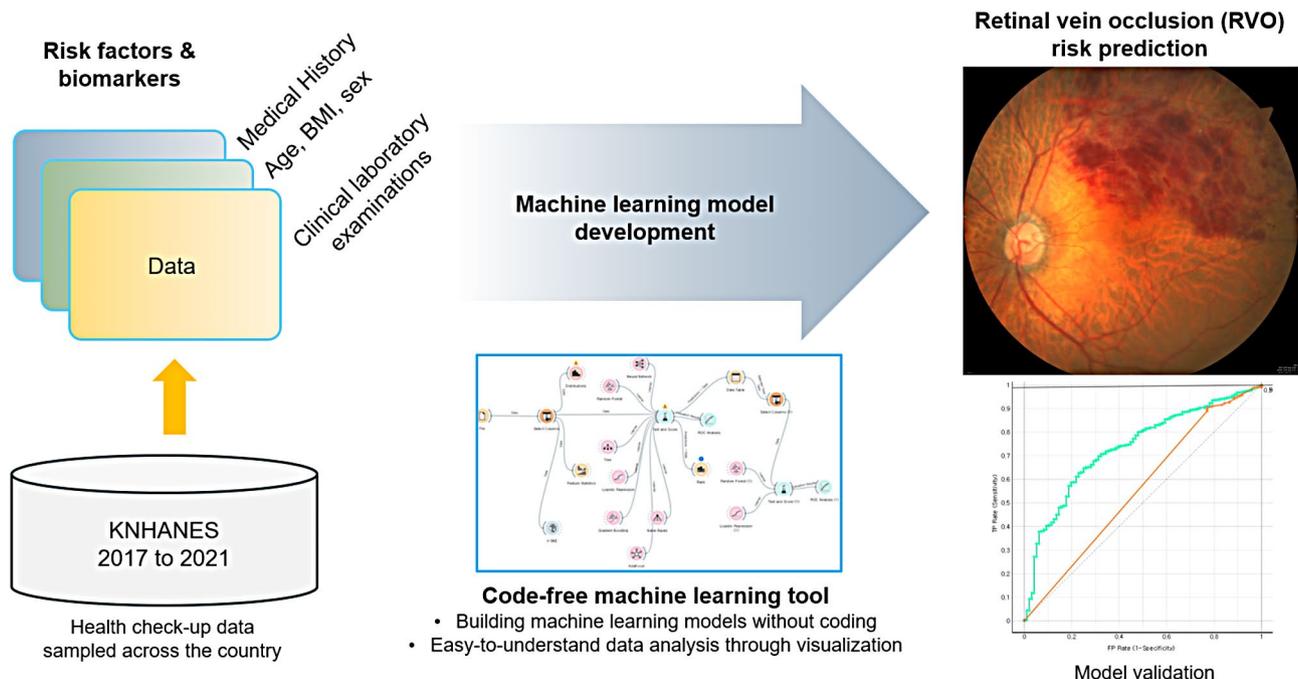


Fig. 1 Schematic diagram of the overview of this study

Determining RVO status

Previous studies reported detailed methods to determine RVO status in KNHANES data collection [17, 18]. Eye examination quality was controlled by the Epidemiologic Survey Committee of the Korean Ophthalmologic Society (KOS). Participating ophthalmologists or residents were periodically trained by acting staff members of the National Epidemiologic Survey Committee of the KOS. The data quality and collection protocols were verified by the KDCA. In KNHANES, non-mydratic fundus photography (VISUCAM, Carl Zeiss Meditec, Jena, Germany), and macular OCT (Cirrus HD-OCT 500, Carl Zeiss Meditec, Jena, Germany) were performed. Experienced retinal specialists certified by the Korean Retina Society graded the presence of RVO. An independent grader graded all the fundus photography and OCT images twice. If there was a disagreement in the primary diagnosis, a reading committee from the Korean Retina Society determined the final diagnosis of RVO. This study did not distinguish between branch RVO (BRVO) and central RVO (CRVO).

Data preprocessing

The input data included demographic, clinical, and laboratory parameters such as age, BMI, household income level, alcohol use, smoking status, systolic blood pressure (SBP), diastolic blood pressure (DBP), and the presence of hypertension, diabetes, dyslipidemia, stroke, heart disease, osteoarthritis, and osteoporosis. Laboratory

evaluations comprised fasting plasma glucose (FPG), total cholesterol, triglycerides (TG), aspartate aminotransferase (AST), alanine aminotransferase (ALT), creatinine, white blood cell count (WBC), hemoglobin, and platelet counts. Figure 2 illustrates the detailed data workflow, including inclusion and exclusion criteria for model development. KNHANES collected RVO diagnosis data between 2017 and 2021; datasets from earlier or later periods were excluded due to the absence of RVO evaluation. Participants were included if they were aged 40 years or older and had complete data from interviews, health examinations, and blood tests. Missing or incomplete data, including null values in key variables, incomplete RVO evaluations, or missing demographic or clinical information, resulted in participant exclusion. These preprocessing steps ensured the dataset's quality and integrity for machine learning analysis, supporting the reproducibility and reliability of the study findings.

We established a research design to develop and validate machine learning models in chronological order (data split by calendar time) [19]. The RVO prediction models were developed using KNHANES data between 2017 and 2020 (development dataset). Because KNHANES randomly resampled participants every year, the performance of the developed models was evaluated in the independent data from KNHANES 2021. We set KNHANES 2021 data as external validation (Fig. 2). This scheme was to design the retrospective development of the machine models and consecutive prospective

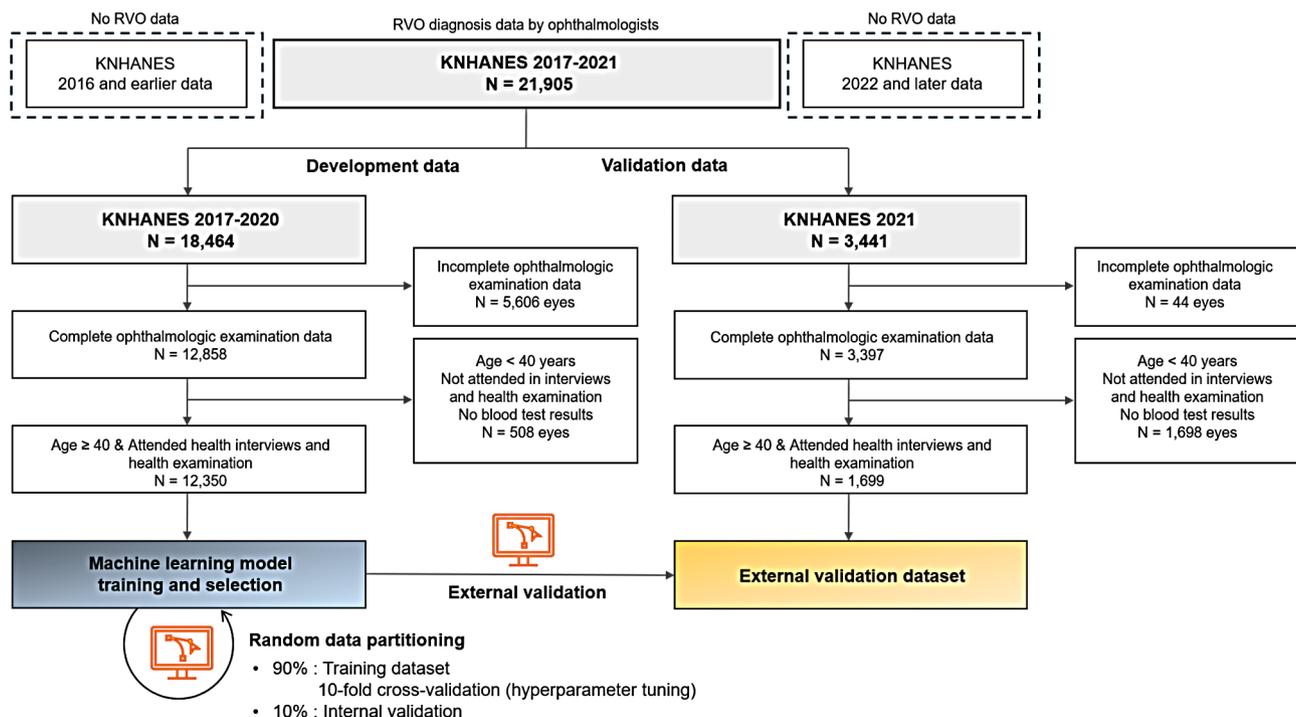


Fig. 2 Data workflow with inclusion and exclusion criteria for the machine learning model development in this study

validation via chronological splitting [20]. Among the development set, 90% of the data was randomly selected as the training dataset, while the remaining 10% was used as the internal validation dataset. Tenfold cross-validation was performed exclusively within the training dataset to optimize hyperparameters and evaluate model performance during training. We searched and trained optimal models using the training data through tenfold cross-validation. After selecting the optimal hyperparameters through tenfold cross-validation, the final model was trained using the entire development dataset (training and internal validation set). This final model was subsequently evaluated on both the internal validation set and the external validation set as independent tests. This approach ensured that no information from the validation sets influenced the model training or hyperparameter tuning, maintaining the integrity of the evaluation. The KNHANES health data is organized at the patient level, ensuring that all data from the same patient is assigned to either the training split or the internal validation split, with no overlap. This thorough partitioning aligns with the rigorous data collection and sampling methodology of KNHANES.

The same dataset was used for both Orange Data Mining and Google Vertex AI to ensure a fair comparison between the two platforms. In Orange Data Mining, tenfold cross-validation was performed to optimize hyperparameters and evaluate model performance during training. In contrast, Google Vertex AI did not require tenfold cross-validation, as it automatically performed hyperparameter tuning as part of its training process.

A fair performance comparison was conducted using manually divided data without coding. Model training and validation were performed without manual normalization of input variables. Manual normalization would complicate the no-code machine learning workflow, as it requires additional preprocessing steps that can be burdensome for users in later applications. The no-code software (Orange Data Mining and Google Vertex AI) is designed to handle input data directly and applies necessary preprocessing internally, depending on the requirements of each algorithm. This approach ensures simplicity and usability for non-technical users while maintaining the performance of the machine learning models.

Machine learning development

We developed machine learning models based on the KNHANES datasets using no-code software. In this study, we adopted the Orange Data Mining version 3.36.2 (Bioinformatics Laboratory, University of Ljubljana, Ljubljana, Slovenia) [21]. Orange Data Mining is a graphic component-based code-free tool that allows machine learning algorithm development (Fig. 3). Unlike the no-code machine learning development services provided by platform companies, Orange Data Mining is free, open-source software. Products made by Orange Data Mining can be modified without any restriction. They can also be distributed with development and documentation under a General Public License, published by the Free Software Foundation (<https://orangedatamining.com/license/>). It provides basic statistical analysis and major

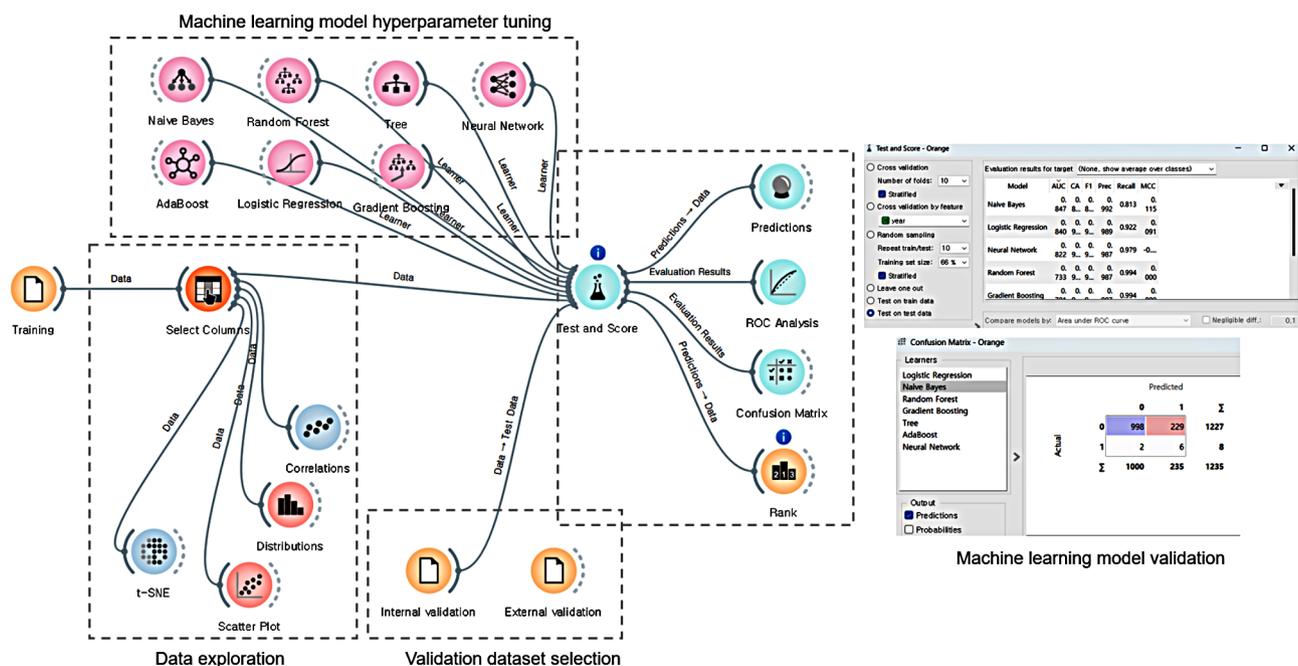


Fig. 3 Graphical component-based Orange Data Mining workflow

predictions by representing the proportion of individuals predicted to be low-risk who are indeed free of RVO. These metrics collectively provide a well-rounded assessment of the model's predictive capabilities, with sensitivity and specificity being particularly critical for balancing the identification of high-risk individuals while avoiding unnecessary misclassification. The data distribution between the two centers was compared using the chi-square test for categorical variables and the t-test for continuous variables. These tests were performed two-sided, with a significance level of P value < 0.05 . Machine learning development and validation were performed using Orange Data Mining and Google Vertex AI. All statistical analyses were performed using MedCalc Version 22.021 (Mariakerke, Belgium).

Results

The data was successfully prepared in a 1:9 ratio using the web-based SMOTE function, eliminating the need for coding. All machine learning training in Orange Data Mining was completed within 10 min without requiring coding expertise, although hyperparameter tuning was necessary. In contrast, Google Vertex AI did not require manual hyperparameter tuning but took approximately 3 h for training. Using grid search via the tenfold cross-validation in Orange Data Mining, we found the optimal hyperparameters of ANN with 100 neurons in hidden layers with ReLu activation function and Adam optimizer. In Random Forest, grid search showed the best performance when the number of decision trees was 500, and the number of attributes considered at each split was five. Gradient Boosting showed the best performance when the number of trees was 1000, the limit depth of trees was 3, and the learning rate was 0.1. Decision Trees showed optimal performance with a maximum tree depth of 250 and a maximum number of features in leaves of 3. We used the naïve Bayes algorithm with default settings since Orange Data Mining did not offer a tunable option. In Google Vertex AI, we adopted the AutoML approach with a maximum of 2 node hours to search and train the machine learning model.

The characteristics and laboratory data of the study participants are summarized in Table 1. The prevalence of RVO in the development and external validation datasets was 0.7% and 0.5%, respectively, with no statistical difference (P -value = 0.419). However, differences in age, smoking, frequency of alcohol consumption, systolic blood pressure, high blood pressure, presence of hypertension, diabetes, hyperlipidemia, stroke, heart disease, osteoarthritis, and osteoporosis between the development and external validation datasets were statistically significant (P -value < 0.001).

Figure 5 shows the development data distribution and trained model (Decision Tree) exploration performed

using the Orange Data Mining software. The t-distributed stochastic neighbor embedding (t-SNE) visualized whether the whole data distributions were predictable by machine learning according to the presence of RVO. In addition, we could search the distribution of data between two variables labeled by the presence of RVO. The detailed decision tree with classification criteria can be displayed using the model viewer. Google Vertex AI did not provide a model viewer to view the internal structure specifically.

We conducted a classical statistical analysis to identify the risk factors of RVO. Table 2 shows the results of the binary logistic regression models for the RVO risk. In the univariate analysis, RVO risk was associated with age, household income, SBP, hypertension, stroke, osteoporosis, and WBC. After stepwise backward feature selection, the final model included age, smoking, DBP, hypertension, diabetes, TG, and WBC. The absence of diabetes or lower TG levels was more strongly associated with RVO. Additionally, we utilized the logistic regression results from Table 2 as input for ChatGPT-4 to create a no-code risk calculator. The code generated by ChatGPT-4 is included in the Supplementary Material 1, and the interactive calculator is publicly accessible at https://taekeunt.oo.github.io/RVO_risk_calc/. This calculator allows users to estimate the risk of RVO based on the identified key predictors, providing a practical tool for further application of our findings.

Figure 6 shows the feature importance calculations using Orange Data Mining and Google Vertex AI. In the case of Orange Data Mining, information gain was measured in the Random Forest and Gradient Boosting algorithms, which showed the highest internal tuning performance. ANN and naïve Bayes did not measure the importance of features inside the model. Age was selected as the most important factor in all algorithms. In addition, blood pressure-related factors and household income were selected as the top important factors.

Figure 7 displays ROC curves of tenfold cross-validation results in the training set, internal validation, and external validation. In Orange Data Mining, the final hyperparameters selected through cross-validation were used to train the final model on the entire training dataset. This final model was then evaluated on the internal and external validation sets as independent tests, ensuring that no information from either validation set influenced model training or hyperparameter tuning. The naïve Bayes algorithm showed the highest performance in predicting RVO in tenfold cross-validation and had an AUC of 0.698. The AUCs of logistic regression, ANN, Random Forest, and Gradient Boosting in tenfold cross-validation were 0.692, 0.695, 0.629, and 0.620, respectively. ANN performed best in the internal and external validation datasets, with 0.856 and 0.784, respectively.

Table 1 Demographics and clinical features in this study

Category	Variable	Development set (training and internal validation) KNHANES 2017 to 2020, N= 12,350)	External validation set (KNHANES 2021, N= 1,699)	P-value
Eye examination	RVO (%)	82 (0.7)	8 (0.5)	0.419
Demographics and BMI	Age (years)	59.79 ± 11.55	49.87 ± 5.80	0.001
	Sex, female (%)	7009 (56.8)	972 (57.2)	0.734
	BMI (kg/m ²)	24.17 ± 3.27	24.25 ± 3.64	0.367
	Household income			< 0.001
	Very low (%)	2738 (22.2)	120 (7.1)	
	Low (%)	3108 (25.2)	354 (20.8)	
	Moderate (%)	3161 (25.6)	524 (30.8)	
	High (%)	3343 (27.1)	693 (40.8)	
Medical History and blood pressure	Current smoker (%)	1963 (15.9)	333 (19.6)	< 0.001
	Alcohol, ≥ 1 drink/week (%)	4866 (39.4)	775 (45.6)	< 0.001
	Blood pressure			
	SBP (mmHg)	122.21 ± 16.68	117.02 ± 14.63	< 0.001
	DBP (mmHg)	76.01 ± 9.94	75.86 ± 10.30	0.565
	Hypertension (%)	3941 (31.9)	277 (16.3)	< 0.001
	Diabetes mellitus (%)	1602 (13.0)	127 (7.5)	< 0.001
	Dyslipidemia (%)	2838 (23.0)	300 (17.7)	< 0.001
	Stroke (%)	271 (2.2)	7 (0.4)	< 0.001
	Heart disease (%)	438 (3.5)	14 (0.8)	< 0.001
	Osteoarthritis (%)	1894 (15.3)	97 (5.7)	< 0.001
	Osteoporosis (%)	1077 (8.7)	37 (2.2)	< 0.001
Laboratory data	Fasting glucose (mg/dL)	104.64 ± 25.2	102.87 ± 22.49	0.006
	Total cholesterol (mg/dL)	192.23 ± 39.55	199.45 ± 39.13	< 0.001
	Triglyceride (mg/dL)	138.34 ± 104.08	137.40 ± 117.04	0.753
	AST (IU/L)	24.91 ± 15.71	25.05 ± 12.04	0.731
	ALT (IU/L)	22.89 ± 16.38	24.90 ± 18.47	< 0.001
	Creatinine (mg/dL)	0.81 ± 0.26	0.78 ± 0.26	< 0.001
	WBC (10 ³ cells/μL)	6.08 ± 1.71	5.92 ± 1.60	< 0.001
	Hemoglobin (mg/dL)	13.89 ± 1.54	13.91 ± 1.56	0.530
	Platelet (10 ⁹ /L)	251.74 ± 64.82	260.04 ± 60.94	< 0.001

AST, aspartate aminotransferase; ALT, alanine aminotransferase; BMI, body mass index; DBP, diastolic blood pressure; RVO, retinal vein occlusion; SBP, systolic blood pressure; WBC, white blood cell

In contrast, Google Vertex AI automatically performed hyperparameter tuning as part of its training process, eliminating the need for manual cross-validation. The final model from Google Vertex AI achieved an AUC of 0.842 in internal validation and 0.781 in external validation. This automated approach provides a streamlined alternative to the manual hyperparameter optimization required in Orange Data Mining.

Table 3 shows detailed prediction performance metrics of all machine learning methods in the internal and external validation datasets. In the internal validation, ANN, naive Bayes, and Google Vertex AI algorithms showed better AUC values than logistic regression, but the differences were not statistically significant. In the external validation, ANN shows a marginally more significant AUC than logistic regression (P-value = 0.044). Random Forest, naive Bayes, and Google Vertex AI algorithms showed better AUC values than logistic regression, but

differences were not statistically significant in the external validation.

Discussion

In this machine learning-based study using national cross-sectional datasets, we built a novel algorithm to predict patients with a high risk of RVO. The machine learning model consistently performed well in the internal and external datasets by integrating the features from demographics, medical history, and laboratory tests. A key innovation of this study is the use of no-code machine learning platforms, which allowed the model to be developed without any coding experience. By leveraging user-friendly tools such as Orange Data Mining and Google Vertex AI, we demonstrated that machine learning research can be conducted even by medical researchers without extensive technical expertise. This lowers the barrier to entry for implementing AI-driven solutions in

Table 2 Logistic regression analysis of clinical features for the risk of RVO in the development dataset

	Univariate model*		Adjusted model†	
	OR (95% CI)	P-value	OR (95% CI)	P-value
Age (year, per 1 unit increase)	1.06 (1.04–1.08)	< 0.001	1.06 (1.03–1.08)	< 0.001
Sex, female	0.95 (0.62–1.44)	0.810		
BMI (kg/m ² , per 1 unit increase)	1.04 (0.98–1.10)	0.198		
Household income				
Very low	Reference			
Low	0.55 (0.32–0.94)	0.029		
Moderate	0.44 (0.25–0.78)	0.005		
High	0.34 (0.19–0.62)	< 0.001		
Current smoker (%)	1.02 (0.58–1.78)	0.934	1.42 (0.78–2.57)	0.252
Alcohol, ≥ 1 drink/week (%)	0.67 (0.43–1.05)	0.081		
Blood pressure				
SBP (mmHg, per 1 unit increase)	1.03 (1.01–1.04)	< 0.001		
DBP (mmHg, per 1 unit increase)	1.02 (0.99–1.04)	0.147	1.03 (1.01–1.05)	0.004
Hypertension	4.07 (2.65–6.25)	< 0.001	2.78 (1.74–4.44)	< 0.001
Diabetes mellitus	0.79 (0.39–1.57)	0.505	0.47 (0.23–0.95)	0.036
Dyslipidemia	1.49 (0.95–2.35)	0.082		
Stroke	2.95 (1.18–7.33)	0.020		
Heart disease	2.16 (0.94–4.98)	0.069		
Osteoarthritis	1.31 (0.76–2.25)	0.327		
Osteoporosis	2.15 (1.21–3.82)	0.009		
Fasting glucose (mg/dL, per 1 unit increase)	0.99 (0.98–1.01)	0.630		
Total cholesterol (mg/dL, per 1 unit increase)	0.99 (0.98–1.00)	0.051		
Triglyceride (mg/dL, per 1 unit increase)	0.99 (0.99–1.00)	0.226	0.99 (0.99–1.00)	0.088
AST (IU/L, per 1 unit increase)	0.99 (0.98–1.02)	0.868		
ALT (IU/L, per 1 unit increase)	0.99 (0.98–1.01)	0.459		
Creatinine (mg/dL, per 1 unit increase)	1.37 (0.95–1.98)	0.096		
WBC (10 ³ cells/μL, per 1 unit increase)	1.12 (1.01–1.25)	0.039	1.10 (0.98–1.23)	0.117
Hemoglobin (mg/dL, per 1 unit increase)	1.00 (0.87–1.14)	0.999		
Platelet (10 ⁹ /L, per 1 unit increase)	0.99 (0.99–1.00)	0.224		

* Univariate logistic regression analysis for each variable

† Multivariable logistic regression with stepwise backward selection model

With the growing popularity of artificial intelligence, code-free tools for deep learning model development, particularly for analyzing medical images, have gained attention [10, 30]. However, there remains a significant need for research focusing on the application of code-free tools to tabular medical data [12]. This study demonstrates that machine learning research can be effectively conducted in the medical domain using tabular data with Orange Data Mining software, a code-free and user-friendly platform that enables machine learning development without requiring programming expertise. Additionally, the software supports external integration of developed models through programming languages like Python, enhancing its versatility. To address class imbalance in tabular data, we introduced an easy-to-use, webpage-based SMOTE function (https://taekeuntoo.github.io/SMOTE_web/) for no-code development, which streamlined the workflow. Compared to Google's AutoML, Orange Data Mining surpasses it in terms of usability and scalability, as applied in this study. In this

study, the ReLu-based ANN and Naive Bayes algorithms performed well, demonstrating robustness on imbalanced data without overfitting [31, 32], outperforming tree-based models in this scenario. Furthermore, there have been recent advances in performing regression analysis without coding by utilizing multimodal chatbots such as ChatGPT-4 [33]. In this study, we validated this approach by successfully creating a logistic regression calculator using ChatGPT-4. While the integration of such chatbots into machine learning workflows is still in its early stages, further developments in this area are anticipated, offering even greater accessibility to AI-driven insights. The use of no-code tools for machine learning development holds the potential to significantly simplify and accelerate the creation of accurate prediction models for various diseases. By removing the technical barrier of coding, these tools can drive the adoption of AI technologies in clinical practice, making them accessible to a broader range of medical researchers and practitioners.

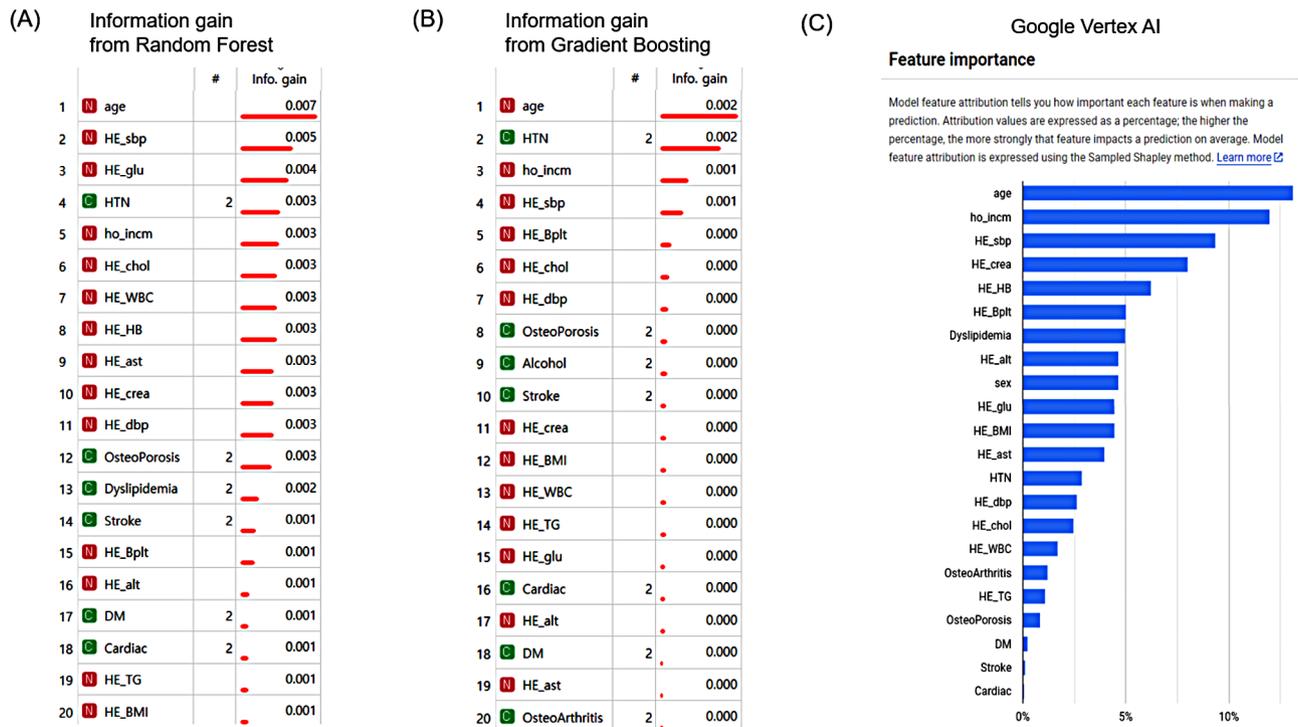


Fig. 6 Feature importance plots. **(A)** Information gain from Random Forest. **(B)** Information gain from Gradient Boosting. **(C)** Feature importance from Google Vertex AI

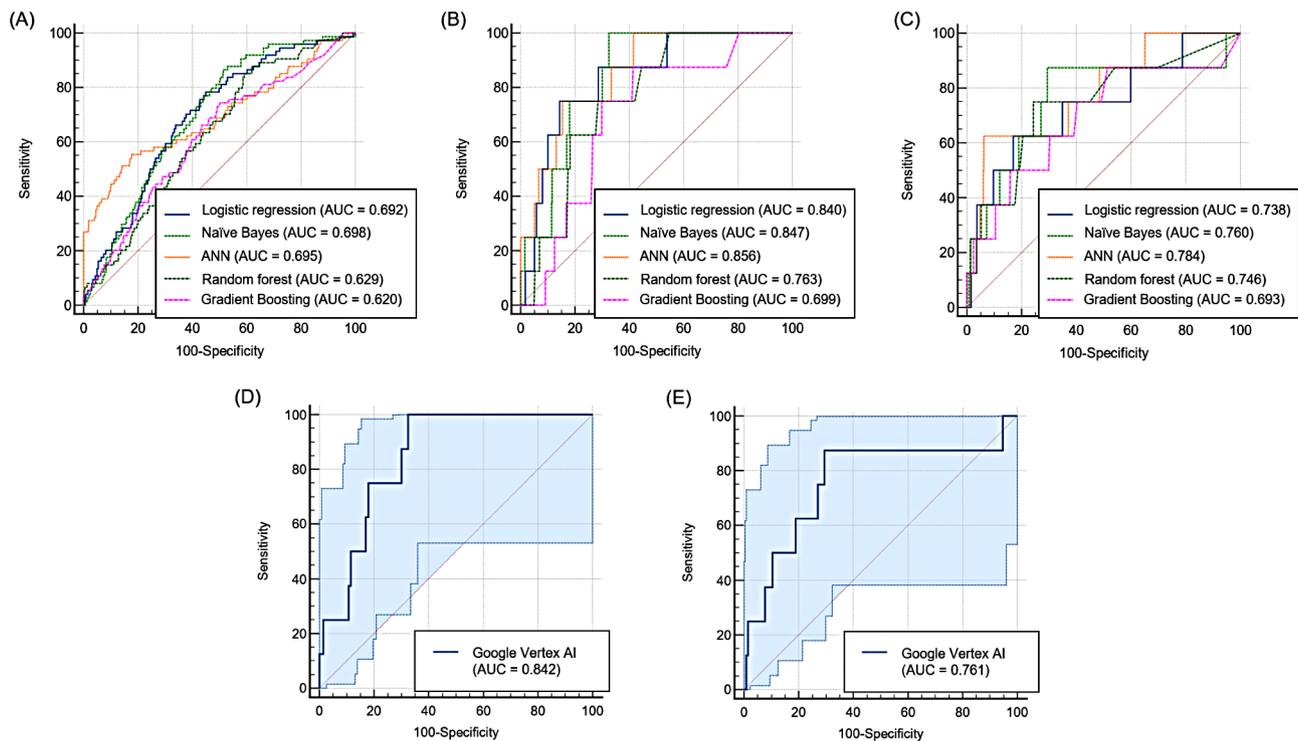


Fig. 7 ROC curves of the developed models to predict RVO. **(A)** Ten-fold cross-validation result from Orange Data Mining. **(B)** Internal validation from Orange Data Mining. **(C)** External validation from Orange Data Mining. **(D)** Internal validation from Google Vertex AI. **(E)** External validation from Google Vertex AI

Table 3 Diagnostic performance in the internal and external validation sets

Validation	Classification algorithm	ROC-AUC (95% CI)	Sensitivity (% 95% CI)	Specificity (% 95% CI)	PPV (% 95% CI)	NPV (% 95% CI)	P-value
Internal validation	Logistic regression	0.840 (0.818, 0.860)	75.0 (34.9, 96.8)	85.6 (83.5, 87.5)	3.3 (2.1, 4.9)	99.8 (99.3, 99.9)	Reference
	Decision tree	0.569 (0.541, 0.597)	100.0 (63.1, 100.0)	13.8 (11.9, 15.9)	0.7 (0.7, 0.8)	100.0 (97.8, 100.0)	< 0.001
	Gradient Boosting	0.699 (0.673, 0.725)	87.5 (47.3, 99.6)	58.5 (55.7, 61.2)	1.3 (1.0, 1.7)	99.8 (99.1, 99.9)	0.003
	Random Forest	0.763 (0.739, 0.787)	75.0 (34.9, 96.8)	71.5 (68.8, 73.9)	1.7 (1.1, 2.5)	99.7 (99.2, 99.9)	0.049
	naïve Bayes	0.847 (0.826, 0.867)	100.0 (63.1, 100.0)	67.4 (64.7, 70.1)	1.9 (1.8, 2.1)	100.0 (99.5, 100.0)	0.837
	ANN	0.856 (0.835, 0.875)	75.0 (34.9, 96.8)	84.5 (82.4, 86.5)	3.1 (2.0, 4.6)	99.8 (99.3, 99.9)	0.601
	Google Vertex AI	0.842 (0.820, 0.861)	100.0 (63.1, 100.0)	66.7 (63.9, 69.3)	1.9 (1.7, 2.0)	100.0 (99.5, 100.0)	0.848
External validation	Logistic regression	0.738 (0.716, 0.759)	62.5 (24.5, 91.4)	82.9 (81.1, 84.7)	1.7 (0.9, 2.9)	99.7 (99.4, 99.9)	Reference
	Decision tree	0.546 (0.522, 0.570)	87.5 (47.3, 99.6)	21.5 (19.5, 23.5)	0.5 (0.4, 0.7)	99.7 (98.3, 99.9)	0.008
	Gradient Boosting	0.693 (0.670, 0.715)	87.5 (47.3, 99.6)	48.5 (46.1, 50.9)	0.8 (0.6, 1.0)	99.8 (99.2, 99.9)	0.612
	Random Forest	0.746 (0.724, 0.766)	75.0 (34.9, 96.8)	75.6 (73.5, 77.7)	1.4 (0.9, 2.1)	99.8 (99.4, 99.9)	0.898
	naïve Bayes	0.760 (0.739, 0.780)	87.5 (47.3, 99.6)	70.5 (68.3, 72.7)	1.4 (1.0, 1.8)	99.9 (99.4, 99.9)	0.765
	ANN	0.784 (0.763, 0.803)	62.5 (24.5, 91.4)	93.7 (92.4, 94.8)	4.5 (2.6, 7.7)	99.8 (99.5, 99.9)	0.044
	Google Vertex AI	0.761 (0.740, 0.781)	87.5 (47.3, 99.6)	70.7 (68.5, 72.9)	1.4 (1.0, 1.8)	99.9 (99.4–99.9)	0.756

ANN, artificial neural networks; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value; ROC-AUC, the areas under the receiver operating characteristic curve

This study demonstrates that RVO risk can be estimated through machine learning by combining multiple risk factors. Previous research has established that various risk factors contribute to the development of RVO [34]. However, the structure and crossing of arteriovenous vessels, particularly in BRVO, play a significant role in its development [35], making it challenging to accurately predict RVO using systemic risk factors alone. Our findings confirmed that factors related to blood pressure have a significant impact on the occurrence of RVO. In contrast, the relationship between diabetes and RVO was either reversed or not significant, suggesting the need for further investigation into the mechanisms underlying RVO development in Koreans. While previous studies have identified both blood pressure and diabetes as critical risk factors for RVO [4, 36], the weaker association with diabetes observed in this study may reflect the active diabetes screening and treatment efforts currently underway in South Korea [37]. Socioeconomic factors could also have influenced the decreased relationship between diabetes and RVO in this population, highlighting the importance of considering regional and demographic contexts in future research.

RVO cannot be perfectly predicted due to its dependence on the structural anatomy of blood vessels, particularly the configuration of arteriovenous crossings [38, 39]. However, predicting vascular health and assessing the risk of RVO occurrence offers a preventive approach. This aligns with the principles of oculomics, which uses retinal analysis as a proxy for evaluating systemic vascular health [40, 41]. By leveraging a machine learning model trained on widely accessible health check-up data, clinicians can identify high-risk individuals without the need for specialized imaging tools, such as fundus photography or OCT. This cost-effective and scalable

approach enables risk stratification in primary care and general health screening settings, facilitating early referrals to ophthalmologic specialists and timely interventions. High-risk individuals could benefit from targeted management of modifiable risk factors, such as hypertension and smoking, reducing their likelihood of developing RVO. Furthermore, the model's no-code design allows seamless integration into routine clinical workflows, making it accessible even to healthcare providers with limited technical expertise [33]. Ultimately, this approach has the potential to improve patient outcomes through proactive risk management, optimized resource allocation, and a reduction in the burden of advanced RVO-related complications on healthcare systems.

There are several limitations in this study. First, the cross-sectional nature of KNHANES data collection hinders us from developing a machine-learning model for future RVO prediction. A longitudinal follow-up research design is required to build the future RVO development prediction model. Second, this study data was based on a single Asian country, raising uncertainty about the generalizability of our developed models to other countries or ethnic groups. Both retinal vasculature and cardiovascular disease risk may differ between races [42, 43]. Third, laboratory tests and body mass index measurements may vary depending on the timing of data collection. Fourth, the lack of information on the type of RVO is another limitation. Because CRVO and BRVO may have different pathophysiology [44], the absence of this information might have confounded our results. Additionally, the use of no-code tools for machine learning, while advantageous for accessibility and ease of use, introduces limitations such as reduced flexibility in model customization and reliance on built-in functionalities [45]. Future studies could explore hybrid approaches that combine

no-code tools with traditional coding environments to balance accessibility and technical control.

This study highlights opportunities for future research. While this study focused on tabular health check-up data, integrating imaging data such as fundus photographs or OCT scans in multimodal approaches could enhance longitudinal prediction accuracy. The rarity of RVO presents challenges in handling imbalanced datasets, even with SMOTE, suggesting that future works explore alternative methods like cost-sensitive learning. Moreover, the lack of differentiation between BRVO and CRVO subtypes limits the applicability of the findings, and future models should address this. Model interpretability remains an issue, particularly with ANN models, warranting further research into explainable AI techniques to improve clinical transparency. Finally, real-time, automated tools that integrate with electronic medical records could streamline RVO risk assessments during routine health check-ups, expanding the practical applications of this approach.

Conclusion

We developed a machine learning model for RVO risk prediction using health check-up data, incorporating demographic information, medical history, and laboratory test results, without requiring fundus examination. To overcome the technical barriers associated with coding-based machine learning model development, we utilized a code-free and user-friendly tool. This approach provides a low-cost and accessible solution for RVO risk prediction in general health check-up settings. Additionally, our methodology demonstrates the potential of using tabular data for developing prediction models for various low-prevalence diseases. Further research using larger RVO datasets and data from diverse regions is needed to validate the feasibility and generalizability of this approach.

Abbreviations

ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
AUC	Areas under the receiver operating characteristic curve
ANN	Artificial neural networks
BMP	Body mass index
DPB	Diastolic blood pressure
FPG	Fasting plasma glucose
KNHANES	Korea National Health and Nutrition Examination Surveys
NPV	Negative predictive value
OCT	Optical coherence tomography
PPV	Positive predictive value
RVO	Retinal vein occlusion
SBP	Systolic blood pressure
SMOTE	Synthetic minority oversampling technique
TG	Triglycerides

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-025-02950-8>.

Supplementary Material 1

Acknowledgements

None.

Author contributions

NHY and DS acquired and analyzed data, interpreted the results, and drafted the manuscript. TKY and KK suggested the original study idea, interpreted the results, contributed to writing. IHR, TKY, and KK analyzed data and contributed to data interpretation and manuscript editing.

Funding

None.

Data availability

All the data supporting the findings of this study are available within the article. The data is available to the public (<https://knhanes.kdca.go.kr/knhanes/eng/index.do>, <https://knhanes.kdca.go.kr/knhanes/eng/main.do>) for research purposes.

Code availability

This study used no-code tools, so there is no code for developing machine learning models.

Declarations

Ethics approval and consent to participate

The KNHANES is a nationwide cross-sectional study conducted by the Korea Disease Control and Prevention Agency (KDCA). The data collection protocol was approved by the institutional review board (IRB) at the KDCA. Ethical approval for this study was waived by the institutional review board of the Korean National Institute for Bioethics Policy and informed consent from the patients was also waived. The study adhered to the guidelines of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

IHR is the director of VISUWORKS and holds company stock. IHR also serves on the Advisory Board for Carl Zeiss Meditec AG and Avellino Lab USA/MAB, as well as for Avellino Lab Korea. The remaining authors declare no conflicts of interest.

Author details

¹Department of Ophthalmology, Kim's Eye Hospital, Konyang University College of Medicine, Seoul, South Korea

²Department of Refractive Surgery, B&VIIT Eye Center, Seoul, South Korea

³Research and Development Department, VISUWORKS, Seoul, South Korea

⁴Department of Ophthalmology, Hangil Eye Hospital, 35 Bupyeong-daero, Bupyeong-gu, Incheon 21388, South Korea

⁵Cornea, Cataract and Refractive Surgery Division Kim's Eye Hospital, Konyang University College of Medicine, 136 Yeongshinro, Youngdeungpogu, Seoul 07301, Republic of Korea

Received: 12 April 2024 / Accepted: 24 February 2025

Published online: 07 March 2025

References

1. Laouri M, Chen E, Looman M, Gallagher M. The burden of disease of retinal vein occlusion: review of the literature. *Eye*. 2011;25:981–8.
2. Lim LL, Cheung N, Wang JJ, Islam FMA, Mitchell P, Saw SM, et al. Prevalence and risk factors of retinal vein occlusion in an Asian population. *Br J Ophthalmol*. 2008;92:1316–9.
3. Rogers S, McIntosh RL, Cheung N, Lim L, Wang JJ, Mitchell P, et al. The prevalence of retinal vein occlusion: pooled data from population studies

- from the united States, Europe, Asia, and Australia. *Ophthalmology*. 2010;117:313–e3191.
4. Chang Y-S, Ho C-H, Chu C-C, Wang J-J, Jan R-L. Risk of retinal vein occlusion in patients with diabetes mellitus: A retrospective cohort study. *Diabetes Res Clin Pract*. 2021;171:108607.
 5. Kolar P. Risk factors for central and branch retinal vein occlusion: a meta-analysis of published clinical data. *J Ophthalmol*. 2014;2014:724780.
 6. Choi JY, Yoo TK. Development of a novel scoring system for glaucoma risk based on demographic and laboratory factors using ChatGPT-4. *Med Biol Eng Comput*. 2024. <https://doi.org/10.1007/s11517-024-03182-0>.
 7. Oh E, Yoo TK, Park E-C. Diabetic retinopathy risk prediction for fundus examination using sparse learning: a cross-sectional study. *BMC Med Inf Decis Mak*. 2013;13:106.
 8. Elkazza SAA. Prognosis prediction in retinal vein occlusion. Thesis. Newcastle University; 2023.
 9. Xing Z, Liu H, Sun Y, Zhang Y, Xing X, Yang K et al. Relationship between retinal volume changes and the prognosis of BRVO-ME treated with Ranibizumab. *Heliyon*. 2024;10.
 10. Korot E, Guan Z, Ferraz D, Wagner SK, Zhang G, Liu X, et al. Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell*. 2021;3:288–98.
 11. Hussain S. Survey on current trends and techniques of data mining research. *Lond J Res Comput Sci Technol*. 2017;17:11.
 12. Shin D, Choi H, Kim D, Park J, Yoo TK, Koh K. Code-Free machine learning approach for EVO-ICL vault prediction: A retrospective Two-Center study. *Translational Vis Sci Technol*. 2024;13:4.
 13. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. 2021;375:n2281.
 14. Xu W, Yan Z, Chen N, Luo Y, Ji Y, Wang M, et al. Development and application of an intelligent diagnosis system for retinal vein occlusion based on deep learning. *Dis Markers*. 2022;2022:4988256.
 15. Fujita A, Hashimoto Y, Okada A, Obata R, Aihara M, Matsui H, et al. Association between proteinuria and retinal vein occlusion in individuals with preserved renal function: a retrospective cohort study. *Acta Ophthalmol*. 2022;100:e1510–7.
 16. Kim JS, Kim M, Kim SW. Prevalence, Survey VII. *Clin Exp Ophthalmol*. 2022;50:2017–8.
 17. Song SJ, Choi KS, Han JC, Jee D, Jeoung JW, Jo YJ, et al. Methodology and rationale for ophthalmic examinations in the seventh and eighth Korea National health and nutrition examination surveys (2017–2021). *Korean J Ophthalmol*. 2021;35:295–303.
 18. Oh TR, Han K-D, Choi HS, Kim CS, Bae EH, Ma SK, et al. Hypertension as a risk factor for retinal vein occlusion in menopausal women. *Med (Baltim)*. 2021;100:e27628.
 19. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, 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.
 20. Yoo TK, Ryu IH, Kim JK, Lee IS, Kim HK. A deep learning approach for detection of shallow anterior chamber depth based on the hidden features of fundus photographs. *Comput Methods Programs Biomed*. 2022;219:106735.
 21. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinović M, et al. Orange: data mining toolbox in Python. *J Mach Learn Res*. 2013;14:2349–53.
 22. Park DJ, Park MW, Lee H, Kim Y-J, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci Rep*. 2021;11:7567.
 23. Yoo TK, Kim SK, Kim DW, Choi JY, Lee WH, Oh E, et al. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Yonsei Med J*. 2013;54:1321–30.
 24. Raghavendran KR, Elragal A. Low-Code machine learning platforms: A fast-lane to digitalization. *Informatics*. 2023;10:50.
 25. Mujahid M, Kina E, Rustam F, Villar MG, Alvarado ES, De La Torre Diez I, et al. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *J Big Data*. 2024;11:87.
 26. Cifci MA, Hussain S. Data mining usage and applications in health services. *JOIV: Int J Inf Visualization*. 2018;2:225–31.
 27. Ren X, Feng W, Ran R, Gao Y, Lin Y, Fu X, et al. Artificial intelligence to distinguish retinal vein occlusion patients using color fundus photographs. *Eye*. 2023;37:2026–32.
 28. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip A-M, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. 2018;125:549–58.
 29. Jaulim A, Ahmed B, Khanam T, Chatziralli IP. BRANCH RETINAL VEIN OCCLUSION: epidemiology, pathogenesis, risk factors, clinical features, diagnosis, and complications. An update of the literature. *RETINA*. 2013;33:901.
 30. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Health*. 2019;1:e232–42.
 31. Kim T, Lee J-S. Maximizing AUC to learn weighted Naive Bayes for imbalanced data classification. *Expert Syst Appl*. 2023;217:119564.
 32. Sen S, Singh KP, Chakraborty P. Dealing with imbalanced regression problem for large dataset using scalable artificial neural network. *New Astron*. 2023;99:101959.
 33. Choi JY, Han E, Yoo TK. Application of ChatGPT-4 to oculosomics: a cost-effective osteoporosis risk assessment to enhance management as a proof-of-principles model in 3PM. *EPMA J*. 2024;15:659–76.
 34. Cugati S, Wang JJ, Rochtchina E, Mitchell P. Ten-Year incidence of retinal vein occlusion in an older population: the blue mountains eye study. *Arch Ophthalmol*. 2006;124:726–32.
 35. Muraoka Y, Tsujikawa A. Arteriovenous crossing associated with branch retinal vein occlusion. *Jpn J Ophthalmol*. 2019;63:353–64.
 36. Ponto KA, Scharer I, Binder H, Korb C, Rosner AK, Ehlers TO, et al. Hypertension and multiple cardiovascular risk factors increase the risk for retinal vein occlusions: results from the Gutenberg retinal vein occlusion study. *J Hypertens*. 2019;37:1372.
 37. Shin DW, Cho J, Park JH, Cho B. National general health screening program in Korea: history, current status, and future direction. *Precision Future Med*. 2022;6:9–31.
 38. Noh SY, Lee JH, Jeong WJ. Branch retinal vein occlusion with arteriovenous crossing. *J Retin*. 2023;8:36–41.
 39. Choi EY, Kim D, Kim J, Kim E, Lee H, Yeo J, et al. Predicting branch retinal vein occlusion development using multimodal deep learning and pre-onset fundus hemisection images. *Sci Rep*. 2025;15:2729.
 40. Kim BR, Yoo TK, Kim HK, Ryu IH, Kim JK, Lee IS, et al. Oculosomics for sarcopenia prediction: a machine learning approach toward predictive, preventive, and personalized medicine. *EPMA J*. 2022;13:367–82.
 41. Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into systemic disease through retinal Imaging-Based oculosomics. *Trans Vis Sci Tech*. 2020;9:6–6.
 42. Li X, Wong WL, Cheung CY, Cheng C-Y, Ikram MK, Li J, et al. Racial differences in retinal vessel geometric characteristics: A multiethnic study in healthy Asians. *Investig Ophthalmol Vis Sci*. 2013;54:3650–6.
 43. Gijsberts CM, den Ruijter HM, Asselbergs FW, Chan MY, de Kleijn DPV, Hofer IE. Biomarkers of coronary artery disease differ between Asians and Caucasians in the general population. *Global Heart*. 2015;10:301–e31111.
 44. Cho B-J, Bae SH, Park SM, Shin MC, Park IW, Kim HK, et al. Comparison of systemic conditions at diagnosis between central retinal vein occlusion and branch retinal vein occlusion. *PLoS ONE*. 2019;14:e0220880.
 45. Khankhoje R. Beyond coding: A comprehensive study of Low-Code, No-Code and traditional automation. *J Artif Intell Cloud Comput*. 2022;1:1–5.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.