JMIR Publications
The Leading eHealth Publisher

# Prediction and Determinants of Visual Impairment among Chinese Middle-aged and Elderly Adults: Machine Learning Algorithms

Lijun Mao, Zhen Yu, Luotao Lin, Manoi Sharma, Hualing Song, Hailei Zhao, Xianglong Xu

Submitted to: JMIR Aging
on: April 23, 2024

**Disclaimer: © The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

# *Table of Contents*



Preprint JMIR Publications

# Prediction and Determinants of Visual Impairment among Chinese Middle-aged and Elderly Adults: Machine Learning Algorithms

Lijun Mao[1]; Zhen Yu[2]; Luotao Lin[3]; Manoi Sharma[4]; Hualing Song[1]; Hailei Zhao[1]; Xianglong Xu[5]

[1]School of Public Health, Shanghai University of Traditional Chinese Medicine Shanghai CN
[2]Monash e-Research Centre, Faculty of Engineering, Airdoc Research, Nvidia AI Technology Research Centre, Monash University Melbourne AU
[3]Nutrition and Dietetics Program, Department of Individual, Family, and Community Education, University of New Mexico Albuquerque US
[4]Department of Social and Behavioral Health, School of Public Health, University of Nevada Las Vegas US

**Corresponding Author:**
Xianglong Xu

## *Abstract*

**Background:** Visual impairment (VI) is a prevalent global health issue, affecting over 2.2 billion people worldwide, with nearly half of the Chinese population aged 60 and above being affected. Early detection of high-risk VI is essential for preventing irreversible vision loss among Chinese middle-aged and elderly adults. While machine learning (ML) algorithms exhibit significant predictive advantages, their application in predicting VI risk among the general middle-aged and elderly population in China remains limited.

**Objective:** We aimed to predict VI and identify its determinants using ML algorithms.

**Methods:** We used 19,047 participants from four waves of the China Health and Retirement Longitudinal Study (CHARLS) that were conducted between 2011 and 2018. To envisage the prevalence of VI, we generated a geographical distribution map. Additionally, we constructed a model using indicators of self-reported questionnaire, physical examination, and blood biomarkers as predictors. Multiple ML algorithms, including gradient boosting machine (GBM), dynamic random forest (DRF), generalised linear model (GLM), deep learning (DL), and stacked ensemble, were used for prediction. We plotted receiver operating characteristics (ROC) and calibration curves to assess the predictive performance. Variable importance analysis was used to identify key predictors.

**Results:** Among 19,047 participants, 33.9% suffered from VI. Qinghai, Chongqing, Anhui, and Sichuan showed the highest VI rates, while Beijing and Xinjiang had the lowest. GLM, GBM, and stacked ensemble achieved acceptable area under curve values of 0.706, 0.710, and 0.715, respectively, with the stacked ensemble performing best. Key predictors included hearing impairment, self-expectation of health status, pain, age, hand grip strength, depression, night sleep duration, haemoglobin, high-density lipoprotein cholesterol, and arthritis or rheumatism.

**Conclusions:** Nearly one-third of middle-aged and elderly adults in China had VI. The prevalence of VI shows regional variations, but no distinct east-west or north-south distribution differences. ML algorithms demonstrate accurate predictive capabilities for VI. The combination of prediction models and variable importance analysis provides valuable insights for the early identification and intervention of VI among Chinese middle-aged and elderly adults.

(JMIR Preprints 23/04/2024:59810)
DOI: https://doi.org/10.2196/preprints.59810

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint
JMIR Publications

# Original Manuscript

Preprint
JMIR Publications

# Prediction and Determinants of Visual Impairment among Chinese Middle-aged and Elderly Adults: Machine Learning Algorithms

**Authors:** Lijun Mao[1], BSc; Zhen Yu[2], PhD; Luotao Lin[3], PhD; Manoi Sharma [4, 5], PhD; Hualing Song[1], MSc;
Hailei Zhao[1*], MD; Xianglong Xu[1, 6, 7*], PhD

**Affiliations:**

[1]School of Public Health, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[2]Monash e-Research Centre, Faculty of Engineering, Airdoc Research, Nvidia AI Technology Research Centre, Monash University, Melbourne, Australia

[3]Nutrition and Dietetics Program, Department of Individual, Family, and Community Education, University of New Mexico

[4] Department of Social and Behavioral Health, School of Public Health, University of Nevada, Las Vegas, USA

[5] Department of Internal Medicine, Kirk Kerkorian School of Medicine, University of Nevada, Las Vegas, USA

[6] School of Translational Medicine, Monash University, Melbourne, VIC, Australia

[7]Artificial Intelligence and Modelling in Epidemiology Program, Melbourne Sexual Health Centre, Alfred Health, Carlton, Victoria, Australia

*Correspondence
Xianglong Xu, PhD
Email: xianglongxu@shutcm.edu.cn
Hailei Zhao, MD
Email: Zhao.hailei@shutcm.edu.cn

## Abstract

**Background:** Visual impairment (VI) is a prevalent global health issue, affecting over 2.2 billion people worldwide, with nearly half of the Chinese population aged 60 and above being affected. Early detection of high-risk VI is essential for preventing irreversible vision loss among Chinese middle-aged and elderly adults. While machine learning (ML) algorithms exhibit significant

predictive advantages, their application in predicting VI risk among the general middle-aged and elderly population in China remains limited.
**Objective:** We aimed to predict VI and identify its determinants using ML algorithms.
**Methods:** We used 19,047 participants from four waves of the China Health and Retirement Longitudinal Study (CHARLS) that were conducted between 2011 and 2018. To envisage the prevalence of VI, we generated a geographical distribution map. Additionally, we constructed a model using indicators of self-reported questionnaire, physical examination, and blood biomarkers as predictors. Multiple ML algorithms, including gradient boosting machine (GBM), dynamic random forest (DRF), generalised linear model (GLM), deep learning (DL), and stacked ensemble, were used for prediction. We plotted receiver operating characteristics (ROC) and calibration curves to assess the predictive performance. Variable importance analysis was used to identify key predictors.
**Results:** Among 19,047 participants, 33.9% suffered from VI. Qinghai, Chongqing, Anhui, and Sichuan showed the highest VI rates, while Beijing and Xinjiang had the lowest. GLM, GBM, and stacked ensemble achieved acceptable area under curve values of 0.706, 0.710, and 0.715, respectively, with the stacked ensemble performing best. Key predictors included hearing impairment, self-expectation of health status, pain, age, hand grip strength, depression, night sleep duration, haemoglobin, high-density lipoprotein cholesterol, and arthritis or rheumatism.
**Conclusions:** Nearly one-third of middle-aged and elderly adults in China had VI. The prevalence of VI shows regional variations, but no distinct east-west or north-south distribution differences. ML algorithms demonstrate accurate predictive capabilities for VI. The combination of prediction models and variable importance analysis provides valuable insights for the early identification and intervention of VI among Chinese middle-aged and elderly adults.
**Keywords:** Visual impairment; China; middle-aged and elderly adults; machine learning; prediction model

Preprint JMIR Publications

## Introduction

Visual impairment (VI) represents a significant global public health challenge. Over the period from 1990 to 2019, the burden index of VI has escalated for individuals aged 50-74 years and those 75 years and older, shifting from 20th and 16th to 19th and 15th positions, respectively [1]. Estimates indicate that 15% of those aged 65 and above experience vision loss, a proportion that climbs to 30% for the 75 and older age group in a California county [2]. The globally observed increase in VI prevalence is primarily attributed to cataracts and uncorrected refractive errors, accounting for 55% of blindness and 77% of VI cases among adults aged 50 and over in 2015 [3]. This trend is further exacerbated by population growth and ageing [3]. According to the National Bureau of Statistics of China's January 2022 report, adults aged 60 and over accounted for 18.9% of the total population by the end of 2021. Meanwhile, adults aged 65 and over exceeded 200 million, representing 14.2% of the total population [4]. It is projected that during the "14th Five-Year Plan" period, the total number of adults aged 60 and over will surpass 300 million, accounting for over 20% of the population, indicating a transition into the moderate ageing phase. By around 2035, China is anticipated to enter the severe ageing phase [4]. Older adults with VI are at heightened risk of falls [5], potentially leading to fractures and severe outcomes such as cerebral haemorrhage. Furthermore, VI can hinder social engagement among the elderly, possibly giving rise to more profound mental health issues, including depression and anxiety [6]. As the population ages, the prevalence of VI is expected to dramatically increase. However, it is estimated that half of all VI cases are preventable or treatable [7]. Hence, bolstering screening efforts and enhancing risk prediction for VI are paramount in stemming the tide of this growing concern.

In response to the crises posed by VI and to improve public visual health, the China National Health Commission has released the 14th Five-Year National Eye Health Plan (2021-2025). The plan focuses on enhancing eye health information platforms and promoting the harmonious integration of big data, artificial intelligence, and ophthalmology services to advance the early detection of eye diseases [8]. Through the development of machine learning (ML) prediction models for VI, the precise determination of VI risk and identification of influencing risk factors can be achieved. ML could offer new insights for early detection and timely intervention of retinopathy, as well as for the integrated management of ocular health in older adults, ultimately enhancing the overall eye health status of the population.

Artificial intelligence has experienced swift progress in recent years, resulting in extensive use of diverse ML algorithms in clinical research [9]. Compared to traditional statistical methods, ML algorithms can handle more complex nonlinear relationships, interactions, and multiple covariances, significantly improving the predictive ability of artificial intelligence models [10, 11].

Despite the advantages of ML algorithms, there is an absence of the use of ML algorithms to predict the risk of VI in the general middle-aged and elderly population in China. Previous studies on predicting VI focused on various topics. From examining trends in the incidence of VI among populations [12, 13] to assessing the risk of VI in specific groups [14-16] and further predicting the risk of developing particular types of VI [17]. However, these studies have not focused on predicting the individual risk of VI among the general public. In current research on predicting individual risk of VI among the public, three studies [18-20] based on traditional statistical methods and two studies [21, 22] based on ML algorithms have achieved good predictive performance. Among these, two studies [19, 21] are based on foreign populations, one focuses on Chinese children [22], one is a single-centre study [18], and one [19] has a small sample size of 133 participants. There has not been an ML algorithms prediction study of VI covering the characteristics of Chinese middle-aged and elderly adults. Therefore, to determine the risk of VI among middle-aged and older adults in China, we utilised a vast amount of sample information from public databases. We constructed our models by incorporating self-reported questionnaire, physical examination, and blood biomarkers data. Using various ML algorithms, we comprehensively evaluated the impact of different predictive variables on model performance. Our objective is to develop an individual risk prediction model for VI, which could be used to assess the risk of VI among China's general middle-aged and older population. Additionally, to identify risk factors that contribute to the occurrence of VI. The research findings could be used to provide personalised intervention guidance for healthcare professionals, aiming to reduce and delay the onset of retinal diseases among middle-aged and older adults.

## Methods

### Analytic sample

The data utilised in our study originates from the China Health and Retirement Longitudinal Study (CHARLS), a longitudinal survey that represents a nationally diverse cohort of Chinese adults aged 45 and above. This survey strives to establish a comprehensive public database documenting Chinese adults' social, economic, and health statuses, thereby bolstering scientific investigations conducted by the National Development Institute of Peking University. The CHARLS project executed a nationwide baseline survey between 2011 and 2012, with subsequent follow-up visits occurring biennially [23]. The CHARLS baseline survey encompassed 450 villages and neighbourhoods spread across 150 counties in China. The sampling process encompassed multiple levels, including counties, villages, households, and individuals, culminating in interviews with 10,257 households, broadly reflecting the general Chinese middle-aged and elderly populace. We used 25,538 observations from four waves of surveys between 2011 and 2018. Excluding age below 45 years (n=4,079) and missing information on self-reported vision conditions (n=2457), a total of 19,047 participants were included in this study.

### Predictors of VI

Drawing on existing literature and expert insights, 42 predictors were used for ML algorithms training. The predictors were categorised as follows: self-reported questionnaire, physical examination, and blood biomarkers. The self-reported questionnaire included (1) demographic factors (gender, age, region); [24] (2) lifestyle (night sleep duration, smoking, drinking); (3) health status factors (pain, weight change, health status during childhood, self-expectations of health status ); [25] (4) disease factors (depression, hearing impairment, hypertension, dyslipidaemia, diabetes, liver disease, heart disease, stroke, kidney disease, stomach or other digestive disease, memory-related disease, arthritis or rheumatism, menopause, prostatic diseases); (5) living environment factors (house structure, heating energy, cooking energy, room temperature); (6) socioeconomic factors (standard of living, education level). Measurement parameters included (1) physical examination data ([26] hand grip strength, waist, body mass index (BMI)) and (2) blood biomarker data (white blood cell, platelets, glycated haemoglobin, haemoglobin, glucose, total cholesterol, triglycerides, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol). The characteristics of the predictor distribution in the study are shown in (Multimedia Appendix 1).

### Measurement of VI

VI in our study was assessed using the following questions from the CHARLS questionnaires: (1) "How well do you see things in the distance? For example, can you recognise a friend across the road (even with glasses on)? Is it excellent, very good, good, fair, or bad?" and (2) "How well do you see things close up? For example, can you read a newspaper with your glasses on? Is it excellent, very good, good, fair, or bad?" Respondents who answered "not good" to any of the questions were categorised as having a visual impairment, while those who answered "excellent" to "fair" were considered to have no VI.

### Statistical analysis

We used R 4.3.1 for statistical analysis and model development. The summary of continuous variables involved the use of the median and interquartile range (25th and 75th percentiles), and categorical variables were summarised by providing the count (n) and proportion (%) for each category. We used the R H2O package to construct various ML predictive models for a dichotomous outcome of VI. As per the No Free Lunch Theorem [27], no algorithm can outperform a linear enumeration of the search space or a purely random search algorithm. Thus, we split the dataset into training (n=14,286) and testing (n=4761) datasets at a 75:25 ratio. The training dataset was used to develop various models, including a generalised linear model (GLM), gradient boosting machine (GBM), dynamic random forest (DRF), deep learning (DL), and stacked ensemble. In this study, the ratio of positive to negative outcomes in the target variable was 2:1, indicating an imbalanced dataset. To address this imbalance, random oversampling of the minority class was initially employed [28]. Furthermore, to mitigate overfitting and enhance model generalisation [29], external 5-fold cross-validation was implemented. However, the model's performance did not show improvement compared to the no-resampling, mixed mode. Therefore, we trained the stacked combinations using the no-resampling and mixed mode, plotted the receiver operating characteristic (ROC) curves, and constructed the confusion matrix. We employed the area under the curve (AUC) to evaluate the best model, with an acceptable AUC of 0.7-0.8, a good AUC of 0.8-0.9, and an excellent AUC of > 0.9 [30]. We calibrated the probabilities predicted by the models to the actual occurrence level in the testing dataset using a logistic function and calculated the Brier score to assess the accuracy of the prediction of VI. The Brier score takes values from 0 to 1, and at a predicted probability of 50%, the Brier score is 0.25 [31]. A model score between 0 and 0.25 indicates correct prediction, and a score closer to 0 indicates better model effectiveness. Additionally, We utilised the GBM model for a variable importance analysis. GBM is a powerful ensemble learning algorithm that iteratively builds multiple decision trees to enhance the model’s predictive performance [32]. Using this framework, we were able to quantitatively assess the contribution of each feature towards model predictions, thereby allowing us to evaluate and compare the significance of various features.

## Results

### Geographical distribution of VI

Figure 1 presents the prevalence of VI by province in China, based on data from the four waves of the China Health and Retirement Longitudinal Study (CHARLS) conducted between 2011 and 2018. Qinghai, Chongqing, Anhui, and Sichuan provinces reported a high prevalence of VI, with rates exceeding 40% (45.0%, 44.1%, 41.9%, and 40.3%, respectively). In contrast, Xinjiang and Beijing had a low prevalence of VI, with rates below 20% (19.8% and 13.8%, respectively). The remaining provinces, municipalities, and autonomous regions exhibited a moderate prevalence of VI, ranging from 20% to 40%.

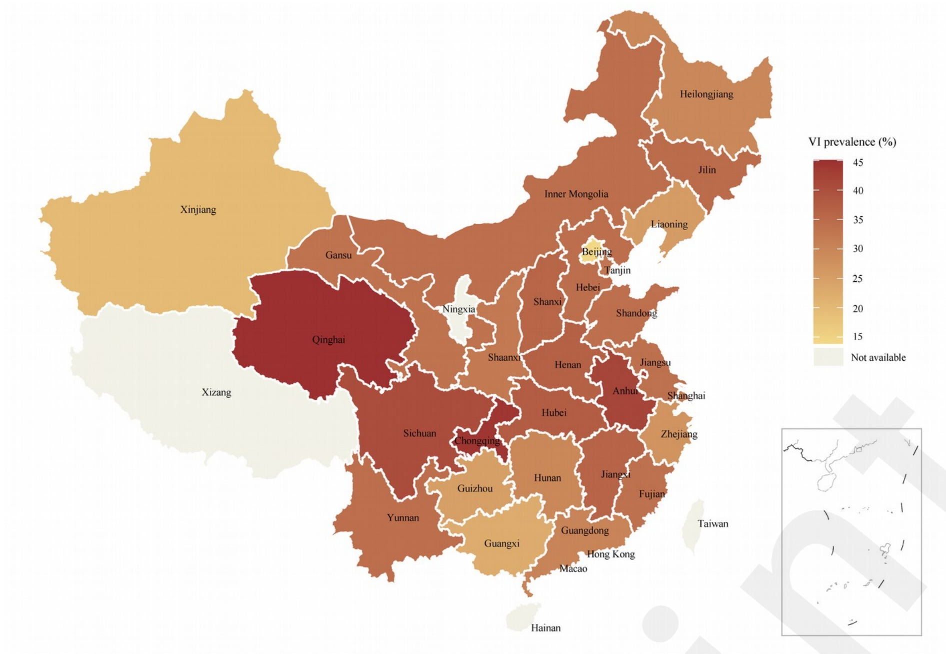



Figure 1. The prevalence of VI by province in China from CHARLS (2011-2018) four waves.

## Characteristics of the study participants

A total of 33.9% (6,449/19,047) participants reported VI, with the training dataset (n=4,837) and the testing dataset (n=1,612). Of the 6,449 cases of VI, 58.8% were male, and 41.2% were female. The age group with the highest prevalence of VI (39.1%) was 55-65 years, followed by those aged ≥65 (31.4%) and 45-55 (29.5%). The selected characteristics of study participants are shown in Table 1. The full characteristics of the predictor distribution in the study are shown in Appendix Table 1 (Multimedia Appendix 1).

Table 1. Selected characteristics of study participants of Chinese adults aged more than 45yrs as drawn from the CHARLS 2011–2018 (n = 19,047).

| Characteristic | Overall, N = 19,047[a] | Non-VI, N = 12,598[a] | VI, N = 6,449[a] | p-value[b] |
|---|---|---|---|---|
| **Gender** | | | | <0.001 |
| Female | 9,927(52.1%) | 6,132(48.7%) | 3,795(58.8%) | |
| Male | 9,120(47.9%) | 6,466(51.3%) | 2,654(41.2%) | |
| **Age** | | | | <0.001 |
| 45-55 | 7,279(38.2%) | 5,375(42.7%) | 1,904(29.5%) | |
| 55-65 | 6,794(35.7%) | 4,274(33.9%) | 2,520(39.1%) | |
| ≥65 | 4,974(26.1%) | 2,949(23.4%) | 2,025(31.4%) | |
| **Region** | | | | <0.001 |
| East | 6,995(36.7%) | 4,486(35.6%) | 2,509(38.9%) | |
| Central | 6,315(33.2%) | 4,355(34.6%) | 1,960(30.4%) | |
| West | 5,737(30.1%) | 3,757(29.8%) | 1,980(30.7%) | |
| **Education level** | | | | <0.001 |
| Less than elementary school | 8,368(43.9%) | 4,869(38.6%) | 3,499(54.3%) | |
| Elementary school | 4,113(21.6%) | 2,805(22.3%) | 1,308(20.3%) | |
| Middle school | 3,971(20.8%) | 2,909(23.1%) | 1,062(16.5%) | |
| High school or above | 2,595(13.6%) | 2,015(16.0%) | 580(9.0%) | |
| **Standard of living** | | | | <0.001 |
| Poor | 2,260(11.9%) | 1,182(9.4%) | 1,078(16.7%) | |
| Relatively poor | 5,824(30.6%) | 3,768(29.9%) | 2,056(31.9%) | |
| Average | 10,420(54.7%) | 7,257(57.6%) | 3,163(49.0%) | |
| Relatively high | 507(2.7%) | 365(2.9%) | 142(2.2%) | |
| Very high | 36(0.2%) | 26(0.2%) | 10(0.2%) | |

[a]n (%).
[b]Pearson's Chi-squared test.

## VI prediction

We applied the trained models to the testing dataset. The distribution of predictor variables between the testing and training datasets is detailed in Appendix Table 2 (Multimedia Appendix 1). The results indicate that the ensemble model demonstrates superior predictive performance compared to individual ML models. Three algorithms, namely GLM, GBM, and stacked ensemble model (GBM-XGBoost-GLM-DL-DRF), achieved acceptable AUC values, with 0.706, 0.710, and 0.715, respectively. The ensemble model exhibited the best performance. However, the DRF model and DL model did not meet the acceptable AUC threshold of 0.70, both achieving an AUC of 0.698. Figure 2 depicts the ROC curves for all the models. Detailed evaluation metrics for all models on the testing dataset are provided in Appendix Table 3 (Multimedia Appendix 1).

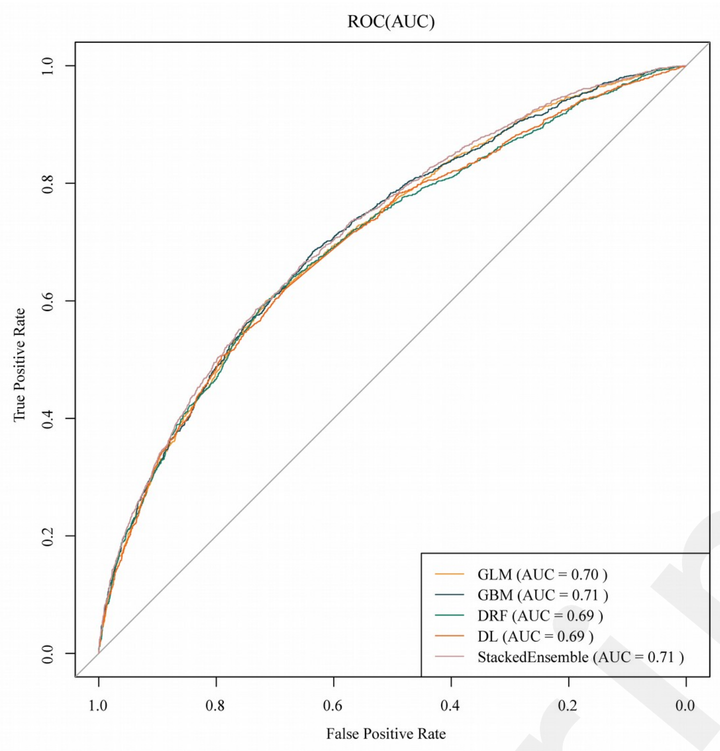



Figure 2. Receiver operating characteristic (ROC) curves of all VI prediction models on the testing dataset.
Note: AUC, area under the curve; GLM, generalised linear model; GBM, gradient boosting machine; DRF, dynamic random forest; DL, deep learning; StackedEnsemble, GBM-XGBoost-GLM-DL-DRF.

## Assessing the efficacy of ML models for VI prediction

Figure 3 presents the calibration curves of all the models on the testing dataset. These curves illustrate the agreement between the predicted probabilities from each model and the observed probabilities of VI in the testing data. The results indicate that all models accurately predicted VI, as evidenced by their Brier scores being less than 0.25.

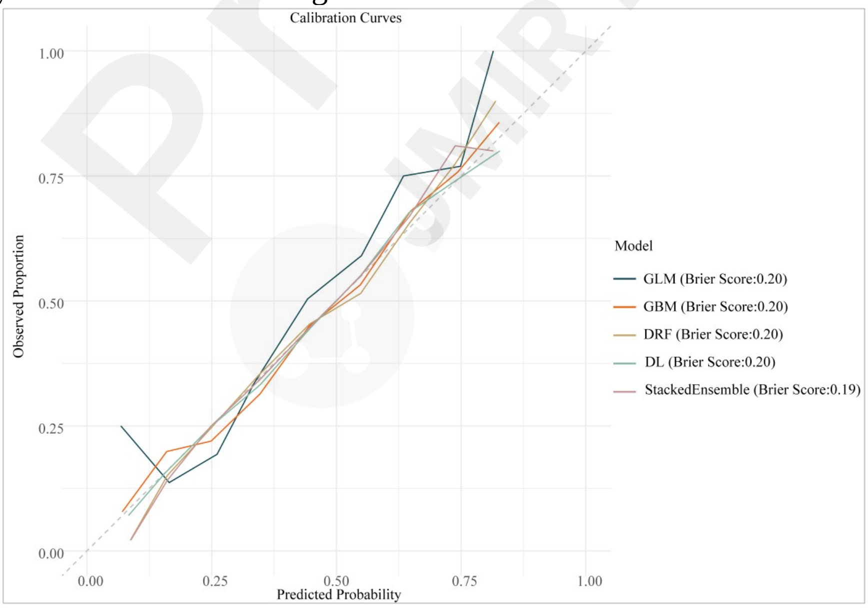



Figure 3. The calibration curves of all VI prediction models on the testing dataset.GLM, generalized linear model; GBM, gradient boosting machine; DRF, dynamic random forest; DL, deep learning; StackedEnsemble, GBM-XGBoost-GLM-DL-DRF.

### Determinants of VI

We plotted the variable importance rankings derived from the GBM model. The top ten important predictors for VI were identified as hearing impairment, self-expectation of health status, pain, age, hand grip strength, depression, night sleep duration, haemoglobin, high-density lipoprotein cholesterol, and arthritis or rheumatism. A comprehensive ranking of the variables' importance is presented in Figure 4.

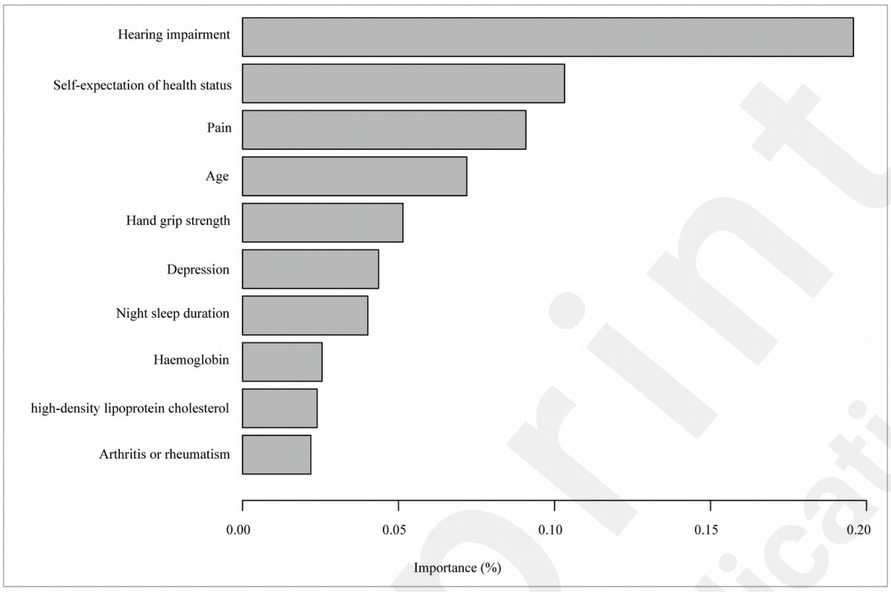



Figure 4. Variable importance analysis performed by gradient boosting machine.

## Discussion

### Principal results

To our knowledge, this study is the first attempt to predict the risk of VI in the general middle-aged and elderly population in China using ML algorithms. The findings indicate that ML algorithms can accurately identify individuals at risk of VI among this demographic. Our results also showed that ensemble algorithms proved superior to individual ML models. Furthermore, we calculated the prevalence of VI and presented it through regional visualisation. The results indicated the existence of regional disparities in the prevalence of VI among Chinese middle-aged and elderly adults, with varying rates across provinces. However, it's worth noting that we did not find any significant north-south or east-west directional differences in prevalence. Based on these results, it is possible to gain a more intuitive understanding of the regional distribution of VI. Additionally, our prediction model can be leveraged to develop a risk assessment tool for early detection of VI in clinical practice. Predictors' importance could help guide personalised early interventions for middle-aged and elderly individuals at risk of VI.

### Comparison with prior work

Our study showed that GBM outperformed the traditional logistic regression model in predicting VI among middle-aged and older Chinese adults. Previous ML studies on VI prediction have predominantly relied on single algorithms without comparing the predictive performance across multiple algorithms. In contrast, our work utilised various ML algorithms for VI prediction and assessed their comparative effectiveness. Furthermore, our findings highlighted the superior performance of ensemble algorithms over individual learning models. Consequently, our research contributes significantly to future investigations of ML algorithms for VI prediction by providing a comprehensive evaluation of multiple algorithms and emphasizing the advantages of ensemble methods. Additionally, our work offers technical guidance for the primary prevention of VI by identifying the most effective predictive models and important risk factors.

Our study has the following strengths compared to previous ML models for VI prediction. Firstly, our model is more adept at identifying potentially modifiable risk factors. Unlike previous studies that primarily relied on image or video data as predictors [21, 22], our approach incorporates easily accessible everyday information, such as lifestyle factors like night sleep duration. Secondly, our predictive model holds greater representativeness and general applicability for the middle-aged and elderly Chinese population. Previous domestic studies on VI prediction tended to focus on specific disease groups (stroke patients) or particular types of VI (anterior retinal visual impairment) [14, 15]. Our research encompasses the general middle-aged and elderly population, and our predicted outcomes pertain to general types of VI.

We found that the best predictors of VI included hearing impairment, self-expectation of health status, pain, age, hand grip strength, depression, night sleep duration, haemoglobin, high-density lipoprotein cholesterol, and arthritis or rheumatism. This finding aligns with those documented in prior research [24, 25, 33-35]. Significantly, our study newly highlighted the crucial role of hearing impairment, self-expectation of health status, and pain in predicting VI. Although hearing impairment does not directly affect vision, its underlying causes may be associated with ocular or neurological disorders. For instance, neurofibromatosis occurring near the inner ear and optic nerve can potentially lead to concurrent hearing and VI [36]. In addition, **s**elf-expectation of health status reveals one's attitudes towards personal health. High expectations often lead to proactive health behaviours, like regular ophthalmologic check-ups for vision issues. In contrast, lower expectations, possibly due to comorbid chronic conditions [25], may cause psychological stress affecting vision [37]. Additionally, pain serves as a vital physiological indicator, revealing certain discomforts or underlying ailments in the body. Prolonged pain can keep an individual in a constant state of stress, disrupting the normal functions of the immune and endocrine systems, thereby indirectly impacting vision. Based on the results of our study, we can enhance our understanding of the mechanisms underlying the occurrence of VI. This understanding allows us to identify high-risk groups for VI early and implement focused interventions tailored explicitly for these individuals. Moreover, our findings provide a valuable reference for selecting variables in constructing VI prediction models. Nevertheless, to ensure the accuracy and credibility of our findings, further studies are required to validate these associations.

### Limitations

This study has several limitations that should be acknowledged. Firstly, the assessment of VI relied on self-reported data, which may be subject to recall bias or subjectivity in responses. The qualitative nature of the VI assessment responses lacks the numerical precision of a quantitative evaluation, potentially affecting the accuracy of the outcome variable. Secondly, while environmental and lifestyle factors were considered, VI is also influenced by genetic factors, which were not included in the model due to the absence of such information in the database. Incorporating genetic data could potentially enhance the model's predictive performance. Finally, the data were sourced from CHARLS, which only represents the Chinese population aged 45 and above. Consequently, the generalizability of the model to other age groups or populations outside of China remains uncertain. Further validation studies are necessary to evaluate the model's effectiveness in diverse populations across different countries and age ranges.

### Conclusions

The prevalence of VI is notably high among middle-aged and elderly Chinese adults, displaying regional disparities but no significant variances between north-south or east-west regions. Our study is the first to use ML algorithms in predicting VI among China's general middle-aged and older population. The findings demonstrate that ML algorithms can accurately predict VI among this demographic. Ensemble algorithms outperform individual learning models in predicting VI. Variable importance analysis highlighted the importance of considering factors such as hearing impairment and individuals' self-expectation of health status when predicting VI risk. By incorporating these predictors, our study facilitates the early identification of individuals at high risk for VI, enabling timely interventions and preventive measures to mitigate the development and progression of VI.

**Acknowledgements**
This document uses data from four waves of CHARLS from 2011 to 2018. We thank the CHARLS research team and each respondent for contributing to this study.

**Ethical approval**
The Peking University Institutional Review Board granted ethical approval for all waves of CHARLS. The IRB approval number for the self-reported questionnaire (including physical examination measurements) was IRB00001052-11015; the IRB approval number for the biomarker collection was IRB00001052-11014.

**Authors' Contributions**
XX and LM conceived and designed the study. LM and XX cleaned the data and built the models and codes. LM wrote the first draft and edited the manuscript. JZ, ZY, FC and LY contributed to data cleaning. HS, HZ, and XX contributed to data validation and supervision. ZX, LL, and MS contributed to data interpretation and manuscript revision. All authors contributed to the preparation of the manuscript and approved the final manuscript.

**Competing interests**
None declared.

**Abbreviations**
**AUC:** area under the curve
**BMI:** body mass index
**CHARLS:** The China Health and Retirement Longitudinal Study
**DL:** deep learning
**DRF:** dynamic random forest
**DCA:** decision curve analysis
**GBM:** gradient boosting machine
**GLM:** generalized linear model
**ML:** machine learning
**ROC:** receiver operating characteristics
**VI:** visual impairment

## Multimedia Appendix 1
Supplementary tables and figures

## References
1. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020 Oct 17;396(10258):1204-22. PMID: 33069326. doi: 10.1016/s0140-6736(20)30925-9.
2. Wallhagen MI, Strawbridge WJ, Shema SJ, Kurata J, Kaplan GA. Comparative impact of hearing and vision impairment

on subsequent functioning. J Am Geriatr Soc. 2001 Aug;49(8):1086-92. PMID: 11555071. doi: 10.1046/j.1532-5415.2001.49213.x.
3. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. Lancet Glob Health. 2017 Dec;5(12):e1221-e34. PMID: 29032195. doi: 10.1016/s2214-109x(17)30393-5.
4. NHC: China made solid progress in elderly care over past decade. Xinhua. 2022. URL:https://english.www.gov.cn/statecouncil/ministries/202209/20/content_WS6329c182c6d0a757729e0446.html [accessed 2023-11-22].
5. Ehrlich JR, Hassan SE, Stagg BC. Prevalence of Falls and Fall-Related Outcomes in Older Adults with Self-Reported Vision Impairment. J Am Geriatr Soc. 2019 Feb;67(2):239-45. PMID: 30421796. doi: 10.1111/jgs.15628.
6. Pardhan S, Smith L, Bourne R, Davis A, Leveziel N, Jacob L, et al. Combined Vision and Hearing Difficulties Results in Higher Levels of Depression and Chronic Anxiety: Data From a Large Sample of Spanish Adults. Front Psychol. 2020;11:627980. PMID: 33536989. doi: 10.3389/fpsyg.2020.627980.
7. Blindness and vision impairment. World Health Organization. 2023. URL:https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment [accessed 2023-11-29].
8. Circular of the National Health Commission on the issuance of the "14th Five-Year Plan" for National Eye Health (2021-2025). Central People's Government of the People's Republic of China. 2022. URL:https://www.gov.cn/zhengce/zhengceku/2022-01/17/content_5668951.htm [accessed 2023-11-28].
9. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018 Dec;284(6):603-19. PMID: 30102808. doi: 10.1111/joim.12822.
10. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak. 2005 Feb 15;5:3. PMID: 15713231. doi: 10.1186/1472-6947-5-3.
11. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina (Kaunas). 2020 Sep 8;56(9). PMID: 32911665. doi: 10.3390/medicina56090455.
12. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health. 2021 Feb;9(2):e130-e43. PMID: 33275950. doi: 10.1016/s2214-109x(20)30425-3.
13. Fricke TR, Jong M, Naidoo KS, Sankaridurg P, Naduvilath TJ, Ho SM, et al. Global prevalence of visual impairment associated with myopic macular degeneration and temporal trends from 2000 through 2050: systematic review, meta-analysis and modelling. Br J Ophthalmol. 2018 Jul;102(7):855-62. PMID: 29699985. doi: 10.1136/bjophthalmol-2017-311266.
14. Hsia Y, Lin YY, Wang BS, Su CY, Lai YH, Hsieh YT. Prediction of Visual Impairment in Epiretinal Membrane and Feature Analysis: A Deep Learning Approach Using Optical Coherence Tomography. Asia Pac J Ophthalmol (Phila). 2023 Jan-Feb 01;12(1):21-8. PMID: 36706331. doi: 10.1097/apo.0000000000000576.
15. Xu J, Wu Z, Nurnberger A, Sabel BA. Interhemispheric Cortical Network Connectivity Reorganization Predicts Vision Impairment in Stroke. Annu Int Conf IEEE Eng Med Biol Soc. 2021 Nov;2021:836-40. PMID: 34891420. doi: 10.1109/embc46164.2021.9630628.
16. Jin HD, Demmler-Harrison GJ, Miller J, Edmond JC, Coats DK, Paysse EA, et al. Cortical Visual Impairment in Congenital Cytomegalovirus Infection. J Pediatr Ophthalmol Strabismus. 2019 May 22;56(3):194-202. PMID: 31116869. doi: 10.3928/01913913-20190311-01.
17. Liu TYA, Ling C, Hahn L, Jones CK, Boon CJ, Singh MS. Prediction of visual impairment in retinitis pigmentosa using deep learning and multimodal fundus images. Br J Ophthalmol. 2023 Oct;107(10):1484-9. PMID: 35896367. doi: 10.1136/bjo-2021-320897.
18. Zhao Y, Wang A. Development and validation of a risk prediction model for visual impairment in older adults. Int J Nurs Sci. 2023 Jul;10(3):383-90. PMID: 37545769. doi: 10.1016/j.ijnss.2023.06.010.
19. Zhao Y, Yu R, Sun C, Fan W, Zou H, Chen X, et al. Nomogram model predicts the risk of visual impairment in diabetic retinopathy: a retrospective study. BMC Ophthalmol. 2022 Dec 8;22(1):478. PMID: 36482340. doi: 10.1186/s12886-022-02710-6.
20. Burkemper B, Torres M, Jiang X, McKean-Cowdin R, Varma R. Factors Associated with Visual Impairment in Chinese American Adults: The Chinese American Eye Study. Ophthalmic Epidemiol. 2019 Oct;26(5):329-35. PMID: 31146615. doi: 10.1080/09286586.2019.1622737.
21. Tham YC, Anees A, Zhang L, Goh JHL, Rim TH, Nusinovici S, et al. Referral for disease-related visual impairment using retinal photograph-based deep learning: a proof-of-concept, model development study. Lancet Digit Health. 2021 Jan;3(1):e29-e40. PMID: 33735066. doi: 10.1016/s2589-7500(20)30271-5.
22. Chen W, Li R, Yu Q, Xu A, Feng Y, Wang R, et al. Early detection of visual impairment in young children using a smartphone-based deep learning system. Nat Med. 2023 Feb;29(2):493-503. PMID: 36702948. doi: 10.1038/s41591-022-02180-9.

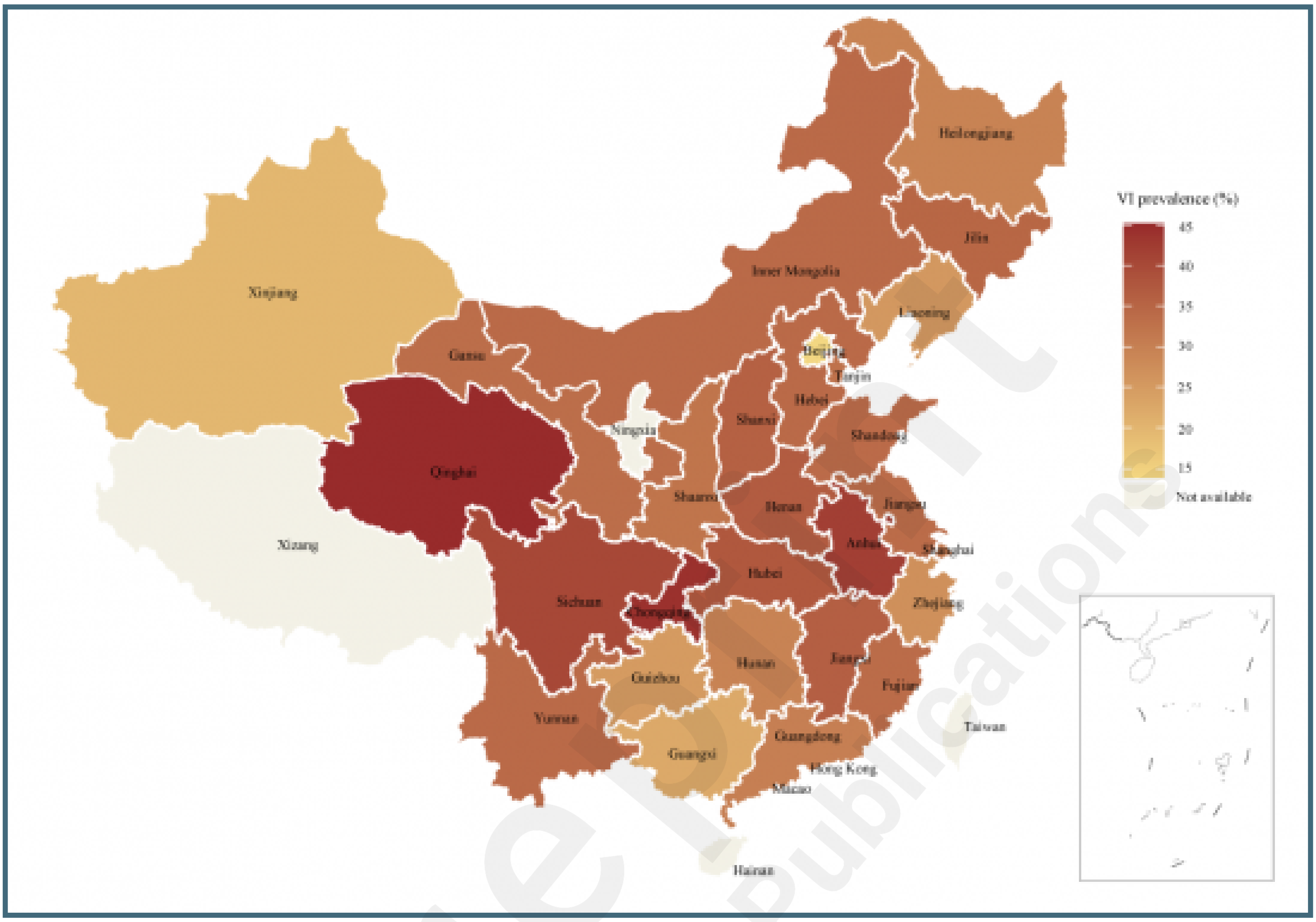
23. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol. 2014 Feb;43(1):61-8. PMID: 23243115. doi: 10.1093/ije/dys203.
24. Sun M, Bo Q, Lu B, Sun X, Zhou M. The Association of Sleep Duration With Vision Impairment in Middle-Aged and Elderly Adults: Evidence From the China Health and Retirement Longitudinal Study. Front Med (Lausanne). 2021;8:778117. PMID: 35004745. doi: 10.3389/fmed.2021.778117.
25. Gu Y, Cheng H, Liu X, Dong X, Congdon N, Ma X. Prevalence of self-reported chronic conditions and poor health among older adults with and without vision impairment in China: a nationally representative cross-sectional survey. BMJ Open Ophthalmol. 2023 Mar;8(1). PMID: 37278435. doi: 10.1136/bmjophth-2022-001211.
26. Shang X, Wu G, Wang W, Zhu Z, Zhang X, Huang Y, et al. Associations of vision impairment and eye diseases with frailty in community-dwelling older adults: a nationwide longitudinal study in China. Br J Ophthalmol. 2022 Dec 19. PMID: 36535748. doi: 10.1136/bjo-2022-322048.
27. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE transactions on evolutionary computation. 1997;1(1):67-82.
28. Longadge R, Dongre S. Class imbalance problem in data mining review. arXiv preprint arXiv:13051707. 2013.
29. Atabaki-Pasdar N, Ohlsson M, Viñuela A, Frau F, Pomares-Millan H, Haid M, et al. Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts. PLoS Med. 2020 Jun;17(6):e1003149. PMID: 32559194. doi: 10.1371/journal.pmed.1003149.
30. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol. 2010 Sep;5(9):1315-6. PMID: 20736804. doi: 10.1097/JTO.0b013e3181ec173d.
31. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010 Jan;21(1):128-38. PMID: 20010215. doi: 10.1097/EDE.0b013e3181c30fb2.
32. Ke G, Qi Meng LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems 2017.
33. Wong PW, Lau JK, Choy BN, Shih KC, Ng AL, Wong IY, et al. Sociodemographic, behavioral, and medical risk factors associated with visual impairment among older adults: a community-based pilot survey in Southern District of Hong Kong. BMC Ophthalmol. 2020 Sep 18;20(1):372. PMID: 32948134. doi: 10.1186/s12886-020-01644-1.
34. Shang X, Wu G, Wang W, Zhu Z, Zhang X, Huang Y, et al. Associations of vision impairment and eye diseases with frailty in community-dwelling older adults: a nationwide longitudinal study in China. Br J Ophthalmol. 2024 Jan 29;108(2):310-6. PMID: 36535748. doi: 10.1136/bjo-2022-322048.
35. Sun R, Huang D, Liu Z, Zhu T, Gu Z, Ma G, et al. Prevalence, Causes, and Risk Factors of Presenting Visual Impairment and Presenting Blindness in Adults Presenting to an Examination Center in Suzhou, China. J Ophthalmol. 2022;2022:2885738. PMID: 36583116. doi: 10.1155/2022/2885738.
36. Alshoabi SA. Neurofibromatosis Type-2 presenting with vision impairment. Pak J Med Sci. 2023 Mar-Apr;39(2):611-5. PMID: 36950434. doi: 10.12669/pjms.39.2.6813.
37. Zhang Q, Cao GY, Yao SS, Wang C, Chen ZS, Hu YH, et al. Self-reported vision impairment, vision correction, and depressive symptoms among middle-aged and older Chinese: Findings from the China health and retirement longitudinal study. Int J Geriatr Psychiatry. 2021 Jan;36(1):86-95. PMID: 32783270. doi: 10.1002/gps.5398.

# Supplementary Files

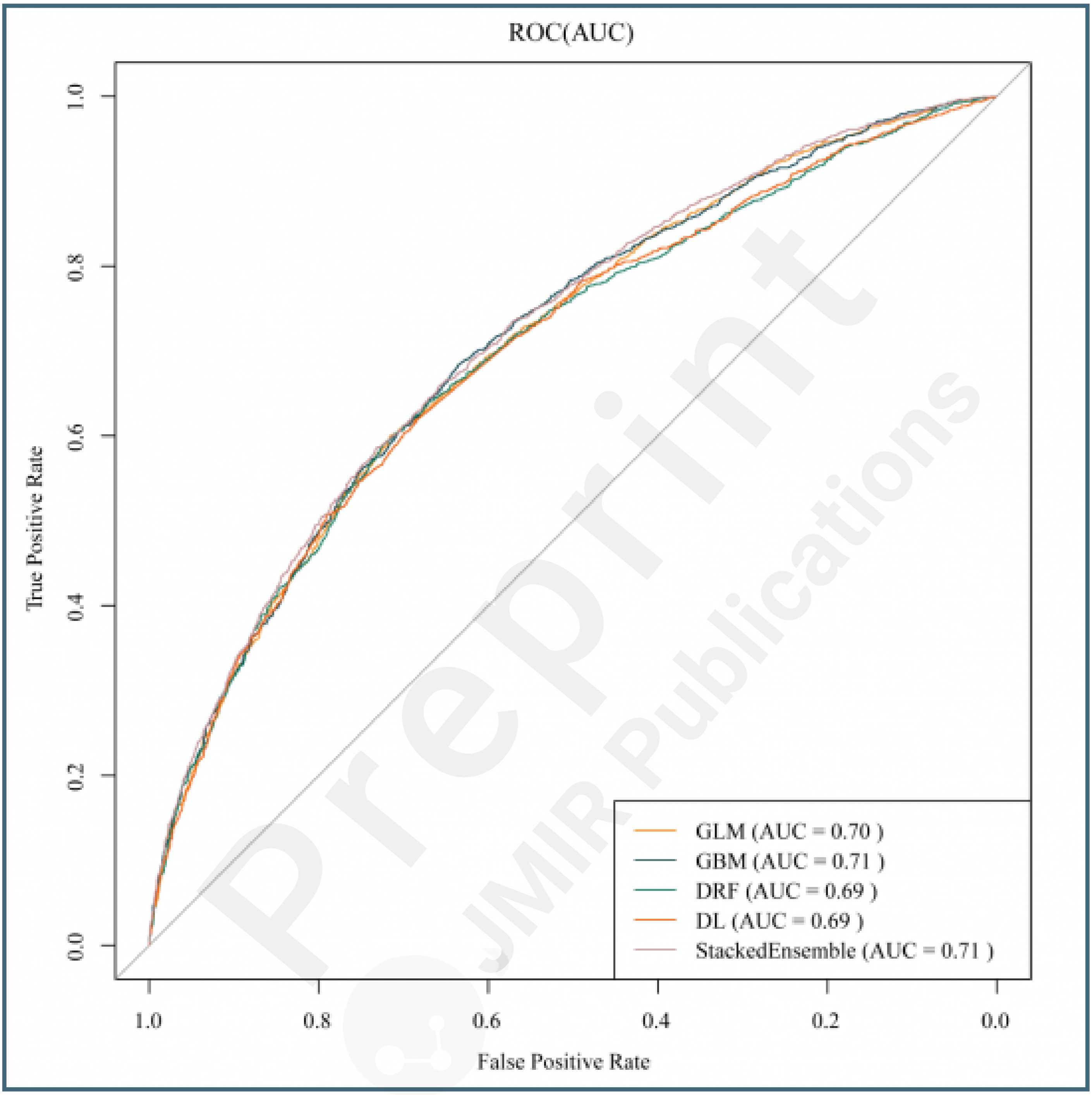
Preprint
JMIR Publications

# Figures

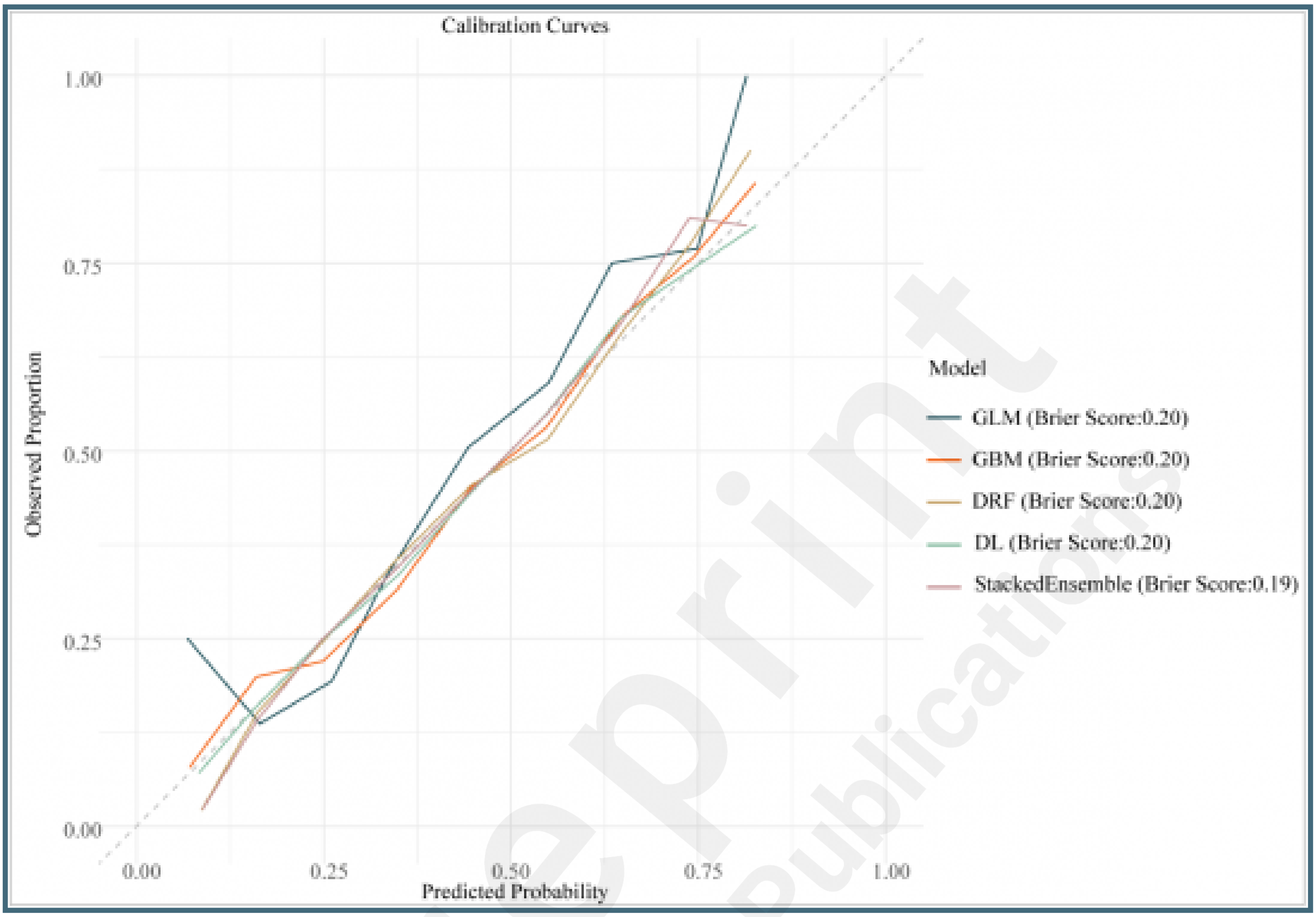
The prevalence of VI by province in China from CHARLS (2011-2018) four waves.



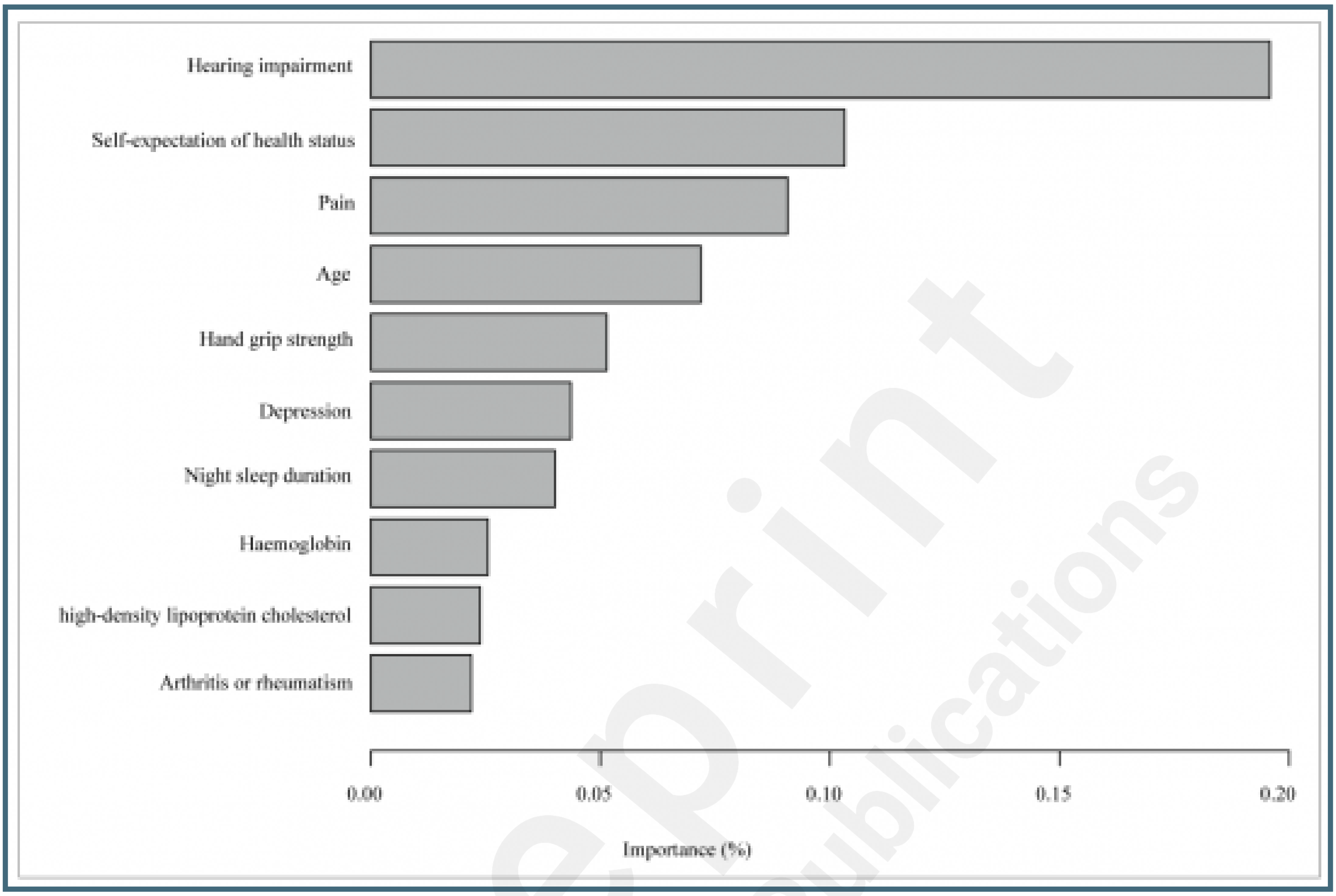
Receiver operating characteristic (ROC) curves of all VI prediction models on the testing dataset.



The calibration curves of all VI prediction models on the testing dataset.



Variable importance analysis performed by gradient boosting machine.



# Multimedia Appendixes

Preprint
JMIR Publications

Untitled.
URL: http://asset.jmir.pub/assets/b1fc90ccf03e809e087a8acc6eb2d567.docx

Preprint JMIR Publications

# TOC/Feature image for homepages

Untitled.