Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Application of Machine Learning Approaches to Identify Factors Associated with Myopia Among Preschool Children

ZhiLong Luo
Guangdong Medical University
Donghao Song
Guangdong Medical University
Mingming Cui
Capital Institute of Pediatrics
Yibo Niu
Guangdong Medical University
Guanghui Sui
Children's Health Department of Maternal and Child Health Hospital
Hao Zhang
Capital Medical University
Zhongxin Zhu
Guangdong Medical University
Tao Yin
Capital Institute of Pediatrics
Yuting Gao
gaoyuting@gdmu.edu.cn

Guangdong Medical University

Article

Keywords: Preschool children, Myopia, Associated factors, Early prevention, Machine Learning

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10551356/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Application of Machine Learning Approaches to Identify Factors Associated with Myopia Among Preschool Children

Zhilong Luo[a,1], Donghao Song[a,1], Mingming Cui[b,1*],Yibo Niu[a], Guanghui Sui[c], Hao Zhang[d], Zhongxin Zhu[a], Tao Yin[b*], Yuting Gao[a*]

a. Department of Environmental and Occupational Health, Dongguan Key Laboratory of Environmental Medicine, School of Public Health, Guangdong Medical University, Dongguan, 523808, China

b. Capital Center for Children's Health, Capital Medical University, Capital Institute of Pediatrics, Beijing, 100020, China

c. Children's Health Department of Maternal and Child Health Hospital, Huaibei, 235000, China

d. Department of ophthalmology,Capital Center for Children's Health, Capital Medical University, Beijing 100020, China

Zhilong Luo: ymenend@163.com

Mingming Cui: bjcmm@126.com

Donghao Song: 2385754698@qq.com

Yibo Niu: 3613843375@qq.com

Guanghui Sui: guanghuisui@126.com

Hao Zhang: 40785840@qq.com

Zhongxin Zhu: zhuzhongxin@gdmu.edu.cn

Tao Yin: gouweicao2911@163.com

Yuting Gao: gaoyuting@gdmu.edu.cn

* Corresponding author

1 Zhilong Luo , Donghao Song and Mingming Cui contributed equally to this study.

Email address:bjcmm@126.com(M. Cui), gouweicao2911@163.com(T. Yin)

, gaoyuting@gdmu.edu.cn(Y. Gao). Prof. Yuting Gao is the primary corresponding author.

Abstract

Objective Myopia has become a major public health concern worldwide, with an increasing trend toward earlier onset. However, factors associated with myopia among preschool children remain insufficiently characterized. This study aimed to identify potential factors associated with preschool myopia using machine learning approaches.

Methods A school-based cross-sectional study was conducted among 1,935 children aged 3 – 7 years in Huaibei, China. Information on demographic characteristics, early life factors, family history, and behavioral exposures was collected through caregiver questionnaires after routine vision screening. Seven machine learning algorithms were applied to identify important factors associated with myopia.

Results The prevalence of caregiver-reported myopia was 2.33%. Among the evaluated models, the Super Learner model demonstrated the best overall performance, with an AUC of 0.742. Variable importance analysis identified parental myopia, close distance television viewing, birth weight, breastfeeding duration, and paternal education as important factors associated with preschool myopia. These findings suggested that early childhood myopia was influenced by multiple domains, including familial, early-life, and behavioral factors.

Conclusions Machine learning approaches provide a useful framework for integrating multidimensional information and identifying potential factors associated with preschool myopia. These findings may support targeted associated assessment and prevention strategies in preschool vision care.

Keywords: Preschool children, Myopia, Associated factors, Early prevention, Machine Learning

1. Introduction

Myopia, a condition of impaired distance vision, has escalated into a major global public health challenge as over half the world's population will be affected by 2050(Flitcroft D. I., et al.,2019; World Health Organization.2019),. In recent decades, the prevalence of myopia has risen sharply, particularly among children, with an apparent shift toward younger ages(Rudnick A. R., et al.2016). This not only threatens immediate visual health but also increases the lifetime risk of sight-threatening complications such as myopic macular degeneration and glaucoma(Liang C. L., et al. 2004; Saw S. M., et al.,2005). In China, the myopia rate among children and adolescents reached 51.9% in 2022(National Bureau For Disease Control Aan Prevention P S R O C.,2024), underscoring an urgent need for effective prevention and control measures(Zhong P.L., et al.,2023; Chen J., et al., 2021).

Myopia is a multifactorial condition involving complex interactions among genetic, early life, environmental, and behavioral factors(Lanca C., Saw S. M.,2020; Liu X. N., et al. 2023; Chakraborty R., et al. 2018; Low W., et al., 2010; He X., et al., 2022; Lin Z., et al.,2015). However, most previous studies have focused on school aged children (≥7 years), resulting in limited understanding of factors associated with myopia during the preschool period(Ma Y., et al.,2018). However, the increasing occurrence of myopia at younger ages has raised particular concern regarding the preschool period, which represents a critical stage of ocular development before formal schooling, early identification of associated factors and timely implementation of preventive measures may help reduce the risk of myopia in later childhood and adulthood(Wolffsohn J. S., et al.,2019).

Conventional epidemiological approaches, such as logistic regression, have substantially contributed to identifying individual factors associated with myopia. However, these methods generally rely on predefined assumptions regarding variable relationships and may have limited ability to capture nonlinear associations and complex interactions among multiple risk factors(Liu L, et al.,2021). Given the multifactorial nature of myopia development, approaches capable of integrating diverse sources of information and identifying

complex patterns may provide additional insights into potential associated factors.

However, preschool children were not included, and whether similar patterns of associated factors exist during early childhood remains insufficiently understood. Furthermore, existing ML based studies have incorporated ophthalmic parameters such as refractive measurements, which can provide strong predictive information but may not be routinely available in large scale preschool screening settings. Therefore, the applicability of ML approaches for identifying myopia associated factors using readily obtainable demographic, familial, early-life, and behavioral information among preschool children remains to be explored.

Machine learning (ML) approaches have emerged as promising tools(Jiang F., et al.,2017; Cao K., et al.,2022; Wu J. H., et al.,2021) for analyzing complex biomedical data by enabling flexible modeling of multidimensional variables and uncovering patterns that may not be readily detected using conventional statistical methods. Previous ML studies on childhood myopia have demonstrated the potential of these approaches for identifying important characteristics and improving classification performance. For example, Tong et al. developed ML models to investigate factors associated with myopia across different school stages and demonstrated age related differences in the contribution of genetic and behavioral factors(Tong H. J., et al.,2023). Similarly, Mu et al. developed several ML models for predicting myopia among children and adolescents, achieving favorable predictive performance(Mu J., et al.,2024)[22]. However, these studies mainly focused on school aged populations, and preschool children, whose visual development and environmental exposures may differ substantially, were not adequately represented. Furthermore, some existing ML based models have incorporated ophthalmic parameters, such as spherical equivalent refraction, axial length, and other biometric indicators, which may enhance predictive accuracy but limit their applicability in large scale preschool screening settings where such measurements are not routinely available(NNunes A. F. et al., 2024; Ying B., et al.,2024).

Therefore, this study aimed to investigate factors associated with caregiver-reported myopia among preschool children and to apply machine learning approaches to identify potentially important factors from demographic, familial, early life, and behavioral domains. Using data from Huaibei City, as one of the pilot regions for myopia prevention and control in Anhui Province, has accumulated valuable experience in combating childhood myopia(OFFICE H M P S G.,,2025). By integrating machine learning methods with a kindergarten based population, this study sought to provide further insights into the multifactorial nature of early childhood myopia and contribute to the identification of children who may benefit from targeted monitoring and preventive strategies.

2. Material and methods

2.1. Study Design and Participants

This study was designed as a school based cross sectional survey. A stratified cluster random sampling method was used to recruit children aged 3-7 years from kindergartens in the Huaibei region. The parents or

legal guardians of the selected children were invited to complete a structured questionnaire. Children with a history of strabismus, amblyopia, congenital eye diseases, allergic conditions, or previous eye surgery were excluded from the study. Additionally, children whose parents declined participation were not included. A total of 1935 valid questionnaires were collected and analyzed. This study was approved by the Ethics Committee of the Capital Institute of Pediatrics (No. SHERLL2023017) and conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from the parents or guardians of all participants prior to data collection.

2.2. Methodology and Diagnostic Criteria

A cross-sectional study was conducted among preschool children undergoing routine health examinations. During the semiannual health check-ups, all participants received standardized vision screening as part of routine pediatric care. Children with abnormal screening results were referred to specialized eye care institutions for further ophthalmic evaluation, including cycloplegic refraction performed by qualified ophthalmologists or optometrists. Children without abnormal screening findings were not routinely referred for cycloplegic examination, as additional clinical assessment was not considered necessary within the screening program. Parents completed the questionnaire during the screening period and reported their child's myopia status and potential associated factors, including demographic characteristics, early-life factors, and visual behaviors. For children who were referred for further ophthalmic evaluation, myopia status was reported based on the diagnosis obtained from the referral examination. This approach reflects the real-world screening and referral pathway, in which clinically identified cases were recorded through caregiver reporting. Myopia was defined as a caregiver-reported diagnosis of myopia, generally corresponding to a spherical equivalent refraction of < -0.50 D based on cycloplegic refraction criteria, consistent with established epidemiological definitions and previous studies(Flitcroft D. I., et al.,2019; Liu S. M. et al.,2026). Clinical examination data were not directly collected or used for analysis. Questionnaire-based assessment of myopia status and associated factors has been widely applied in population-based epidemiological studies(Tian L., et al., 2025; Xie X., et al.,2025).

2.3. Quality Control

The questionnaire was reviewed by experts to ensure accuracy and clarity. A standardized survey tool was used for the on site data collection, completed by parents and submitted to medical staff. During data collection, strict supervision ensured the accurate distribution and retrieval of questionnaires. Following the collection, data cleaning was performed to identify and correct errors or anomalies. Additionally, all data entry was double checked by an independent data entry clerk to minimize the risk of transcription errors.

2.4. Statistical Analysis

All statistical analyses were performed using R version 4.5.0 (R Foundation for Statistical Computing, Vienna, Austria). Normally distributed data are presented as mean ± standard deviation, while non-normally

distributed variables are expressed as median (Q1,Q3) and analyzed with nonparametric tests (Mann-Whitney U test for two-group comparisons). Categorical variables presented as counts and percentages, and group differences were evaluated using the chi square test or Fisher exact test, as appropriate. Binary logistic regression was used to examine associations between potential associated factors and myopia, with associations reported as odds ratios (OR) with 95% confidence intervals (95%*CI*). A two tailed *P*-value < 0.05 was considered statistically significant.

2.5 *Machine Learning Models*

Seven distinct machine learning algorithms were implemented and evaluated for the prediction of myopia status: a Generalized Linear Model (GLM), Light Gradient Boosting Machine (LGB), a single layer Neural Network (NNET), Random Forest (RF), eXtreme Gradient Boosting (XGB), a Bayesian Regularized Neural Network (BRNN), and a Super Learner (SL) ensemble model. All analyses were conducted in R version 4.5.0.

2.5.1 Model Development and Training Pipeline

A standardized preprocessing and feature selection pipeline was implemented to ensure robust model development. The first step addressed the significant class imbalance in the dataset. To mitigate model bias toward the majority class, the Random Over-Sampling Examples (ROSE) was applied exclusively to the training set. ROSE generates synthetic samples for the minority class through interpolation between existing instances, balancing the class distribution without simple duplication and providing a fair foundation for classifier training. Secondly, A hybrid feature selection strategy was employed to identify the most predictive and non-redundant variables, reducing dimensionality and enhancing model interpretability. This strategy combined two complementary methods:

1.Recursive Feature Elimination (RFE) based on XGBoost Importance: An XGBoost model was trained to rank features by importance scores. RFE was then used to iteratively remove the least important features, identifying an optimal subset that preserves predictive power.

2.Feature Selection via LASSO Regression: Simultaneously, a LASSO regression model was fitted. Due to its L1 penalty, LASSO performs feature selection by shrinking coefficients of less informative variables to zero.

The final feature set was determined by integrating results from both methods. Features highly ranked by both XGBoost-RFE and LASSO were prioritized, while variables showing exceptional prominence in either method (e.g., top RFE rank or large LASSO coefficient) were also retained. This approach combines the strengths of tree-based importance measures and linear sparsity constraints, ensuring a robust and informative feature subset.

Using the processed training data (balanced via ROSE and comprising the selected features), hyperparameter tuning for each machine learning algorithm was conducted. A 5-fold cross-validation grid

search was performed to systematically optimize model parameters, maximizing classification performance while minimizing the risk of overfitting.

2.5.2 Super Learner Ensemble Construction

A Super Learner ensemble was developed to integrate the predictions from multiple base algorithms, with the aim of obtaining a more balanced and robust classification model. The Super Learner was constructed as a stacked ensemble, integrating predictions from three base learners: Random Forest (RF), eXtreme Gradient Boosting (XGB), and Generalized Linear Model (GLM). This ensemble method differs from individual model training by combining the outputs of the base learners to improve overall performance. These were implemented using the SuperLearnerR package. The meta-learner was specified as a non-negative least squares (NNLS) regression model, which assigns optimal weights to the base learner predictions by minimizing the prediction error. The combination of these three base learners was selected based on model performance evaluation, as it provided a balanced ensemble incorporating both linear and nonlinear learning approaches.

The training procedure utilized 5-fold cross-validation to generate unbiased inputs for the meta-learner and prevent overfitting. Specifically, each base learner was fit on the ROSE-processed training folds, and their corresponding out-of-fold predictions were collected. These cross-validated predictions were then stacked to form a new feature set, which was used to train the NNLS meta-learner. The final Super Learner model produces predictions by first obtaining predictions from each base learner (fit on the full training set) and then calculating their weighted average, with the weights derived from the trained meta-learner.

2.5.3 Model Evaluation

All final models, the six individually tuned models and the Super Learner ensemble were trained on the entire processed training set and evaluated on a strictly held-out independent test set. Performance was assessed using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). For analyses across the five multiply imputed datasets, performance metrics are reported as averages.

2.5.4 Variable Importance Analysis

To enhance model interpretability, variable importance was assessed within the Super Learner framework. Variable importance was evaluated to identify influential factors contributing to model classification. Variable importance measures were extracted from individual base learners using model-specific metrics, including absolute coefficients for GLM, mean decrease in accuracy for RF, and gain for XGBoost.

These importance values were normalized to ensure comparability and subsequently weighted

according to the corresponding Super Learner coefficients. The aggregated importance scores were then averaged across the five imputed datasets, with standard deviations calculated to reflect variability.

3. Results

3.1. Baseline characteristics of the study population

A total of 1,935 preschool children (982 boys and 953 girls) were included in the final analysis. The children ranged from 3-7 years, with a mean age of 3.99 ± 1.03 years, and a nearly equal sex distribution (50.75% male, 49.25% female). Based on refractive status, participants were categorized into myopic (n = 45, 2.33%) and non-myopic (n = 1890, 97.67%) groups. No significant differences were observed between groups in terms of sex ($p = 0.514$). However, a significant difference was observed in age distribution ($p = 0.043$). Additionally, having both parents with myopia significantly increased the risk of myopia in children($p = 0.026$) . Clinical and demographic characteristics are detailed in Table 1.

3.2. Univariate analysis of myopia associated factors

As shown in Table 2, comparative analysis between the myopia and non-myopia groups revealed a significant association between maternal age at childbirth and myopia prevalence in preschool children ($P = 0.029$).Additionally, birth weight showed a significant statistical difference between the groups, with myopia prevalence higher in children with abnormal birth weights ($p = 0.039$).

As presented in Table 3, a significant association was found between the frequency of TV watching at a distance of less than 3 meters and myopia. The myopia rate was 1.49% among children who never watched TV at this distance, compared to 6.49% among those who always did ($p < 0.001$). Similarly, a significantly association was observed between computer use at a distance of less than 50 cm and myopia ($p = 0.024$), with rates of 1.80% (never), 2.36% (occasional), 4.09% (often), and 9.38% (always). However, no significant association was found between outdoor activity duration and myopia ($p = 0.642$).

3.3. Multiple logistic regression analysis of related factors of myopia

Multiple logistic regression analysis was performed with myopia status as the dependent variable. Independent variables included child's age,, frequency of watching television at a distance of less than 3 meters, maternal age at childbirth, and parental myopia and birth weight. Children aged 6 to 7 years had a significantly higher risk of myopia compared to those aged 3 years (OR = 2.982, 95% *CI*: 1.292 – 6.884). Frequent of watching television at a distance of less than 3 meters was associated with a higher risk of myopia. Children who "always" watched TV less than 3 meters had an increased risk (OR = 4.051, 95% *CI*: 1.261 – 13.011,), as did those who "often" (OR = 3.171, 95% *CI*: 1.280 – 7.855). Additionally, a positive correlation was observed between maternal age at childbirth and myopia incidence (OR = 1.076, 95% *CI*: 1.010 – 1.148), and children with myopic parents had a significantly higher risk (OR = 1.953, 95% *CI*:

1.048 – 3.640,).Children with low birth weight had a higher risk of myopia compared to those with normal birth weight (OR = 3.169, 95% *CI*: 1.045 – 9.606, $p = 0.042$). No significant difference was observed between children with macrosomia and those with normal birth weight ($p > 0.05$).

In conclusion, age, frequency of close distance TV viewing, maternal age, parental myopia, and low birth weight were significant factors associated with myopia risk in preschool children. Detailed results are presented in Table 4.

3.4 Machine Learning Model Results

During the feature selection phase, candidate variables were screened by integrating three methods: univariate analysis, XGBoost-based Recursive Feature Elimination (XGBoost-RFE), and LASSO regression. The variables selected by the two machine learning methods showed a certain degree of overlap (see Tables 5 and 6). By synthesizing the importance rankings from each method and incorporating existing research evidence, several key variables were identified for subsequent model construction. Ultimately, ten predictive features were selected for the final model, including: TV (frequency of watching television with eyes less than 3 meters from the screen), ALLM (whether both parents are myopic), MABB (maternal age at childbirth), chsweight (birth weight (kg)), sleeptime (average daily sleep time), age (age (years)), chxuyy (frequency of continuous near work without breaks for >30 minutes), weiyangm (duration of exclusive breastfeeding (months)), Physicalactivity (average daily physical activity time), and FW (father’s education level).

The classification performance of each model, averaged across the five imputation sets, is summarized in Table 5 and Figure 1. Although the Generalized Linear Model (GLM) achieved the highest AUC, indicating strong overall discrimination between children with and without caregiver-reported myopia, other performance metrics were also considered, where both correctly identifying true cases (high recall) and ensuring prediction reliability (high precision) are critical.The Super Learner (SL) ensemble demonstrated a more balanced performance across multiple classification metrics. Although its AUC (0.742 ± 0.0036) was slightly lower than GLM's, SL excelled in all other key classification metrics: it attained the highest accuracy (0.943), near-perfect recall (0.963), and consequently, the highest F1-score (0.971), which harmonizes precision and recall. In contrast, GLM, despite its high precision (0.991), showed substantially lower recall (0.724) and accuracy (0.724), indicating a potential bias towards conservative predictions that might miss a proportion of children with caregiver-reported myopia. Furthermore, the low standard deviation of SL's AUC (0.0036) suggests commendable performance stability across imputed datasets. Therefore, the SL model was selected for subsequent variable importance analysis, providing an additional machine learning-based perspective on factors associated with preschool myopia.

The specific feature importance scores within the SL ensemble are detailed in Figure 2. As illustrated, occasional television viewing at a close distance (＜3m) was showed the highest relative importance within

the SL model, followed by both parental myopia. Other significant associated factors included paternal education level, prolonged near work , and breastfeeding duration, while these factors were not recognized as significant by the logistic regression model. Overall, both approaches highlighted the importance of familial and behavioral characteristics, while the machine learning framework identified additional variables that may contribute to the classification of preschool myopia. Compared with logistic regression, which provides interpretable effect estimates based on predefined relationships, the SL model may capture complex patterns among multiple variables and provide complementary information. Notably, some variables identified as important by the SL model were not statistically significant in conventional regression analyses. This discrepancy may be attributed to differences between statistical inference and machine learning approaches. Logistic regression evaluates predefined associations under specific model assumptions, whereas machine learning methods may identify variables that contribute to classification through complex, nonlinear, or multivariable patterns even when their individual marginal associations are weak.

4. Discussion

This study investigated factors associated with caregiver reported myopia among preschool children and further applied machine learning approaches as a complementary strategy to identify potentially important factors associated with myopia classification. The prevalence of caregiver reported myopia was 2.33% in this population. Our findings are consistent with previous studies in preschool children, supporting the roles of age, parental myopia, and screen related behaviors in early refractive development(Matsumura S., et al., 2022). Beyond these established factors, our analysis highlighted several early life and sociodemographic characteristics, including birth weight, breastfeeding duration, and paternal education level, as potentially relevant factors associated with myopia status. These findings further reinforce the multifactorial nature of early childhood myopia, suggesting that genetic susceptibility, early life conditions, and behavioral exposures may collectively contribute to variations in refractive development.

The observed prevalence in our study was relatively low, which is expected given the younger age distribution of the study population and is consistent with findings from studies using comparable methodologies(Huang L., et al, 2021). Previous studies have demonstrated that differences in age structure, diagnostic criteria, and study settings may also contribute to variations across studies(Li T., et al, 2019; Alvarez-Peregrina C., et al, 2020). In our study, the prevalence of caregiver-reported myopia increased with age, ranging from 1.70% to 5.08% across different age groups. Older preschool children showed a higher prevalence compared with younger children, consistent with previous observations(LI Q., et al., 2023). This pattern likely reflects the cumulative nature of myopia development and increasing exposure to environmental factors over time. Rather than indicating an intrinsic effect of age itself, the observed age-related differences may be attributable to prolonged exposure to near work, screen use, and early educational activities among older children. Therefore, age related variation in myopia prevalence may represent the accumulation of genetic susceptibility and environmental influences during early

childhood(Zhang C., et al., 2022).

Beyond age related differences, several behavioral and environmental factors were identified as important variables associated with myopia among preschool children. Among these factors, visual behavior patterns, including close-distance television viewing and prolonged near work activities, may represent modifiable exposures during early childhood. Frequent television viewing at a distance of less than 3 meters and sustained near work activities were highlighted by the machine learning model, suggesting that the cumulative characteristics of children's visual environments may contribute to early refractive development.

A study conducted in Liyang, China, reported that closer viewing distance to computer screens was associated with childhood myopia(Chen C., et al.,2022), while a longitudinal study from Rotterdam further suggested that computer use may increase myopia risk among school-aged children(Enthoven C.A., et al.,2020). Prolonged near work activities may increase accommodative demand and reduce opportunities for visual relaxation, thereby potentially promoting myopic changes. Electronic device use has increasingly become an important component of children's visual environments. Although computer use was not identified as a major factor in our preschool population, the discrepancy may reflect differences in age groups, usage patterns, and exposure contexts. Compared with computers, television viewing is more commonly used as passive entertainment among preschool children and may better represent habitual screen exposure within the home environment. Furthermore, screen related behaviors may reflect broader lifestyle patterns rather than isolated device exposure. Prolonged television viewing is often accompanied by extended sedentary time and reduced opportunities for outdoor activities, which may collectively influence refractive development(Pagani L. S., et al.,2010). During preschool years, prolonged near work may represent broader familial and environmental contexts associated with children's visual habits(Song Y., et al.,2026), largely shaped by household routines and parental supervision. A study among primary and middle school students across six provinces in China showed that insufficient parental control over children's screen time was associated with increased myopia risk(Zhou J., et al.,2016). These findings highlight the importance of establishing healthy visual habits during early childhood, including appropriate screen use, maintaining suitable viewing distances, and encouraging balanced daily activities.

In addition to screen-related behaviors and near work activities, outdoor exposure and physical activity represent other important components of children's daily lifestyle. Increased outdoor time has been consistently reported as a protective factor against myopia, particularly among school aged children, potentially through mechanisms involving light exposure and regulation of ocular growth(Zhao L., et al.,2024; Kaiti R., et al.,2021; He M., et al., 2015; Wu P. C., et al., 2018). However, outdoor activity duration was not identified as a dominant factor in our preschool population, a Japanese study similarly reported no association in preschool children(Matsumura S., et al., 2022). This difference may be related to variations in age distribution, measurement approaches, and the relatively low prevalence of myopia during early childhood. Nevertheless, physical activity were retained as informative variables in the machine learning framework, suggesting that broader lifestyle patterns may contribute to the identification of children

with different myopia related profiles.

While behavioral and environmental exposures are potentially modifiable targets for myopia prevention, they may not fully explain the development of myopia during preschool years. Compared with school aged children, preschool children have experienced shorter periods of intensive educational exposure and near work demands, whereas genetic susceptibility and early developmental factors may remain particularly relevant during this stage. Among the identified factors, parental myopia was one of the important variables associated with myopia in preschool children. This finding is consistent with previous epidemiological studies(PäRSSINEN O., et al., 2019), demonstrating a strong relationship between parental myopia and childhood myopia, suggesting the contribution of inherited susceptibility to refractive developmen. However, parental myopia may also represent shared family environments, as parents with myopia may transmit not only genetic predisposition but also visual habits, such as increased near work exposure or reduced outdoor activities, to their children(Martínez-Albert N., et al., 2023). Therefore, the association between parental myopia and childhood myopia likely reflects the combined influence of genetic and environmental pathways.

In addition to familial predisposition, several early life characteristics were identified as potentially important factors associated with preschool myopia, including birth weight and breastfeeding duration. Early life development may influence ocular growth through complex biological pathways, as prenatal and postnatal growth conditions can affect tissue development and maturation processes. The identification of birth weight as an important variable in our machine learning framework, suggests that early developmental information may provide additional insights into myopia related profiles in preschool children. In addition, Previous studies have suggested that lower birth weight has been linked to less hyperopic refractive error in early childhood and continues to correlate with myopia into adulthood, highlighting the long term influence of early life determinants(Fieß A., et al., 2021). Maternal age at childbirth was positively associated with myopia in children, with each additional year of maternal age increasing the risk of myopia. This finding is consistent with previous research from Beijing(Lin Z., et al. 2015). Factors associated with advanced maternal age, such as poor nutrition, hormonal changes, and lifestyle during pregnancy or postpartum, may associated with fetal or infant eye development. Delayed intrauterine growth has been associated with myopia(Rahi J. S., et al. 2011). The underlying mechanisms remain unclear and may involve complex interactions between prenatal conditions, socioeconomic factors, and early developmental environments. Breastfeeding duration was another early life factor highlighted in our analysis. Breastfeeding may influence childhood development through nutritional, immune, and neurodevelopmental pathways, and its potential relationship with refractive development has received increasing attention. However, evidence regarding the association between breastfeeding and myopia remains inconsistent(Zhao M. Y., et al. 2022), possibly due to differences in study populations and adjustment for confounding factors. Our findings suggest that early life nutritional history may be considered when exploring the multifactorial mechanisms underlying preschool myopia. Furthermore, paternal education was identified as an important demographic factor.

Educational attainment may act as a proxy for socioeconomic status, health awareness, parenting practices, and children's learning environments. Higher parental education may be associated with greater emphasis on early educational activities, potentially increasing near work exposure, while also reflecting better health literacy and preventive behaviors. Therefore, the relationship between parental education and childhood myopia is likely complex and influenced by multiple social and behavioral pathways(Xi Y., et al. 2025).

The application of machine learning in this study provided an integrated framework for evaluating heterogeneous information from familial, early life, and behavioral domains. By identifying potentially important factors associated with preschool myopia, this approach may contribute to a more comprehensive understanding of the multifactorial characteristics of early refractive development(Peng W., et al. 2023).

The study has several limitations. First, the cross sectional design prevents causal inference regarding the identified associations. Although factors such as parental history, early life characteristics, and behavioral exposures were found to be associated with preschool myopia, longitudinal studies are needed to determine whether these factors contribute to the subsequent development of myopia over time. In addition, myopia status and several exposure variables were obtained through caregiver-reported questionnaires, which may have introduced potential misclassification of myopia status and limited the accuracy of exposure assessment. Future studies incorporating standardized ophthalmic examinations and objective measurements of behavioral and biological factors are needed to improve the validity of exposure and outcome assessment. Moreover, the relatively small number of children with myopia (n = 45, 2.33%) limited the statistical power of conventional regression analyses and may have constrained the ability of machine learning models to fully capture relevant patterns. The limited number of positive cases may also affect the stability and precision of model performance estimates. Furthermore, although ROSE was applied to address the severe class imbalance between myopic and non-myopic children, synthetic resampling generates artificial samples that may not fully represent the characteristics and distribution of real-world cases. Therefore, the high recall observed for the SL model should be interpreted with caution, as model performance may vary across populations with different prevalence levels and case distributions. Future studies should compare different imbalance-handling strategies using larger datasets with sufficient numbers of myopic cases. Finally, although internal validation procedures were performed, the generalizability of the developed models remains uncertain. External validation in independent populations is required before the models can be considered for application in routine preschool vision screening settings. In addition, this study was conducted in a single region of China, and further validation across populations with diverse genetic, environmental, and lifestyle characteristics is needed.

From a public health perspective, identifying factors associated with preschool myopia may provide important implications for early prevention and population-based vision care. Preschool years represent a critical developmental period during which visual behaviors, lifestyle patterns, and early developmental conditions remain potentially modifiable. By integrating information from multiple sources of information, the machine learning framework explored in this study may provide additional insights into children with

different myopia related profiles and support the development of more targeted monitoring strategies in preschool vision care(Ron Y., et al., 2025; Qi Z, et al.,2024). Such an approach may support the optimization of preschool vision screening programs by highlighting children who may benefit from closer monitoring and early behavioral guidance, while maintaining routine surveillance within population-based screening systems. Ultimately, the integration of data-driven approaches into preschool vision care may contribute to provide insights for future resource planning and prevention research and support evidence-based strategies for childhood myopia prevention(Brennan N. A., et al., 2020; Bullimore M. A., et al.,2020).

5. Conclusions

In this cross sectional study, we identified several factors associated with myopia among preschool children by integrating machine learning approaches with multidimensional epidemiological information. The identified factors encompassed familial characteristics, early life conditions, and behavioral exposures, highlighting the multifactorial nature of myopia development during early childhood. The application of machine learning provided a comprehensive framework for integrating heterogeneous information and identifying potentially important factors associated with preschool myopia. These findings may contribute to a better understanding of early childhood myopia and provide evidence to support targeted prevention strategies and vision care practices in preschool populations.

References

1 Flitcroft D. I., et al. IMI - Defining and Classifying Myopia: A Proposed Set of Standards for Clinical and Epidemiologic Studies. *Investigative ophthalmology & visual science*, 2019, 60(3): M20-m30.

2 World Health Organization. WHO launches first World report on vision. From https:/ wwwwhoint/zh/news-room/detail/08-10-2019-who-launches-first-world-report-on-vision(2019).

3 Rudnick A. R., et al. Global variations and time trends in the prevalence of childhood myopia, a systematic review and quantitative meta-analysis: implications for aetiology and early prevention. *The British journal of ophthalmology*, 2016, 100(7): 882-90.

4 Liang C. L., et al. Impact of family history of high myopia on level and onset of myopia. *Investigative ophthalmology & visual science*, 2004, 45(10): 3446-52.

5 Saw S. M., Gazzard G., Shih-Yen E. C., Chua W. H.. Myopia and associated pathological complications. *Ophthalmic & physiological optics*, 2005, 25(5): 381-91.

6 National Bureau For Disease Control Aan Prevention P S R O C. The National Bureau for Disease Control and Prevention has actively promoted the prevention and control of myopia in children and adolescents. From https://www.ndcpa.gov.cn/jbkzzx/c100008/common/content/content_1764617954927783936.html. (2024)

7 Zhong P.L., et al. Trend of the detection rate of myopia among Chinese Han children and adolescents aged 7-18 years from 2010 to 2019. *Chinese journal of preventive medicine*, 2023 Feb 28;57:20-26. Chinese. doi: 10.3760/cma.j.cn112150-20221008-00964.

8 Chen J., et al. Forcasting the prevalence of myopia among students aged 6-18 years in China from 2021 to 2030 . *Chinese journal of ophthalmology*, 2021 Apr 11;57(4):261-267. Chinese. doi: 10.3760/cma.j.cn112142-20201228-000851.

9 Lanca C., Saw S. M.. The association between digital screen time and myopia: A systematic review. *Ophthalmic & physiological optics*, 2020 Mar;40(2):216-229. doi: 10.1111/opo.12657.

10 Liu X. N., Naduvilath T. J., Sankaridurg P. R.. Myopia and sleep in children-a systematic review. *Sleep*,. 2023 Nov 8;46(11):zsad162. doi: 10.1093/sleep/zsad162.

11 Chakraborty R., Ostrin L. A., Nickla D. L., Iuvone P. M., Pardue M. T., Stone R. A.. Circadian rhythms, refractive development, and myopia. *Ophthalmic Physiol Opt*. 2018 May;38(3):217-245. doi: 10.1111/opo.12453.

12 Low W., et al. Family history, near work, outdoor activity, and myopia in Singapore Chinese preschool children. *Br J Ophthalmol*. 2010 Aug;94(8):1012-6. doi: 10.1136/bjo.2009.173187.

13 He X., et al. Time Outdoors in Reducing Myopia: A School-Based Cluster Randomized Trial with Objective Monitoring of Outdoor Time and Light Intensity. *Ophthalmology*, 2022 Nov;129(11):1245-1254. doi: 10.1016/j.ophtha.2022.06.024.

14 Lin Z., et al. The Association between Maternal Reproductive Age and Progression of Refractive Error in Urban Students in Beijing. *PLoS One*. 2015 Sep 30;10(9):e0139383. doi: 10.1371/ journal.pone.0139383..

15 Ma Y., et al. Cohort study with 4-year follow-up of myopia and refractive parameters in primary schoolchildren in Baoshan District, Shanghai. *Clinical & experimental ophthalmology*, 2018 Nov;46(8):861-872. doi: 10.1111/ceo.13195.

16 Wolffsohn J. S., et al. IMI - Myopia Control Reports Overview and Introduction. *Investigative ophthalmology & visual science*, 2019 Feb 28;60(3):M1-M19. doi: 10.1167/iovs.18-25980.

17 Liu L, et al. Prediction of premyopia and myopia in Chinese preschool children: a longitudinal cohort. *BMC ophthalmology*, 2021 Jul 21;21(1):283. doi: 10.1186/s12886-021-02045-8..

18 Jiang F., et al. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2017 Jun 21;2(4):230-243. doi: 10.1136/svn-2017-000101.
19 Cao K., Verspoor K., Sahebjada S., Baird P. N.. Accuracy of Machine Learning Assisted Detection of Keratoconus: A Systematic Review and Meta-Analysis. *J Clin Med*. 2022 Jan 18;11(3):478. doi: 10.3390/jcm11030478.
20 Wu J. H., Liu T. Y. A., Hsu W. T., Ho J. H., Lee C. C.. Performance and Limitation of Machine Learning Algorithms for Diabetic Retinopathy Screening: Meta-analysis. Journal of medical Internet research, 2021 Jul 3;23(7):e23863. doi: 10.2196/23863.
21 Tong H. J., Huang Z. M., Li Y. L., Chen Y. M., Tian B., Ding L. L., Zhu L. L.. Machine learning to analyze the factors influencing myopia in students of different school periods. *Frontiers in public health*,2023 Jun 1;11:1169128. doi: 10.3389/fpubh.2023.1169128.
22 Mu J., Zhang Z., Wu X., Chen S., Geng H., Duan J.. Refraction and ocular biometric parameters in 3-to 6-year-old preschool children : a large-scale population-based study in Chengdu, China. *BMC ophthalmology*, 2024 May 6;24(1):207. doi: 10.1186/s12886-024-03467-w.
23 NNunes A. F., Cunha M., Sousa M. C., Godinho C.A.. Prevalence, sociodemographic risk factors, and coverage of myopia correction among adolescent students in the central region of Portugal. *BMC public health*, 2024 Sep 12;24(1):2490. doi: 10.1186/s12889-024-19914-8.
24 Ying B., Chandra R.S., Wang J., Cui H., Oatts J.T.. Machine Learning Models for Predicting Cycloplegic Refractive Error and Myopia Status Based on Non-Cycloplegic Data in Chinese Students. *Translational vision science & technology*, 2024 Aug 1;13(8):16. doi: 10.1167/tvst. 13.8.16.
25 OFFICE H M P S G. Interpretation of the Implementation Plan of the Myopia Prevention and Control Reform Experimental Area for Children and Adolescents in Huaibei City. Aviliable at: https://www.huaibei.gov.cn/zwgk/public/15/59948171.html (2025),
26 Liu S. M., Wang Y. T., Chen J., Wang F., Shih S. F.. Parent engagement in children's eye care behavior and vision-related quality of life: a cross-sectional study. *BMC Public Health* **26**, 879 (2026). https://doi.org/10.1186/s12889-026-26277-9
27 Tian L., Zhu M., Song Y., Wang Y., Jiang Y.. The prevalence of myopia and eye health behaviors among 3 to 18 years: a cross-sectional survey study. *BMC Public Health.* 2025 May 7;25(1):1688. doi: 10.1186/s12889-025-22906-x.
28 Xie X., Ma J., Chen Q., Li X., Jia L., Cao J.. Evaluation of myopia status and eye use behavior in school-age and preschool children. *PLoS One*. 2025 Jun 11;20(6):e0322569. doi: 10.1371/journal.pone.0322569.
29 Matsumura S., et al. Prevalence of Myopia and Its Associated Factors Among Japanese Preschool Children. Front Public Health. 2022 Jun 22;10:901480. doi: 10.3389/fpubh.2022.901480.
30 Huang L., et al. Combination Effect of Outdoor Activity and Screen Exposure on Risk of Preschool Myopia: Findings From Longhua Child Cohort Study. *Frontiers in public health*, 2021 Mar 5;9:607911. doi: 10.3389/fpubh.2021.607911.
31 Li T., Zhou X., Chen X., Qi H., Gao Q.. Refractive Error in Chinese Preschool Children: The Shanghai Study . *Eye & contact lens*, 2019 May;45(3):182-187. doi: 10.1097/ICL.0000000000000555.
32 Alvarez-Peregrina C., Sánchez-Tena M.Á., Martinez-Perez C., Villa-Collar C.. The Relationship Between Screen and Outdoor Time With Rates of Myopia in Spanish Children. *Frontiers in public health*, 2020 Oct 14;8:560378. doi: 10.3389/fpubh.2020.560378.
33 Li Q., et al. Prevalence Trend of Myopia during the Post-COVID-19 Epidemic Period among Preschoolers: A Prospective School-based Study. *Optom Vis Sci*, 2023 Nov 1;100(11):727-734. doi: 10.1097/OPX.0000000000002069.

34 Zhang C., Li L., Jan C., Li X., Qu J.. Association of School Education With Eyesight Among Children and Adolescents. *JAMA network open*, 2022 Apr 1;5(4):e229545. doi: 10.1001/jamanetworkopen.2022.9545.
35 Chen C., et al. Investigation on the prevalence and influencing factors of myopia among children and adolescents in Liyang city. *American journal of translational research*, 2022 Oct 15;14(10):7164-7171.
36 Enthoven C.A., Tideman J. W. L., Polling J. R., Yang-Huang J., Raat H., Klaver C. C. W.. The impact of computer use on myopia development in childhood: The Generation R study. *Preventive medicine*, 2020 Mar;132:105988. doi: 10.1016/j.ypmed.2020.105988.
37 Pagani L. S., Fitzpatrick C., Barnett T. A., Dubow E.. Prospective associations between early childhood television exposure and academic, psychosocial, and physical well-being by middle childhood. *Archives of pediatrics & adolescent medicine*, 2010 May;164(5):425-31. doi: 10.1001/archpediatrics.2010.50.
38 Song Y., Zheng W., Yang X., Yang S., Du F., Tian H.. Near-work activity duration and myopia in children: An updated systematic review and meta-analysis. *Biomed Rep*. 2026 Mar 17;24(5):58. doi: 10.3892/br.2026.2131.
39 Zhou J., et al. Prevalence of myopia and influencing factors among primary and middle school students in 6 provinces of China. *Chinese Journal of Epidemiology*, 2016 Jan;37(1):29-34. Chinese. doi: 10.3760/cma.j.issn.0254-6450.2016.01.006.
40 Zhao L., et al. Prevalence and risk factors of myopia among children and adolescents in Hangzhou. *Scientific reports*, 2024 Oct 19;14(1):24615. doi: 10.1038/s41598-024-73388-7.
41 Kaiti R., Shyangbo R., Sharma I. P., Dahal M.. Review on current concepts of myopia and its control strategies. *International journal of ophthalmology*, 2021 Apr 18;14(4):606-615. doi: 10.18240/ijo.2021.04.19.
42 He M., et al. Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial. *Jama*, 2015 Sep 15;314(11):1142-8. doi: 10.1001/jama.2015.10803.
43 Wu P. C., et al. Myopia Prevention and Outdoor Light Intensity in a School-Based Cluster Randomized Trial. *Ophthalmology*, 2018 Aug;125(8):1239-1250. doi: 10.1016/j.ophtha.2017.12.011.
44 PäRSSINEN O., KAUPPINEN M.. Risk factors for high myopia: a 22-year follow-up study from childhood to adulthood. *Acta ophthalmologica*, 2019 Aug;97(5):510-518. doi: 10.1111/aos.13964.
45 Martínez-Albert N., Bueno-Gimeno I., Gené-Sampedro A.. Risk Factors for Myopia: A Review. *Journal of clinical medicine*, 2023 Sep 19;12(18):6062. doi: 10.3390/jcm12186062.
46 Fieß A., et al. The relationship of ocular geometry with refractive error in normal and low birth weight adults. *Journal of optometry*, 2021 Jan-Mar;14(1):50-57. doi: 10.1016/j.optom.2020.08.004. .
47 Rahi J. S., Cumberland P. M., Peckham C. S.. Myopia over the lifecourse: prevalence and early life influences in the 1958 British birth cohort. *Ophthalmology*, 2011 May;118(5):797-804. doi: 10.1016/j.ophtha.2010.09.025.
48 Zhao M. Y., Zhu Y., Xu S. J., Tao F. B.. Research progress on influencing factors of hyperopia reserve in children. *Chinese Journal of Epidemiology*, 2022 Oct 10;43(10):1671-1676. Chinese. doi: 10.3760/cma.j.cn112338- 20220420-00324.
49 Xi Y., Zhu W., Yan W., Wei X., Zhao Q., Dai S.. Optimizing myopia prediction in children and adolescents using machine learning: a multi-factorial risk assessment model. *Frontiers in medicine*, 2025 Nov 13;12:1672432. doi: 10.3389/fmed.2025.1672432.

50 Peng W., Wang F., Sun S., Sun Y., Chen J., Wang M.. Does multidimensional daily information predict the onset of myopia? A 1-year prospective cohort study. *Biomedical engineering online,* 2023 May 13;22(1):45. doi: 10.1186/s12938-023-01109-8.

51 Ron Y., Ron T., Fridman N., Goldstein A.. Predicting onset of myopic refractive error in children using machine learning on routine pediatric eye examinations only. *Sci Rep*. 2025 Aug 17;15(1):30055. doi: 10.1038/s41598-025-13990-5.

52 Qi Z, Li T, Chen J, Yam JC, et al. A deep learning system for myopia onset prediction and intervention effectiveness evaluation in children. *NPJ Digit Med*. 2024 Aug 7;7(1):206. doi: 10.1038/s41746-024-01204-7.

53 Brennan N. A., Toubouti Y. M., Cheng X., Bullimore M. A.. Efficacy in myopia control. *Progress in retinal and eye research*, 2021 Jul;83:100923. doi: 10.1016/j.preteyeres.2020.100923.

54 Bullimore M. A., Richdale K.. Myopia Control 2020: Where are we and where are we heading?. *Ophthalmic Physiol Opt*. 2020 May;40(3):254-270. doi: 10.1111/opo.12686.

Authorship contribution

Manuscript drafting: Z. L. and D. S.；Data acquisition: Y. N., T. Y., and H. Z.；Statistical analysis and data interpretation: Z. L and Z. Z.；Discussion and manuscript drafting: Z. L., D. S., and M. C.；Manuscript editing and critical revision: M. C. and Y. G.；Resources, supervision, and funding acquisition: T. Y., M. C., and Y. G.. All authors reviewed and approved the final manuscript.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author upon reasonable request. Due to ethical restrictions and privacy considerations, the data cannot be publicly deposited.

Ethics approval and consent to participate

The research was performed in accordance with the Declaration of Helsinki. This study was approved by the Ethics Committee of Capital Institute of Pediatrics under the code (NO. SHERLL2023017). Informed consent was obtained from children's parents or guardians.

Declaration of Interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This study was supported by Capital Institute of Pediatrics Grant Fund for Institutional Projects (SKYJ-2023-07), Guangdong Medical University Undergraduate Innovation and Entrepreneurship Education Base Program (JDXM2024055) and Innovation Training Program for Undergraduates of Guangdong Medical University (GDMUCX2024183, GDMUCX2024142). Beijing Municipal Hospital Research Training Program (No.PG2024023), National Health Commission Capacity Building and Continuing Education Center (No. GWJJMB 202510030006). The funding body had no role in the design, analysis or writing of this article.

Table 1. Demographics and parental characteristics of preschool children

Variables	Children N (%), Mean (SD)	Myopia N (%), Mean (SD)	Non-myopia N (%), Mean (SD)	$\chi 2$	*P*
Age(year)[a]	3.99±1.03	4.24±1.17	3.98±1.03	8.061	0.043*
3~	849 (43.88)	16 (1.98)	833 (98.12)		
4~	478 (24.70)	12 (2.51)	466 (97.49)		
5~	411 (21.24)	7 (1.70)	404 (98.30)		
6~7	197 (10.18)	10 (5.08)	187 (94.92)		
Sex[a]				0.426	0.514
Male	982 (50.75)	25 (2.55)	746 (97.45)		
Female	953 (49.25)	20 (2.10)	723 (97.90)		
Father's educational level[a]				3.667	0.160
junior high school or below	203 (10.49)	8 (3.94)	195 (96.06)		
senior high school	347 (17.93)	10 (2.88)	337 (97.12)		
College graduate or above	1385 (71.58)	27 (1.95)	1358 (98.05)		
Mother's educational level[a]				1.413	0.493
junior high school or below	217 (11.21)	7 (3.23)	210 (96.77)		
senior high school	367 (18.97)	10 (2.72)	357 (97.28)		
College graduate or above	1351 (69.82)	28 (2.07)	1323 (97.93)		
Parents has myopia[a]				7.264	0.026*
No parental myopia	604 (31.21)	12 (1.99)	592 (98.01)		
One parent with myopia	770 (39.79)	12 (1.56)	758 (98.44)		
Both parents with myopia	561 (28.99)	21 (3.74)	540 (96.26)		
Total	1935 (100.00)	45 (2.33)	1890 (97.67)		

a:Chi square test, b:Fisher exact test, c:Rank sum test

*:$P<0.05$, **:$P<0.001$.

Table 2. Analysis results of Prenatal Conditions, Birth and Feeding Practices in preschool children

Variables	Children N (%), Mean (SD)/ Median(Q1,Q3)	Myopia N (%), Mean (SD)/ Median(Q1,Q3)	Non-myopia N (%), Mean (SD)/ Median(Q1,Q3)	$\chi 2/Z$	*P*
Paternal age at childbirth (years)[c]	30(28,35)	32(28.5,36.5)	30(28,35)	-1.533	0.125
Advanced paternal age[a]				1.238	0.266
Yes	487 (25.17)	14 (2.87)	473 (97.13)		
No	1448 (74.83)	31 (2.14)	1417 (97.86)		
Maternal age at childbirth (years)[c]	30(27,33)	31(28.5,35)	30(27,33)	-2.190	0.029*
Advanced maternal age[a]				1.953	0.162
Yes	364 (18.81)	13 (3.57)	351 (96.43)		
No	1571 (81.19)	32 (2.04)	1539 (97.96)		
Maternal anemia[a]				0.006	0.938
Yes	293 (15.14)	7 (2.39)	286 (96.47)		
No	1642 (84.86)	38 (2.31)	1604 (97.96)		
Gestational hypertension[b]					0.368
Yes	55 (2.84)	2 (3.64)	53 (96.36)		
No	1880 (97.16)	43 (2.29)	1837 (97.71)		
Gestational diabetes[b]					0.239
Yes	73 (3.77)	3 (5.36)	70 (95.89)		
No	1862 (96.23)	42 (2.34)	1820 (97.74)		
Maternal work stress[b]					0.325
Yes	50 (2.58)	2 (4.00)	48 (96.00)		
No	1875 (97.42)	43 (2.28)	1842 (97.72)		
Birth weight(kg)[c]	3.50 (3.15,3.85)	3.40 (3.0,4.0)	3.50 (3.17,3.85)	-0.769	0.442
Birth weight (group)[b]					0.039*
Low birth weight infants	70 (3.62)	4 (5.71)	66 (94.29)		
Normal birth weight infants	1390 (74.73)	26 (1.87)	1364 (98.13)		
High birth weight infants	400 (21.51)	13 (3.25)	387 (96.75)		
Gestational weeks[b]					1.000
<37weeks	87 (4.50)	2 (2.30)	85 (97.70)		
37~42weeks	1848 (96.19)	43 (2.33)	1805 (97.67)		
Mode of delivery[a]				0.148	0.701
Eutocia	1214 (62.74)	27 (2.22)	1187 (97.78)		
Cesarean delivery	721 (37.26)	18 (2.50)	703 (97.50)		
Exclusive breastfeeding[a]				0.551	0.458
Yes	823 (54.65)	18 (2.19)	805(97.81)		
No	683 (45.35)	19 (2.78)	664(97.22)		
Breastfeeding duration (months)[c]	13(10,18)	12(6.50,18)	13(10,18)	-1.471	0.141
Breastfeeding duration (group)[a]				1.277	0.259
Less or equal to 12 months	927 (49.81)	24 (2.59)	903 (97.41)		
12 months or more	934 (50.19)	17 (1.82)	917 (98.18)		
Decsyed teeth[a]				0.223	0.637

No	1599 (82.64)	36 (2.25)	1563 (97.75)	
Yes	336 (17.36)	9 (2.68)	327 (97.32)	
Allergic Conjunctivitis[b]				0.668
No	1868 (96.54)	43 (2.30)	1825 (97.70)	
Yes	67 (3.46)	2 (2.99)	65 (97.01)	
Total	1935 (100.00)	45 (2.33)	1890 (97.67)	

Advanced Paternal age at childbirth: At the time of this child birth, Whether the father being 35 years or older.

Advanced maternal age at childbirth: At the time of this child birth, Whether the mather being 35 years or older.

All pregnancy-related conditions (e.g., anemia, gestational diabetes, hypertension, work stress) refer to the mother's status during pregnancy

a:Chi square test, b:Fisher exact test, c:Rank sum test

*:$P<0.05$, **:$P<0.001$.

Table 3. Results of factor analysis of Lifestyle in preschool children

Variables	Children N (%), Mean (SD)/ Median(Q1,Q3)	Myopia N (%), Mean (SD)/ Median(Q1,Q3)	Non-myopia N (%), Mean (SD)/ Median(Q1,Q3)	$\chi 2/Z$	*P*
Dietary bias[a]				0.003	0.955
No	1110 (57.36)	26 (2.34)	1048 (97.66)		
Yes	825 (42.64)	19 (2.30)	806 (97.00)		
Sweet food consumption[a]				1.913	0.384
Occasional	193 (9.97)	5 (2.59)	188 (97.41)		
Often	1176 (60.78)	23 (19.56)	1153 (98.04)		
Daily	566 (29.25)	17 (3.00)	549 (97.00)		
Outdoors activity[a]				0.216	0.642
<2h/day	668 (34.52)	17 (2.54)	651 (97.46)		
≥2h/day	1267 (65.48)	28 (2.21)	808(97.79)		
Physical activity (time)[c]	2(2,4)	2(2,4)	2(1.5,3)	-0.652	0.514
Physical activity[a]				0.009	0.925
<3h/day	998 (52.97)	23 (2.30)	975 (97.70)		
≥3h/day	886 (47.03)	21 (2.37)	865 (97.63)		
Night lamp use [a]				0.058	0.809
No	1322 (68.32)	30 (2.27)	1292 (97.73)		
Yes	613 (31.68)	15 (2.45)	598 (97.55)		
Electronics device use[b]				0.032	0.857
<1h/day	1309 (67.65)	31 (2.37)	97.63 (95.98)		
≥1h/day	626 (32.35)	22 (2.24)	97.76 (97.74)		
Viewing distance from TV (<3 m)[a]				17.894	＜0.001**
Never	538 (27.80)	8 (1.49)	530 (98.51)		
Occasional	1039 (53.70)	18 (1.73)	1021 (98.27)		
Often	281 (14.52)	14 (4.98)	267 (95.02)		
Always	77 (3.98)	5 (6.49)	72 (93.51)		
Viewing distance from computer (<50 cm)[b]				9.036	0.024*
Never	1053 (54.42)	19 (1.80)	1034 (98.20)		
Occasional	679 (35.09)	16 (2.36)	663 (97.64)		
Often	171 (8.84)	7 (4.09)	164 (95.91)		
Always	32 (1.65)	3 (9.38)	29 (90.62)		
Screen use in dim lighting[b]				2.258	0.431
Never	1516 (78.35)	32 (2.11)	1484 (97.89)		
Occasional	319 (16.49)	10 (3.13)	309 (96.87)		
Often	27 (1.91)	1 (2.70)	27 (97.30)		
Always	52 (3.26)	2 (3.17)	51 (96.83)		
Prolonged near work[b]				5.431	0.122
Never	525 (27.13)	7 (1.50)	518 (98.50)		
Occasional	1085 (56.07)	31 (3.05)	1054 (96.95)		
Often	283 (14.63)	5 (1.77)	278 (98.23)		

Always	42 (2.17)	2 (3.45)	40 (96.55)		
Sleep duration (hours/day)[c]	10(10.12)	10(9.5,11)	10(10.12)	1.782	0.41
Sleep deprivation[a]				0.362	0.330
No	363 (18.76)	10 (2.75)	353 (97.25)		
Yes	1572 (81.24)	35 (2.23)	1537 (97.77)		
Sleep quality[b]				2.915	0.088
Poor	368 (19.02)	13 (3.53)	355 (96.48)		
Good	1567 (80.98)	32 (2.04)	1535 (97.96)		
Difficulty initiating sleep[a]				2.915	0.070
Yes	368 (19.92)	13 (4.33)	287(95.67)		
No	1567 (80.08)	32(1.99)	1182(98.01)		
Total	1935 (100.00)	45 (2.33)	1890(97.67)		

Occasional: Refers to engaging in the behavior 1or less than 1 times per week.

Often: Refers to engaging in the behavior 2-5 times per week.

Always: Refers to engaging in the behavior more than 5 times per week.

Viewing distance from TV (<3 m): Frequency of TV use with a viewing distance < 3 m

Viewing distance from computer (<50 cm): Frequency of computer use with a viewing distance < 50 cm

Screen use in dim lighting : Turn off roof lights when watching TV or computer at night

Prolonged near work : Continuous near-work for more than 30 minutes without rest.

Sleep deprivation: Refers to the individual duration of sleep for less than 10 hours.

Poor (Sleep quality): Refers to a sleep quality score of 1-4 based on the survey questionnaire

Good (Sleep quality): Refers to a sleep quality score of 5-7 based on the survey questionnaire.

Variables related to visual behaviors (e.g.,**Sweet food, Viewing distance from TV (<3 m), Viewing distance from computer (<50 cm), Screen use in dim lighting and Prolonged near work**) were categorized based on frequency.

a:Chi square test, b:Fisher exact test, c:Rank sum test

*:$P<0.05$, **:$P<0.001$.

Table 4. Multivariable logistic regression analysis for factors that affected myopia among children

Variables	β	Wald	*P*	OR	95% *CI*
Age(years)		8.162	0.043*		
4~	0.301	0.573	0.449	1.352	0.619~2.951
5~	-0.171	0.120	0.730	0.843	0.321~2.218
6~7	1.093	6.552	0.010*	2.982	1.292~6.884
Watching TV less than 3 m		12.862	0.005*		
Occasional	0.117	0.720	0.788	0.947	0.479~2.641
Often	1.154	6.214	0.013*	3.171	1.280~7.855
Always	1.399	5.520	0.019*	4.051	1.261~13.011
Maternal age at childbirth (years)	0.074	5.089	0.024*	1.076	1.010~1.148
Parents has myopia (Both parents with myopia)	0.669	4.442	0.035*	1.953	1.048~3.640
Birth weight(group)[b]					
Low birth weight infants	1.153	4.115	0.042*	3.169	1.045~9.606
High birth weight infants	0.502	0.376	0.182	1.652	0.825~3.271

*$P<0.05$.

Table 5. The feature selection results of the Lasso regression model on different imputed datasets

Rank	Interpolation Set 1	Interpolation Set 2	Interpolation Set 3	Interpolation Set 4	Interpolation Set 5
1	Age	Age	Age	Age	Age
2	Father's Education Level	Father's Education Level	Father's Education Level	Father's Education Level	Father's Education Level
3	Both Parents Are Myopic	Both Parents Are Myopic	Mather's Education Level	Both Parents Are Myopic	Both Parents Are Myopic
4	Maternal Age at Childbirth	Maternal Age at Childbirth	Both Parents Are Myopic	Maternal Age at Childbirth	Maternal Age at Childbirth
5	Birth Weight (Category)	Viewing distance from TV (<3 m)	Maternal Age at Childbirth	Birth Weight (Category)	Birth Weight (Category)
6	Breastfeeding Duration >12 months	Viewing distance from computer (<50 cm)	Birth Weight (kg)	Breastfeeding (months)	Breastfeeding (months)
7	Viewing distance from TV (<3 m)	Prolonged near work	Birth Weight (Category)	Viewing distance from TV (<3 m)	Viewing distance from TV (<3 m)
8	Viewing distance from computer (<50 cm)	Sleep duration (hours/day)	Breastfeeding (months)	Viewing distance from computer (<50 cm)	Viewing distance from computer (<50 cm)
9	Prolonged near work	Sleep qality	Breastfeeding Duration >12 months	Electronic device use (Category)	Electronic device use (Category)
10	Sleep qality	Difficulty Initiating Sleep	Viewing distance from computer (<50	Prolonged near work	Prolonged near work

cm)

Viewing distance from TV (<3 m): Frequency of TV use with a viewing distance < 3 m
Viewing distance from computer (<50 cm): Frequency of computer use with a viewing distance < 50 cm
Screen use in dim lighting : Turn off roof lights when watching TV or computer at night
Prolonged near work : Continuous near-work for more than 30 minutes without rest.
Sleep deprivation: Refers to the individual duration of sleep for less than 10 hours.
Poor (Sleep quality): Refers to a sleep quality score of 1-4 based on the survey questionnaire
Good (Sleep quality): Refers to a sleep quality score of 5-7 based on the survey questionnaire.
Variables related to visual behaviors (e.g.,**Sweet food, Viewing distance from TV (<3 m), Viewing distance from computer (<50 cm), Screen use in dim lighting and Prolonged near work**) were categorized based on frequency.

Table 6. The feature selection results of XGBoost-RFE on different imputed datasets

Rank	Interpolation Set 1	Interpolation Set 2	Interpolation Set 3	Interpolation Set 4	Interpolation Set 5
1	Birth Weight (kg)	Birth Weight (kg)	Birth Weight (kg)	Birth Weight (kg)	Birth Weight (kg)
2	Maternal Age at Childbirth	Maternal Age at Childbirth	Maternal Age at Childbirth	Maternal Age at Childbirth	Maternal Age at Childbirth
3	Paternal Age at Childbirth	Both Parents Are Myopic	Breastfeeding (months)	Breastfeeding (months)	Breastfeeding (months)
4	Both Parents Are Myopic	Sleep duration (hours/day)	Both Parents Are Myopic	Paternal Age at Childbirth	Paternal Age at Childbirth
5	Breastfeeding (months)	Paternal Age at Childbirth	Paternal Age at Childbirth	Sleep duration (hours/day)	Both Parents Are Myopic
6	Physical activity time (hours/day)	Breastfeeding (months)	Physical activity time (hours/day)	Both Parents Are Myopic	Physical activity time (hours/day)
7	Sleep duration (hours/day)	Age	Sleep duration (hours/day)	Physical activity time (hours/day)	Age
8	Viewing distance from TV (<3 m)	Physical activity time (hours/day)	Age	Age	Sleep duration (hours/day)
9	Age	Viewing distance from TV (<3 m)	Viewing distance from TV (<3 m)	Viewing distance from TV (<3 m)	Viewing distance from TV (<3 m)
10	Father’s education level	Father’s education level	Father’s education level	Father’s education level	Father’s education level
11	Sex	Prolonged near work	Prolonged near work	Mode of delivery	Sex
12	Mather’ s education level	Sex	Mather’ s education level	Prolonged near work	Prolonged near work
13	Prolonged near work	Mather’ s education level	Mode of delivery	Screen use in dim lighting	Mode of delivery
14	Mode of delivery	Mode of delivery	Electronic device use (Category)	Sex	Electronic device use (Category)

Viewing distance from TV (<3 m): Frequency of TV use with a viewing distance < 3 m
Viewing distance from computer (<50 cm): Frequency of computer use with a viewing distance < 50 cm

Screen use in dim lighting : Turn off roof lights when watching TV or computer at night
Prolonged near work : Continuous near-work for more than 30 minutes without rest.
Sleep deprivation: Refers to the individual duration of sleep for less than 10 hours.
Poor (Sleep quality): Refers to a sleep quality score of 1-4 based on the survey questionnaire
Good (Sleep quality): Refers to a sleep quality score of 5-7 based on the survey questionnaire.
Variables related to visual behaviors (e.g.,**Sweet food, Viewing distance from TV (<3 m), Viewing distance from computer (<50 cm), Screen use in dim lighting and Prolonged near work**) were categorized based on frequency.

Table 7. Average Predictive Performance Metrics of the Machine Learning Models on the Test Set

Model	AUC(*95%CI*)	*SD*	*Accuracy*	*Precision*	*Recall*	*F1*
GLM	0.776(0.769-0.783)	0.0056	0.726	0.988	0.728	0.838
BRNN	0.759(0.705-0.814)	0.0438	0.788	0.985	0.796	0.881
SL	0.742(0.738-0.747)	0.0036	0.973	0.980	0.993	0.986
RF	0.733(0.725-0.743)	0.0072	0.950	0.981	0.968	0.974
LGB	0.726(0.698-0.755)	0.0229	0.968	0.979	0.989	0.984
XGB	0.724(0.703-0.746)	0.0173	0.966	0.979	0.987	0.983
NNET	0.689(0.606-0.773)	0.0672	0.830	0.984	0.840	0.905

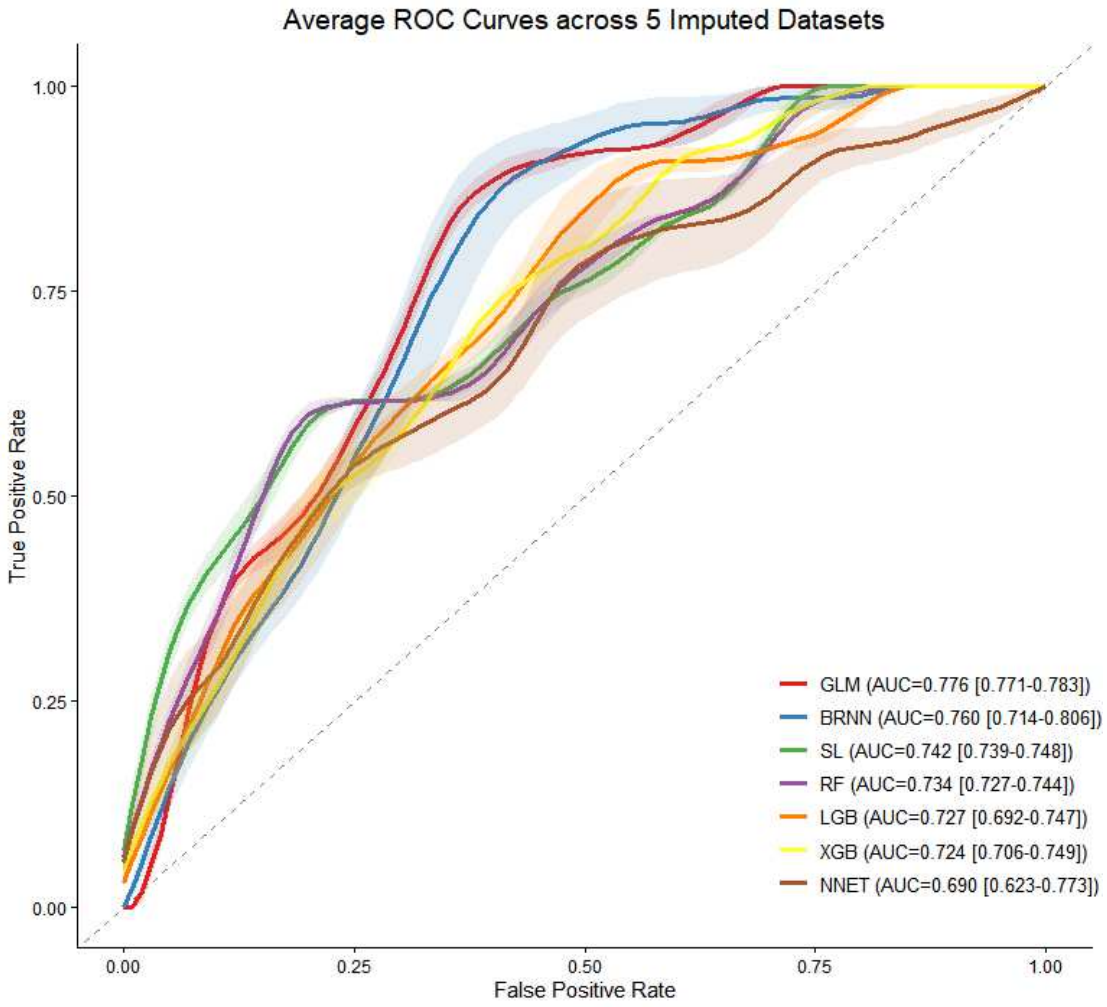


Figure 1. Comparison of Average ROC Curves Based on 5 imputation sets

Thin lines: individual imputation sets

Thick lines: average curves

Shaded: 95% confidence intervals

GLM：Generalized Linear Model

LGB: Light Gradient Boosting Machine

NNET: Neural Network

RF: Random Forest

XGB: eXtreme Gradient Boosting

BRNN: Bayesian Recurrent Neural Network

SL: Super Learner

AUC: area under the receiver operating characteristic curve

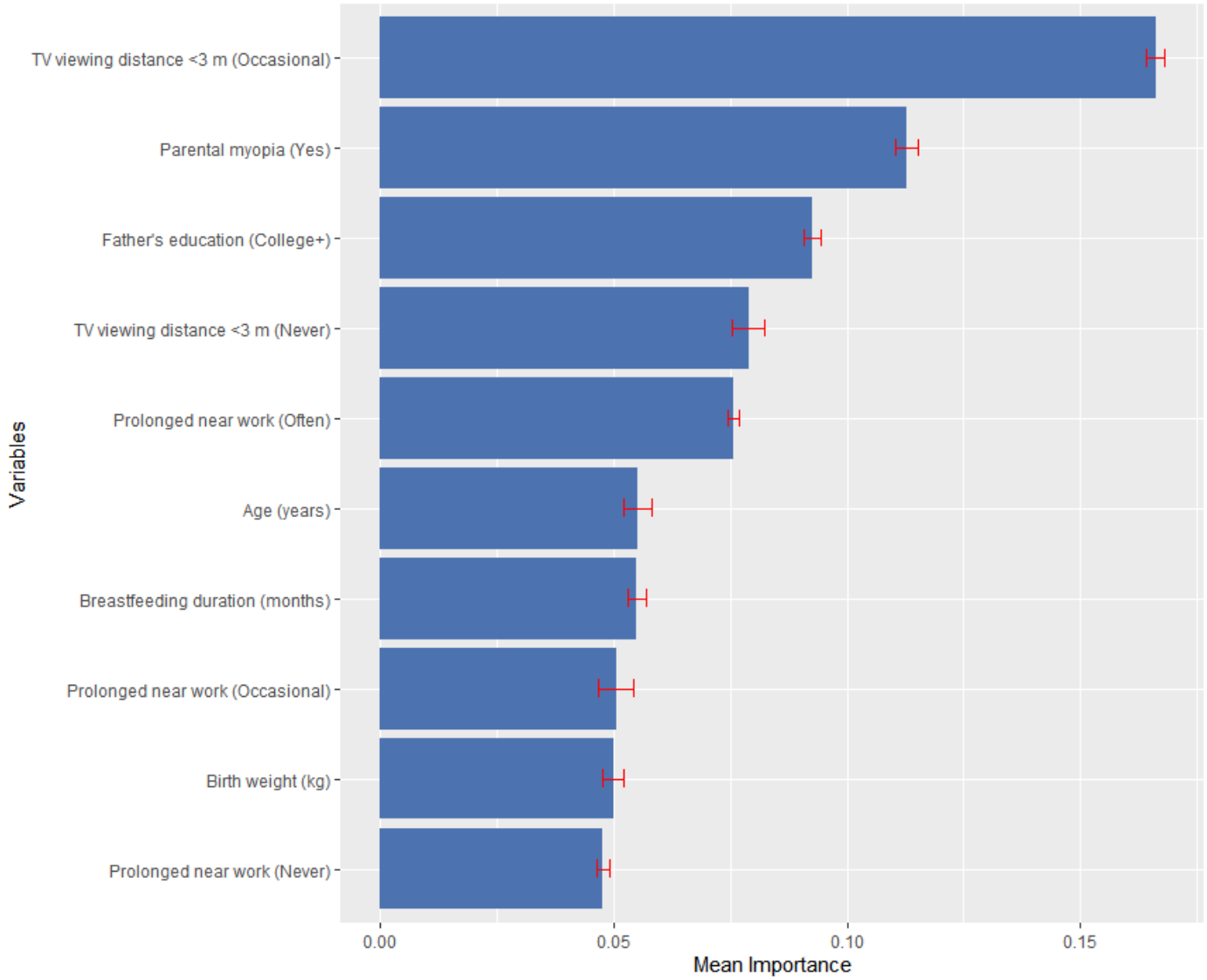


Figure 2. Variable importance plot of the SuperLearner model (SL) (Based on 5 Imputation Sets).

Occasional: Refers to engaging in the behavior 1or less than 1 times per week.
Often: Refers to engaging in the behavior 2-5 times per week.
Always: Refers to engaging in the behavior more than 5 times per week.
Prolonged near work : Continuous near-work for more than 30 minutes without rest.
Viewing distance from TV (<3 m): Frequency of TV use with a viewing distance < 3 m