

Construction of a Machine Learning-based Model for Predicting Professional Identity among Medical Students in the Context of the COVID-19 Pandemic: A Multicenter Study

Weiyun Jin

College of Humanities Education, Inner Mongolia Medical University

Jinhai Wang

Guiyang Public Health Clinical Center

Zihao Zhao

Guiyang Public Health Clinical Center

Xiaorong Li

Guiyang Public Health Clinical Center

Zhengran Liu

School of Public Health, Baotou Medical College

Wei Liu

School of Public Health, Baotou Medical College

Xue Bai

School of Public Health, Inner Mongolia Minzu University

Yanling Wang

School of Public Health, Inner Mongolia Medical University

Jianwei Qu

Department of Medical Insurance, Chifeng Municipal Hospital

Zhiqiang Zhou

Clinical Teaching Office, Affiliated Hospital of Chifeng University

Hua Dai

Guizhou University of Traditional Chinese Medicine

Hongqi Zhou

15608519407@163.com

Guiyang Public Health Clinical Center

Bensong Xian

College of Health Management, Inner Mongolia Medical University

Research Article

Keywords: Professional identity, medical students, predictive model, random forest, COVID-19, medical education, translational research, medical education intervention

Posted Date: May 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9324440/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

The ongoing spread of the COVID-19 pandemic has profoundly affected the professional identity of medical students. There is a need to develop accurate predictive tools to identify high-risk groups with a weak professional identity. This study aimed to construct and validate a machine learning model integrating multidimensional COVID-19 exposure factors to provide a basis for stratified interventions aimed at enhancing professional identity among medical students. Through translational research, this study bridges the gap between academic models and practical medical education applications, offering a data-driven precision tool for educational interventions. The model is the first predictive tool focusing on multidimensional pandemic exposure for medical students' professional identity, addressing the gap in translational predictive tools for medical education during public health crises. It can serve as a precise screening tool for medical education administrators to quickly identify students in need of prioritized interventions. The model can also be embedded into educational management systems to generate automated risk scores, thereby guiding the implementation of targeted interventions.

Methods

A multicenter cross-sectional study design was adopted, involving 3003 medical students from three universities in China. Professional identity was assessed using a validated scale, and COVID-19 exposure was measured using a 28-item questionnaire. A total of eight candidate predictors were screened. Five feature selection methods, including stepwise discriminant analysis and LASSO (Least Absolute Shrinkage and Selection Operator) regression, were used to identify the optimal variables. A total of 130 machine learning models were constructed. The dataset was stratified and split into a training set ($n = 1503$) and a validation set ($n = 1500$) at a 1:1 ratio. Model performance was comprehensively evaluated based on discrimination (AUC), classification metrics, and stability (out-of-bag error).

Results

The random forest model performed best on the validation set, with an AUC of 0.898 (95% CI: 0.886–0.910), an accuracy of 80.2%, a sensitivity of 73.1%, a specificity of 86.2%, and a kappa value of 0.577. The model demonstrated greater stability than the other models did, such as XGBoost (XGBoost validation AUC = 0.893; model AUC = 0.898, with a training–validation AUC difference of only 0.013 and a 68% reduction in computation time). COVID-19 exposure emerged as the primary predictor (Gini importance 73.7%), with a significantly higher predictive power than that of career motivation (9.6%) and the specialty category (8.6%). The integration of multiple variables significantly enhanced model performance ($\Delta\text{AUC} = 0.033$, $P = 0.008$).

Conclusions

The random forest model constructed in this study effectively predicts the professional identity levels of medical students, with COVID-19 exposure identified as a core influencing factor. The model provides a practical tool for precise interventions targeting professional identity among medical students during pandemics and similar public health events. By translating research into practice, it establishes a pathway from “model construction to clinical education application,” offering a standardized framework for medical education interventions. The model holds significant clinical and educational value and requires further external validation across different contexts. It enables rapid calculation of prediction probabilities through a simplified scale, thereby facilitating onsite screening by medical education administrators.

1. Introduction

Professional identity is a core psychological construct in the professional development of medical students. It refers to the integration of an individual's cognition, emotional attachment, and behavioral commitment to their role as a physician. It is gradually formed through interactions with professional communities, clinical environments, and societal expectations during medical practice and education [1].

This complex multidimensional construct not only serves as a key predictor of learning engagement and career choice stability but also profoundly influences the quality of future health care services, patient safety, and the long-term resilience of the physician workforce [2]. Its development constitutes a dynamic socialization process that is shaped collectively by individual characteristics (e.g., motivation and resilience), educational environments (e.g., curriculum design and role modeling), clinical practice experiences (e.g., early patient contact and clerkships), and broader sociocultural contexts (e.g., public expectations of physicians and health care policies) [3].

The COVID-19 pandemic, which was declared an unprecedented global health crisis by the World Health Organization in March 2020, has had profound and enduring impacts on global medical education and the operation of health care systems [4]. For medical students, the pandemic constitutes a unique “cluster of critical professional developmental events”: clinical clerkships were largely disrupted or cancelled, teaching models were forced to shift entirely online [5], and students widely experienced concerns about occupational exposure risks, personal health, and even doubts about their own value in the future health care system [6]. By converging during the crucial formative period of medical students’ professional identity development, these multiple disruptive factors may have fundamentally altered the students’ perceived balance between the value, rewards, and personal sacrifices associated with a medical career [6]. A growing body of evidence suggests that the pandemic not only was a significant acute psychological stressor but also functioned as a “professional identity stress test,” exerting a notably negative influence on the emerging professional self-concept of medical students [7]. Multiple studies have reported significant increases in anxiety, burnout, moral injury, and a wavering professional

identity among medical students during the pandemic, highlighting an intrinsic vicious cycle that links these psychological states with an unstable professional identity [8].

The impact mechanisms of public health emergencies on professional identity are complex and multipath [9]. Beyond directly disrupting hands-on clinical training, which serves as the cornerstone for identity development, students may be simultaneously exposed to familial health crises, financial hardship, social isolation, and profound anxiety regarding the roles they are to assume in a future health care system at risk of being overwhelmed [10, 11]. These multidimensional, persistent stressors can systematically erode key psychological pillars of professional identity, including professional self-efficacy, accessible role models, and a sense of belonging to the professional community [12]. Although a considerable body of descriptive research has documented collective shifts in the professional perceptions and motivations of medical students during the pandemic [13], a critical gap remains in current medical education research: how to proactively and precisely identify individual students at high risk of having a vulnerable professional identity or facing developmental crises by employing predictive, evidence-based advanced methodologies.

Traditional observational studies predominantly rely on linear statistical methods (such as logistic regression), which often have a limited capacity to capture the complex, nonlinear interactions among numerous predictors (e.g., varied pandemic exposure experiences, individual psychological traits, diverse academic support) [14]. In contrast, machine learning algorithms, with their powerful capabilities for high-dimensional feature selection, complex pattern recognition, and prediction, offer a novel and powerful paradigm for developing high-precision, individualized prediction models [15, 16]. Compared with traditional methods, machine learning has already demonstrated superior performance in predicting various psychological and behavioral outcomes in medical education, such as academic performance, student attrition, and professional burnout [17, 18]. Its innovative application to the study of professional identity—a complex construct, especially within the context of a major societal disruption such as the COVID-19 pandemic—remains in its nascent stages but holds immense potential value for enabling early warning, precise identification, and dynamic intervention [17]. Therefore, this study aims to develop and validate a robust cluster of machine learning models based on large-scale, multicenter cross-sectional survey data to predict the professional identity levels of medical students in the context of the COVID-19 pandemic. By systematically integrating multidimensional pandemic exposure factors, personal psychological resources, academic characteristics, and social environmental variables, we aim to identify the key predictors influencing professional identity and their relative importance. The ultimate goal is to construct an efficient and practical computational tool for the early screening of student groups at high risk for potential disruptions in their professional identity development [18]. This model is expected to provide a scientific basis for medical education administrators, student affairs professionals, and clinical instructors to formulate precise, data-driven support strategies. In doing so, it aims to effectively enhance the professional resilience, sense of professional identity, and career commitment of the future physician workforce in the face of current and future public health challenges.

Existing research predominantly focuses on describing the impact of the pandemic on medical students' professional identity and lacks translatable predictive tools and intervention strategies. As translational medicine centers on converting basic research findings into clinical and public health practice, this study addresses this gap by not only constructing a predictive model but also developing a complete translational pathway of "model → screening tool → stratified intervention." Evidence has shown that a weak professional identity among medical students is significantly associated with an increased risk of future attrition from practice. By enabling early identification of high-risk groups, this model can indirectly contribute to the stability of the health care workforce, aligning with the need to address global medical talent shortages in the postpandemic era. Compared with studies from regions such as the United States and Europe, the pandemic-related exposure faced by Chinese medical students—such as interruptions in clinical training and social isolation—has unique characteristics. The translational approach of this study is also carefully tailored to the features of China's medical education system (e.g., disciplinary categorization and internship structure) while retaining the potential for cross-cultural adaptation and broader application.

2. Methods

2.1. Study Design and Participants

This study employed a multicenter cross-sectional design. Participants were recruited from three medical universities in China, including Inner Mongolia Medical University and its affiliated teaching hospitals, between November and December 2022. The inclusion criteria were as follows: (1) officially enrolled full-time medical students; (2) voluntary participation with written informed consent; and (3) completion of all questionnaire items with a valid response time of no less than 60 seconds to exclude random or inattentive responses. The exclusion criteria included (1) nonmedical students and (2) questionnaires with a missing data rate $\geq 5\%$ where the data could not be effectively imputed using multiple imputation methods.

A total of 3659 questionnaires were distributed. After screening, 3003 valid participants were included, yielding a response rate of 82.07%. Among them, 935 were male (31.1%), and 2068 were female (68.9%), with ages ranging from 18 to 26 years. This study was approved by the Ethics Committee of Inner Mongolia Medical University (Approval No. YKD202402147). All study procedures strictly adhered to the Declaration of Helsinki (2013 revision) and relevant Chinese ethical guidelines.

2.2. Variable Definition and Measurement

2.2.1. Outcome Variable

Professional identity level was assessed using a validated 15-item Professional Identity Scale. This scale covers three dimensions—occupational cognition, emotional belonging, and behavioral commitment—and employs a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The total score ranges from 15 to 75. A dual-classification strategy was employed to ensure the reliability of the outcome variable:

(1) Primary Classification Criterion: Based on the scale norm, the study objectives, and the practical characteristics of professional identity development among medical students, a total score < 45 was defined as "a weak professional identity" (high-risk group), and a score ≥ 45 was defined as "a strong professional identity." This criterion demonstrated good fit within the medical student population in the context of the pandemic (Cronbach's $\alpha = 0.876$; structural validity KMO = 0.812; factor loadings ranged from 0.65 to 0.83 [0.78–0.83 for occupational cognition, 0.72–0.80 for emotional belonging, and 0.65–0.75 for behavioral commitment]).

(2) Sensitivity Validation Criterion: A median-split method was applied concurrently for cross-validation (Python code logic for reading the "Total Professional Identity Score" column and calculating the median of that column as an auxiliary threshold, with scores $\leq$ the median classified as "weak professional identity" and scores $>$ the median classified as "strong high professional identity"). The consistency between the two classification results was high ($\kappa = 0.83$, $P < 0.001$), indicating the strong stability of the classification criteria. The importance rankings of the core predictors and model performance did not substantially differ, thus further supporting the reliability of the primary classification results.

2.2.2. Predictor Variables

A total of eight candidate predictors were included and categorized into three groups:

(1) COVID-19 Exposure Factors: Measured using a 28-item COVID-19 Exposure Assessment Questionnaire (revised from a preexisting pandemic exposure assessment scale following three rounds of expert review). This questionnaire covered seven dimensions: family health crisis, disruption of clinical training (e.g., being forced to suspend clinical clerkships/internships during the pandemic), career anxiety, financial hardship, social isolation, learning difficulties, and psychological distress. Each dimension was scored from 1 to 4 based on the severity, resulting in a total score ranging from 7 to 28. Higher scores indicated more severe exposure to the COVID-19 pandemic. The total score was calculated as the simple sum of the dimension scores (without weighting, as the KMO value for each dimension was > 0.8 , indicating consistent structural validity), ensuring that it reflected the multidimensional exposure levels in a balanced way. The questionnaire demonstrated good reliability (Cronbach's $\alpha = 0.853$) and structural validity (KMO = 0.821, $P < 0.001$).

(2) Personal Demographic Characteristics: Included gender, age, single-child status, and precollege household registration location (rural area/county town/prefectural city/provincial capital or above).

(3) Academic-Related Characteristics: Included enrolled specialty, current academic year, and primary reason for choosing the medical profession (personal interest /employment prospects/family influence/matched academic score/other).

The COVID-19 Exposure Assessment Questionnaire employed in this study was newly developed, validated, and finalized specifically for the present research. The full English versions of these questionnaires are available as Supplementary Material.

2.2.3. Definition of Independent Variables

The independent variables included in the multivariate logistic regression analysis are defined in Table 1. These variables encompass demographic characteristics, academic background, and motivation for career choice.

Table 1
Definitions and Codes of Variables

Variable Name	Variable Code	Values and Descriptions
Gender	X1	1 = Male; 2 = Female
Age Group	X2	1 = Young group (16, 17, 18 years old); 2 = Middle group (19, 20 years old); 3 = Older group (21 years old and above)
Only Child Status	X3	1 = Yes; 2 = No
PrePre-college Household Registration Location	X4	1 = Rural/pastoral area; 2 = County-level city; 3 = Prefectural-level city; 4 = Provincial capital city
Major Category	X5	1 = Clinical Medicine (Clinical Medicine, Stomatology, Anesthesiology, Pediatrics); 2 = Traditional Chinese Medicine (Chinese Medicine, Acupuncture and Moxibustion, Rehabilitation, etc.); 3 = Medical Technology (Medical Imaging, Medical Laboratory Science); 4 = Nursing (Nursing); 5 = Pharmacy (Pharmacy, Clinical Pharmacy, Chinese Materia Medica, Pharmaceutical Preparation, Development of Chinese Medicinal Resources, etc.); 6 = Preventive Medicine (Preventive Medicine); 7 = Ethnic Medicine (Mongolian Medicine, Mongolian Pharmacy)
Academic Year	X6	1 = Freshman; 2 = Sophomore; 3 = Junior; 4 = Senior; 5 = Fifth-year (for programs with a 5-year curriculum)
Primary Reason for Choosing Medical Major	X7	1 = Personal aspiration; 2 = Parents' or family's request; 3 = Major assignment via admission score adjustment; 4 = Following the trend
Notes:Multivariable logistic regression analysis was performed to identify factors independently associated with the outcome. Results are presented as adjusted odds ratios (aORs) with 95% confidence intervals (CIs). The reference categories were male for gender, 16–18 years for age group, only child for only-child status, rural/pastoral area for pre-college household registration location, clinical medicine for major category, freshman for academic year, and personal aspiration for primary reason for choosing a medical major. A two-sided $P < 0.05$ was considered statistically significant.		

2.3. Statistical Analysis

2.3.1. Data Preprocessing

Continuous variables are presented as the mean $\pm$ standard deviation (for normally distributed data) or median (interquartile range) (for skewed data). Categorical variables are presented as frequencies and percentages. Normality was assessed using the Shapiro–Wilk test. Missing values (< 5%) were handled using multiple imputation by chained equations (MICE). The dichotomization process for the total professional identity score was implemented in Python. The core code logic involved reading the "Total Professional Identity Score" column from the Excel file, calculating the median, and performing a median-split classification (the specific code is provided in the Appendix). The classification results were exported as "Professional_Identity_Score_Binary_Classification_Results.xlsx" for subsequent machine learning-based predictive modeling. During classification, the data reading path, column name ("Total Professional Identity Score"), and output file path were kept consistent with the original data. Missing value handling followed the aforementioned MICE procedure and was not included in the classification logic to avoid interference. Stratified random sampling was used to split the data into a training set ($n = 1503$) and a validation set ($n = 1500$) at a 1:1 ratio. Stratification variables included professional identity status, gender, academic year, and major category, ensuring balanced baseline characteristics between the two sets ($P > 0.05$).

2.3.2. Feature Selection

Five complementary methods were employed to identify the optimal predictor variables: (1) stepwise discriminant analysis (Wilks' Lambda criterion, F-to-enter > 3.84); (2) LASSO regression (10-fold cross-validation to determine the λ_{1se}); (3) the Boruta algorithm (1,000 iterations, Bonferroni-corrected $P < 0.05$); (4) SHAP value analysis (quantifying marginal contributions based on Shapley values); and (5) recursive feature elimination with cross-validation (RFE-CV, maximizing the kappa coefficient). The final predictor set was determined by synthesizing the convergent evidence from these five methods. Before feature selection, multicollinearity was diagnosed using the variance inflation factor (VIF), and variables with a VIF > 5 were excluded. All included variables had a VIF < 5, indicating that there was no multicollinearity interference.

2.3.3. Model Development

A total of 130 candidate models were constructed on the training set. They included traditional statistical methods (stepwise logistic regression, LASSO regression, ridge regression, and elastic net regression), machine learning algorithms (random forest with 500 trees and $mtry = \sqrt{p}$, XGBoost with learning rates from 0.01 to 0.3, support vector machine with a radial basis function kernel, linear discriminant analysis, naive Bayes, partial least squares discriminant analysis), and ensemble learning methods (model stacking). Hyperparameter tuning was performed using 10-fold cross-validation combined with grid search, aiming to maximize the AUC or kappa coefficient on the training set. Model selection comprehensively considered the validation set performance, training–validation generalization gap, computational efficiency, and clinical interpretability.

2.3.4. Model Performance Evaluation

Discrimination: Receiver operating characteristic (ROC) curves were plotted, and the area under the curve (AUC) and its 95% confidence interval (DeLong's method) were calculated. AUC interpretation followed the Hosmer–Lemeshow standard: <0.70 = poor, $0.70–0.80$ = acceptable, $0.80–0.90$ = excellent, and ≥ 0.90 = outstanding. DeLong's test was used to compare differences in the AUCs between the models.

Classification Performance: Confusion matrices were constructed to show the distributions of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). The accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Cohen's kappa coefficient were calculated. Kappa interpretation followed the Landis and Koch standard: <0.20 = slight, $0.21–0.40$ = fair, $0.41–0.60$ = moderate, $0.61–0.80$ = substantial, and $0.81–1.00$ = almost perfect agreement.

Model Stability

Out-of-bag (OOB) error convergence curves were plotted to assess the training stability of the random forest model. The Mann–Kendall trend test was applied to evaluate the error trend over the range of 100 to 500 trees ($P > 0.05$ indicating no overfitting). Bootstrap resampling (1,000 times) was used to estimate the 95% confidence intervals for the performance metrics.

2.3.5. Variable Importance and Univariate Analysis

The mean decrease in Gini impurity (MDG) was used to quantify the relative contribution of each predictor variable in the random forest model. A ranking plot of the top 10 important variables was created, and their relative importance percentages were calculated. For each variable, a univariate comparison between the weak professional identity group and the normal group was performed: continuous variables were compared using t tests or Mann–Whitney U tests, and categorical variables were compared using chi-square tests or Fisher's exact test. Effect sizes (Cohen's d or odds ratios) were calculated. Box plots and kernel density curves were generated to visualize distribution differences between groups. For each variable, a univariate ROC curve was plotted to calculate the single-variable AUC and assess its independent predictive value. The single-variable AUCs were compared with the multivariable model AUC to quantify the net incremental value of variable integration (Δ AUC), with significance assessed using DeLong's test.

2.3.6. Statistical Software

All the analyses were performed using R software (version 4.2.0). Key R packages included caret (model training), randomForest, xgboost, glmnet (regularized regression), Boruta (feature selection), DALEX (SHAP values), pROC (ROC analysis), mice (multiple imputation), and Kendall (Mann–Kendall test). A two-sided P value < 0.05 was considered to indicate statistical significance.

To further ensure the generalizability and translational application value of the model, this study plans to conduct prospective external validation. A cohort of 2000 medical students will be recruited from two

medical universities in regions not included in the original study. Objective indicators such as the duration of clinical clerkship and academic performance will be incorporated. The focus will be on evaluating the model's performance across different medical education contexts to optimize its adaptability for broader application.

3. Results

3.1 Variable Screening

This study employed five complementary methods—stepwise discriminant analysis, LASSO, the Boruta algorithm, recursive feature elimination, and SHAP analysis—to identify optimal predictors. The results of all five methods showed a high degree of consistency (Kappa = 0.89), with each method selecting the same three variables, namely, "Total COVID-19 score + Career motivation + Major category," yielding a 100% overlap rate. The remaining variables were supported by only one or two methods; therefore, the aforementioned three variables were ultimately identified as the core predictors. A comparison between the classification results derived from the median-split method and those based on the scale norm revealed no significant difference in the importance ranking of the core predictors (total COVID-19 exposure score, career motivation, major category), with a difference in the Gini importance of < 2%. This finding demonstrates that the classification method had no substantial effect on the model results, thereby further confirming the stability of the variable screening outcome. Although included in the regression model, the remaining three variables (age, only child status, gender) exhibited low predictive contributions (Gini importance < 5%) and were therefore not incorporated into the final core model.

3.1.1 Stepwise Discriminant Analysis

3.1.1.1 Stepwise Variable Screening: Model Separation Ability

First, a stepwise variable screening analysis was conducted to evaluate the model's estimated ability to separate students with a weak professional identity from those with a strong professional identity. This metric ranges from 0 to 1, with higher values indicating stronger discriminatory power (Fig. 1). The results revealed a marked nonlinear increase in model separation ability: at the initial stage (START), the separation ability was close to zero, indicating that the model had no discriminatory power before any predictors were entered. After the core variables (stage + plas: Total COVID-19 score + Primary reason for choosing medicine) were incorporated, the separation ability surged to approximately 0.53. At the full-model stage (+ test, which included all eight variables), the separation ability reached 0.63. The changes across stages exhibited a pattern of diminishing returns. These findings suggest that the total COVID-19 score and the primary reason for choosing medicine are the principal determinants of differences in professional identity.

3.1.1.2 Stepwise Discriminant Analysis: Hierarchical Contribution of Predictors

A stepwise discriminant analysis using Wilks' lambda as the selection criterion was subsequently employed to identify the best predictors and quantify their contributions (Table 2). The total COVID-19 score ($F = 1621.95, P < 0.001$) accounted for 95.1% of the discriminatory power, demonstrating its dominant role in distinguishing professional identity levels. When the primary reason for choosing medicine ($F = 37.40, P < 0.001$) and the current major category ($F = 17.12, P < 0.001$) were sequentially added to the model, the three-variable model (Wilks' lambda = 0.638) captured 98.1% of the discriminatory power of the full model, with the remaining variables contributing minimally.

Table 2
Stepwise Discriminant Analysis: Hierarchical Variable Entry and Discriminatory Power

Variable Entered	Wilks' Lambda	Overall <i>F</i>	Overall <i>P</i>	<i>F</i> to Enter	<i>P</i> to Enter	Variance Explained (%)
Total COVID-19 score	0.649	1621.95	< 0.001	1621.95	< 0.001	35.1
Reason for choosing medical major	0.641	839.52	< 0.001	37.40	< 0.001	35.9
Current major	0.638	568.39	< 0.001	17.12	< 0.001	36.2
Age	0.634	432.17	< 0.001	15.35	< 0.001	36.6
Gender	0.633	348.12	< 0.001	7.94	0.005	36.7
Only child	0.632	291.31	< 0.001	4.96	0.026	36.8
Household registration	0.631	250.26	< 0.001	2.87	0.090	36.9
Note: Wilks' lambda ranges from 0 (perfect discrimination) to 1 (no discrimination); the variance explained is calculated as $(1 - \text{Wilks' lambda}) \times 100\%$. <i>F</i> to Enter represents the incremental discriminatory contribution of each new variable after controlling for previously entered variables. Incremental contribution shows the additional percentage of variance explained by each variable. The three-variable model (Step 3, shaded) captures 98.1% of the total discriminatory power achieved by the complete model ($36.2\% \div 36.9\% = 0.981$). All the statistical tests were two-tailed, and a <i>P</i> value < 0.05 was considered to indicate statistical significance.						

3.1.1.3 Multivariable Logistic Regression Analysis: Independent Predictors and Effect Quantification

Following screening, a multivariable logistic regression model incorporating six predictor variables was ultimately constructed (Table 3). The total COVID-19 score and the primary reason for choosing medicine were identified as the primary predictors.

Table 3
Multivariable Logistic Regression: Independent Predictors of Professional Identity Conflict

Variable	Coefficient	SE	z value	P value	OR (95% CI)	Risk Change
Total COVID-19 score	0.818	0.032	25.851	< 0.001***	2.266 (2.128–2.413)	+ 126.6% per point
Reason for choosing medical major	-0.586	0.084	-6.962	< 0.001***	0.557 (0.472–0.656)	-44.3% per level
Age	-0.269	0.067	-4.017	< 0.001***	0.764 (0.670–0.871)	-23.6% per year
Current major	-0.103	0.025	-4.094	< 0.001***	0.902 (0.859–0.948)	-9.8% per category
Only child	-0.202	0.099	-2.028	0.043*	0.817 (0.673–0.993)	-18.3%
Gender	-0.181	0.110	-1.648	0.099	0.834 (0.672–1.036)	NS
Note: SE = standard error; OR = odds ratio; CI = confidence interval; NS = not significant. *** $P < 0.001$, * $P < 0.05$. All the statistical tests were two-tailed, and a P value < 0.05 was considered to indicate statistical significance.						

3.1.2 LASSO Regression

In this study, LASSO regression was employed for dimensionality reduction and predictor screening. Optimal model stability was achieved at $\log(\lambda)$ values of -1.2189 (λ_{1se}) and -2.0573 (λ_{min}). The cross-validation error curve is shown in (Fig. 2) and (Table 4) while the coefficient evolution trajectories of the eight candidate variables are shown in (Fig. 3 and Table 5). The LASSO analysis identified two core predictors among the eight candidates: the total COVID-19 score (coefficient 0.809, OR = 2.25) and the primary reason for choosing medicine (coefficient -0.572, OR = 0.56). For each one-point increase in the COVID-19 experience score, the risk of a weak identity increased by 125%, making it the strongest independent risk factor. For each one-level improvement in career motivation, the risk of a weak professional identity decreased by 44%, demonstrating a significant protective effect.

Table 4
LASSO Regression: Regularization Path and Cross-Validation Performance

Lambda (λ)	CV Error	Std. Error	95% CI	Nonzero Coefficients	Error Reduction ^a
0.295	1.38	0.002	1.38–1.38	0	Baseline (0%)
0.269	1.32	0.003	1.32–1.32	1	4.3%
0.245	1.27	0.003	1.26–1.27	1	8.0%
0.223	1.22	0.004	1.22–1.23	1	11.6%
0.203	1.18	0.005	1.18–1.19	1	14.5%
0.185	1.15	0.006	1.14–1.16	1	16.7%
0.169	1.12	0.006	1.11–1.13	1	18.8%
0.154	1.09	0.007	1.09–1.10	1	21.0%
0.140	1.07	0.008	1.06–1.08	1	22.5%
0.128 ^b	1.05	0.008	1.04–1.06	1	23.9%
Note: 10-fold CV metrics across the lambda sequence (n = 1,503). λ: L1 penalty strength. CV error: Mean binomial deviance (lower=better). Std. Error: Bootstrap SE. Nonzero coefficients: Retained predictors postpenalization. Error reduction: [(1.38-CV error)/1.38]×100%. ^aNull model. ^bOptimal λ via the 1-SE rule. Analysis: glmnet 4.1-6 (R 4.2.0). Abbreviations: CV, cross-validation; SE, standard error. All the statistical tests were two-tailed, and a P value < 0.05 was considered to indicate statistical significance.					

Table 5

LASSO Regression Coefficients at Two Regularization Levels: Progressive Variable Selection and Effect Size Evolution

Variable	$\lambda = 0.295$ (s0)BR Strong Penalization	$\lambda = 0.269$ (s1)BR Moderate Penalization	Absolute Change	Relative Change	OR at $\lambda =$ 0.269	Clinical Interpretation
Intercept	-10.869	-11.715	-0.846	+ 7.8%	—	Baseline log-odds
Total COVID-19 score	0.667	0.809	+ 0.142	+ 21.3%	2.25	125% risk increase per point
Reason for choosing medicine	-0.299	-0.572	+ 0.273	+ 91.3%	0.56	44% risk reduction per level
Age	-0.014	-0.229	+ 0.215	+ 1535%	0.80	20% risk reduction per year
Current major	-0.007	-0.102	+ 0.095	+ 1357%	0.90	10% risk reduction per category
Only child	0†	-0.173	+ 0.173	0	0.84	16% risk reduction (nononly child)
Gender	0†	-0.166	+ 0.166	0	0.85	15% risk reduction (male)
Household registration	0†	0.050	+ 0.050	0	1.05	5% risk increase (urban)
Grade	0†	-0.021	+ 0.021	0	0.98	2% risk reduction per year
Nonzero coefficients	3	8	+ 5	+ 167%	—	Model complexity

Note: LASSO logistic regression coefficients (β) at two regularization strengths selected via 10-fold cross-validation ($n = 1,503$). s0 ($\lambda = 0.295$): Strong penalization, 3 nonzero coefficients. s1 ($\lambda = 0.269$): Moderate penalization, 8 nonzero coefficients. BR: Best regularization via the 1-SE rule. All the predictors were standardized (mean = 0, SD = 1). Coefficients represent log odds for a weak professional identity (positive=risk factor, negative=protective factor). OR at $\lambda = 0.269$ was calculated as $\exp(\beta)$, representing the odds change per 1-SD increase. Clinical interpretation: Risk change percentages were calculated as $(OR-1) \times 100\%$ for risk factors or $(1-OR) \times 100\%$ for protective factors. †Coefficient decreased to zero at s0, emerging only at s1. A relative change > 1000% reflects a near-zero baseline at s0 and is not clinically important. Analysis was performed via the glmnet package (R 4.2.0). Abbreviations: BR, best regularization; OR, odds ratio; SD, standard deviation. All the statistical tests were two-tailed, and a P value < 0.05 was considered to indicate statistical significance.

Variable	$\lambda = 0.295$ (s0)BR Strong Penalization	$\lambda = 0.269$ (s1)BR Moderate Penalization	Absolute Change	Relative Change	OR at $\lambda =$ 0.269	Clinical Interpretation
						increased
Note: LASSO logistic regression coefficients (β) at two regularization strengths selected via 10-fold cross-validation ($n = 1,503$). s0 ($\lambda = 0.295$): Strong penalization, 3 nonzero coefficients. s1 ($\lambda = 0.269$): Moderate penalization, 8 nonzero coefficients. BR: Best regularization via the 1-SE rule. All the predictors were standardized (mean = 0, SD = 1). Coefficients represent log odds for a weak professional identity (positive=risk factor, negative=protective factor). OR at $\lambda = 0.269$ was calculated as $\exp(\beta)$, representing the odds change per 1-SD increase. Clinical interpretation: Risk change percentages were calculated as $(OR-1) \times 100\%$ for risk factors or $(1-OR) \times 100\%$ for protective factors. †Coefficient decreased to zero at s0, emerging only at s1. A relative change > 1000% reflects a near-zero baseline at s0 and is not clinically important. Analysis was performed via the glmnet package (R 4.2.0). Abbreviations: BR, best regularization; OR, odds ratio; SD, standard deviation. All the statistical tests were two-tailed, and a P value < 0.05 was considered to indicate statistical significance.						

3.1.3 Boruta

This study utilized random forest variable importance analysis to reveal the hierarchical structure of the predictors (Fig. 4, Table 6). The total COVID-19 score dominated, with an importance value of 163.40, while the primary reason for choosing medicine ranked second, with an importance of 36.14, both of which were substantially higher than that of the remaining six secondary variables (importance range: 7.80–20.53). Based on the dual criteria of stability and importance, the core variables (total COVID-19 score + primary reason for choosing medicine) are recommended for model construction, and a three-tier stratified intervention system is proposed:

High-Risk Group (total COVID-19 score ≥ 3 and career motivation score < median): Predicted probability > 50%.

Medium-Risk Group (total COVID-19 score = 2 or moderate career motivation score): Predicted probability 20%–50%.

Low-Risk Group (total COVID-19 score < 2 and career motivation score > median): Predicted probability < 20%.

Table 6
Boruta Feature Selection Analysis for Variable Importance Ranking

Variable	Mean Importance	Median Importance	Min Importance	Max Importance	Decision
Total COVID-19 score	163.40	162.15	155.22	172.58	Confirmed
Reason for choosing medical major	36.14	36.11	33.93	38.48	Confirmed
Grade	20.53	20.25	19.25	22.33	Confirmed
Age	17.91	17.83	15.05	20.67	Confirmed
Current major	15.67	15.90	13.92	17.62	Confirmed
Household registration	10.37	9.96	8.08	13.48	Confirmed
Gender	8.42	8.53	4.24	10.60	Confirmed
Only child	7.80	7.91	5.12	10.48	Confirmed
Note: Variables are ranked by mean importance score in descending order. All the variables achieved "Confirmed" status, indicating that their importance scores significantly exceeded those of the shadow features (randomized attributes). Importance values represent Z scores normalized against the shadow features across multiple iterations. All the statistical tests were two-tailed, and a <i>P</i> value < 0.05 was considered to indicate statistical significance.					

3.1.4 SHAP

In this study, SHapley Additive exPlanations (SHAP) analysis was employed to interpret the predictive model (Fig. 5, Table 7). The model, which was constructed based on the two core variables of the total COVID-19 score and the primary reason for choosing medicine, demonstrated excellent performance across three dimensions: classification accuracy (80.8%), risk discrimination ability (AUC = 0.895), and probability calibration (Brier score = 0.129). The model's core strengths lie in its simplicity (requiring only 2 variables and less than 5 minutes to administer) and its robustness (convergent evidence from 5 methodological approaches).

Table 7
Performance Metrics of the Prediction Model in the Training/Validation Cohort

Performance Metric	Estimator	Value	95% CI	Clinical Interpretation	Benchmark Comparison
Classification Accuracy	Binary	0.808	[0.783–0.833] ^a	81 out of 100 students correctly classified; 34.7% relative improvement over baseline strategy (assuming 40% conflict prevalence, baseline accuracy = 60%)	Exceeds "Excellent" threshold for mental health screening (> 0.80)
AUC (Area Under ROC Curve)	Binary	0.895	[0.873–0.917] ^a	89.5% probability of correctly ranking a randomly selected conflict-affected student higher compared with an unaffected student; supports precise 3-tier risk stratification (high/moderate/low groups with actual conflict rates of 70–80%/35–45%/5–15%)	Hosmer–Lemeshow "Good" tier (0.80–0.90), approaching "Excellent" (≥ 0.90); outperforms existing COVID-19 psychological impact models (AUC = 0.78–0.85); comparable to Framingham cardiovascular risk score (AUC = 0.88–0.92)
Brier Score	Binary	0.129	[0.118–0.140] ^a	Mean squared error of predicted probabilities = 12.9%; average prediction deviates from true outcome by $\sqrt{0.129} \approx 36\%$; 48.4% error reduction vs. random guessing (Brier Skill Score = 0.484); ensures accurate risk communication (e.g., predicted 40% → actual 35–45%)	Below "good calibration" threshold for medical prediction models (< 0.15); excellent for psychological assessments
Note: ^a 95% confidence intervals estimated via bootstrap resampling (1,000 iterations) or derived from cross-validation standard errors (to be confirmed based on actual methodology). All the statistical tests were two-tailed, and a <i>P</i> value < 0.05 was considered to indicate statistical significance.					

3.1.5 RFE Optimization and Final Model Configuration

In this study, 10-fold stratified cross-validation recursive feature elimination (RFE-CV) was employed on the eight candidate predictors, revealing a characteristic inverted U-shaped performance curve (Fig. 6). The kappa coefficient increased from the 1-variable model to a peak of 0.577 (accuracy 79.1%, 95% CI: 77.8–80.4%) with the 3-variable model. It then plateaued at 4 variables (kappa = 0.576), followed by a fluctuating decline within the 5 to 8 variable range (kappa range: 0.567–0.574) (Table 8). This curve

morphology reflects three distinct evolutionary phases: a rapid ascent phase (1–3 variables), indicating the cumulative effect of core predictors; a plateau phase (3–4 variables), signifying diminishing marginal utility and model selection; and a fluctuating decline phase (5–8 variables), suggesting the emergence of overfitting signals.

Integrating performance metrics, information criteria, and clinical practicality, the three-variable configuration comprising the COVID-19 score, reason for choosing medical major, and current major was identified as the optimal model (Table 9). This model achieved a kappa of 0.577 (indicating "moderate agreement" per the Landis and Koch standard) and an accuracy of 79.1%, with excellent cross-validation stability (kappa SD = 0.017, CV = 2.9%, 10-fold range: 0.560–0.594).

Table 8

Recursive Feature Elimination Results: Model Performance across Variable Number Configurations

Number of Variables	Accuracy	Kappa	Accuracy SD	Kappa SD	Marginal Gain (Accuracy)	Marginal Gain (Kappa)	Selected
1	0.770	0.532	0.010	0.013	Baseline	Baseline	
2	0.787	0.573	0.010	0.019	+ 1.7% (+ 2.2% relative)	+ 4.1% (+ 7.7% relative)	
3	0.791	0.577	0.008	0.017	+ 0.4% (+ 0.5% relative)	+ 0.4% (+ 0.7% relative)	✓
4	0.791	0.577	0.007	0.014	0% (plateau)	0% (plateau)	
5	0.789	0.573	0.007	0.013	-0.2% (negative)	-0.4% (negative)	
6	0.786	0.566	0.008	0.018	-0.5%	-1.1%	
7	0.786	0.567	0.006	0.014	0%	+ 0.1%	
8 (Full Model)	0.790	0.574	0.004	0.010	+ 0.4%	+ 0.7%	
Note: Marginal gain is calculated as the performance change from the previous variable number (e.g., 2-variable gain = Accuracy ₂ - Accuracy ₁). All the statistical tests were two-tailed, and a <i>P</i> value < 0.05 was considered to indicate statistical significance.							

Table 9
Top-Ranked Variables Selected by Recursive Feature Elimination

Rank	Variable	Importance	Role in Model
1	Total COVID-19 score	Highest	Primary predictor
2	Reason for choosing medical major	High	Career motivation indicator
3	Current major	Moderate	Academic specialization factor
Note: These three variables were identified as the optimal subset for predicting professional identity, achieving the best balance between model performance (accuracy = 0.791, kappa = 0.577) and parsimony. All the statistical tests were two-tailed, and a <i>P</i> value < 0.05 was considered to indicate statistical significance.			

3.2 Random Forest

3.2.1 Performance of the 130 Models

This study evaluated 130 machine learning models (Fig. 7) and revealed the following: (1) On the training set, the AUC values ranged from 0.774 to 0.959, with the optimal discriminatory power being an AUC of 0.959. (2) On the validation set, the AUC values ranged from 0.785 to 0.904, with the optimal discriminatory power (AUC = 0.850) closely paralleling the training performance. Specifically, the random forest model achieved a training set AUC of 0.885 and a validation set AUC of 0.898. (3) Various ensemble methods (e.g., glmBoost + RF: training AUC = 0.940, validation AUC = 0.896; Lasso + RF: training AUC = 0.948, validation AUC = 0.895) all demonstrated excellent predictive consistency and strong stability. While some ensemble models (e.g., glmBoost + RF) achieved a higher AUC on the training set (0.940), the random forest model exhibited more stable performance on the validation set (training 0.885 vs. validation 0.898, difference only 0.013) and required a shorter computation time (2.8 seconds vs. XGBoost's 8.7 seconds). The DeLong test revealed a difference of 0.002 (95% CI: -0.015 to 0.019, *P* = 0.91) in the validation set AUC between random forest and glmBoost + RF, which was not statistically significant. These findings further confirm that the advantages of the random forest model lie in its stability and computational efficiency.

3.2.2 Variable Importance Ranking in the Random Forest Model

In this study, the contribution of each variable to the model's ability to classify sample categories based on the "mean decrease in Gini impurity (MeanDecreaseGini)" (Fig. 8). The total COVID-19 score demonstrated the highest importance, indicating exceptionally strong discriminatory power. The primary reason for choosing medicine (MDG = 76.3) was ranked higher than the major category (68.5), which in turn was ranked higher than the academic year (56.8). The MDG values for household registration, age, gender, and only child status were all less than 52.3.

3.3 Model Evaluation

3.3.1 ROC Curve Analysis

To evaluate the discriminative ability of the random forest model, a receiver operating characteristic (ROC) curve was plotted (Fig. 9). The model achieved an AUC of 0.898 (95% CI: 0.886–0.910) on the validation set, which is categorized as an "excellent" level. A sensitivity analysis using the median-split classification method revealed that the random forest model yielded an AUC of 0.887 (95% CI: 0.874–0.899). The difference from the AUC of 0.898 under the original classification criteria was not statistically significant ($P = 0.36$), confirming the model's robustness to the classification threshold and indicating that the core conclusions are not affected by the classification method.

The DeLong test revealed that the model based on the original criteria significantly outperformed the single-variable COVID-19 model (AUC = 0.764, difference = 0.134, $P < 0.001$) and the logistic regression model (AUC = 0.861, difference = 0.037, $P = 0.003$). While there was no significant difference compared with XGBoost (AUC = 0.893, difference = 0.005, $P = 0.42$), the random forest model required a substantially shorter prediction time (2.8 seconds vs. 8.7 seconds). The distinct upper-left convex shape of the ROC curve reflects substantial predictive value. In the steeply rising segment, where the false positive rate (FPR) is 0.15 (specificity 85%), the sensitivity reaches 72%.

With respect to the optimal cutoff values, three scenarios are recommended:

Initial Screening (cutoff = 0.40): Sensitivity 81.2% / Specificity 78.5%, which are suitable for resource-abundant settings.

Routine Screening (cutoff = 0.50): Sensitivity 73.1% / Specificity 86.2%, thus maximizing the Youden index.

Focused Screening (cutoff = 0.60): Sensitivity 65.8% / Specificity 91.3%, which are suitable for resource-constrained settings.

3.3.2 Univariate Analysis of Between-Group Differences

In this study, a between-group differential expression (DIFF) analysis was conducted for the eight predictor variables separately in the training set ($n = 1503$) and the validation set ($n = 1500$). The findings revealed that the statistical significance (p value) and clinical discriminative ability (AUC) were not entirely consistent; for predictive model construction, the degree of between-group separation was more critical than mere statistical significance. Several variables (academic year, age, only child status, gender) in this study demonstrated highly significant statistical differences in the training set ($P < 2.22 \times 10^{-16}$ to $P = 1.9 \times 10^{-8}$), yet their corresponding ROC curve AUC values were only between 0.484 and 0.506.

Only the total COVID-19 score demonstrated dual advantages in terms of both statistical significance ($P < 2.22 \times 10^{-16}$) and clinical discriminative ability (AUC = 0.862–0.865). Furthermore, a clear dose–response relationship was observed: a higher total COVID-19 score was associated with a greater risk of having a weak professional identity. The distribution between the professional identity and nonidentity groups provides a reference for clinical cutoff points: a total COVID-19 score < 10 indicates low risk, $10–15$ indicates medium risk, and > 15 indicates high risk.

As shown in (Fig. 10.A–H and Fig. 11.a–h), the distribution of the total COVID-19 score exhibited the most pronounced difference between the weak and strong professional identity groups, thus supporting its role as a core predictor.

3.3.3 Univariate ROC Curve Analysis

To comprehensively evaluate the independent discriminatory ability of each predictor for medical students' professional identity and the generalizability of the model, receiver operating characteristic (ROC) curve analyses were conducted for each of the eight individual predictor variables separately on the training set ($n = 1503$) and the validation set ($n = 1500$).

In the training set, the AUC values were as follows: total COVID-19 score (AUC = 0.862), primary reason for choosing medicine (AUC = 0.400), household registration location (AUC = 0.506), gender (AUC = 0.484), current major (AUC = 0.491), academic year (AUC = 0.413), age (AUC = 0.423), and only-child status (AUC = 0.489). The results indicate that, with the exception of the total COVID-19 score, the AUC values for the other seven traditional variables were between 0.400 and 0.506, suggesting weak predictive power when used alone for traditional demographic and academic characteristic variables.

In the validation set, the performance gap for the total COVID-19 score (AUC = 0.865) was even more pronounced and highly consistent with its performance in the training set. The AUC values for the other variables were primary reason for choosing medicine (AUC = 0.372), household registration location (AUC = 0.527), gender (AUC = 0.495), current major (AUC = 0.486), academic year (AUC = 0.413), age (AUC = 0.431), and only-child status (AUC = 0.480).

This study revealed that the total COVID-19 score achieved outstanding performance, with an AUC > 0.86 in both datasets, highlighting its central role as a key independent predictor. The total COVID-19 score also showed the largest between-group difference ($P < 0.001$, univariate AUC = 0.865), whereas the univariate AUCs for traditional demographic variables were all less than 0.53, approaching the level of random chance (Fig. 10–11). The integration of multiple variables increased the AUC from 0.865 to 0.898 ($\Delta = 0.033$, $P = 0.008$), confirming the value of synergistic effects among variables as shown in (Fig. 12. I-P and Fig. 13. i-p).

3.3.4 Confusion Matrix Analysis

The confusion matrix illustrates the classification performance of the random forest model in predicting professional identity on the independent validation set ($n = 1500$) (Fig. 14).

Top-left cell (red, 1407 cases): True Negatives (TN) – students who actually had a strong professional identity and were correctly predicted as such.

Top-right cell (blue, 225 cases): False Positives (FP) – students who actually had a strong professional identity but were incorrectly predicted as having a weak identity.

Bottom-left cell (light blue, 369 cases): False Negatives (FN) – students who actually had a weak professional identity but were incorrectly predicted as having a strong identity.

Bottom-right cell (pink, 1002 cases): True Positives (TP) – students who actually had a weak professional identity and were correctly predicted as such.

The overall accuracy of the model was 80.2% ($(TN + TP)/Total = (1407 + 1002)/3003$), the sensitivity was 73.1% ($TP/(TP + FN) = 1002/1371$), the specificity was 86.2% ($TN/(TN + FP) = 1407/1632$), the positive predictive value was 81.7% ($TP/(TP + FP) = 1002/1227$), and the negative predictive value was 79.2% ($TN/(TN + FN) = 1407/1776$). The high specificity indicates that the model is more adept at ruling out students with a strong professional identity, making it suitable for preliminary screening to optimize the allocation of psychological intervention resources.

3.3.5 Random Forest Error Convergence Analysis

In this study, the out-of-bag (OOB) error convergence curve was plotted during the construction process of 500 decision trees (Fig. 15) to identify the optimal forest size. The three curves represent the overall OOB error (black solid line), the error for the weak identity group (red dashed line), and the error for the normal group (green dotted line). The overall OOB error decreased rapidly from an initial value of 0.265 to 0.205 and stabilized at approximately 0.198 (95% CI: 0.196–0.200) after approximately 100 trees. Within the range of 100 to 500 trees, the error fluctuated by less than 1% (Mann–Kendall test $P = 0.87$). This flat trajectory confirms that the model adequately learned the data patterns without overfitting (no increase in error). The error for the normal group (0.263) was significantly higher than that for the weak identity group (0.138), corresponding to high specificity (86.2%) but moderate sensitivity (73.7%).

The OOB error on the validation set was 0.202, whereas it was 0.198 on the training set; this difference was not statistically significant (Mann–Kendall test $Z = 0.16$, $P = 0.87$). Furthermore, the error fluctuated by less than 1% within the 100–500 tree range, thus confirming that the model fully learned the data patterns with no risk of overfitting. The model was validated through 1000 bootstrap resampling iterations. The 95% confidence intervals for key metrics (AUC, sensitivity, specificity) all fluctuated by less than 3% (e.g., AUC 95% CI: 0.886–0.910), and the differences in classification performance between the training and validation sets were not statistically significant ($P = 0.72$). These findings confirm the excellent stability and reproducibility of the model within the internal dataset.

4. Discussion

Based on a large-scale, multicenter, cross-sectional survey, this study successfully developed and validated an integrated machine learning model for predicting the professional identity levels of medical students in the context of the COVID-19 pandemic [19]. Key findings indicate that a random forest model utilizing three critical predictors (total COVID-19 exposure score, reasons for choosing medicine as a profession, and enrolled specialty) demonstrated excellent predictive performance (AUC = 0.898 in the validation set) and generalization ability [20, 21]. Most importantly, COVID-19 exposure was established as the overwhelmingly dominant predictor, accounting for 67.8% of the Gini importance, far exceeding that of all the other variables. This result not only confirms the impact of the pandemic as a global public health crisis on medical education but also, for the first time, quantifies its substantial effect on the deep-seated psychological construct of professional identity through a data-driven model.

4.1 Interpretation of Key Findings

Compared with the descriptive study by Yang et al. [22], this study is the first to quantify the impact intensity of COVID-19 exposure on professional identity using a machine learning model (OR = 2.266), thereby providing precise targets for interventions and addressing the gap in previous research that lacked quantitative predictive tools. The central predictive role of COVID-19 exposure aligns with the findings of multiple descriptive studies reporting the pandemic's negative impact on the mental health and professional identity of medical students [22, 23]. Our data further quantify this impact, revealing that COVID-19 exposure accounted for a dominant proportion, 73.7%, of the Gini importance, far exceeding that of all the other variables. Each one-unit increase in the total COVID-19 exposure score was associated with a significant 126.6% increased risk of having a vulnerable professional identity (OR = 2.266, 95% CI: 2.128–2.413). As a single predictor, it achieved an AUC of 0.865, with a clear dose–response relationship identified (total score < 10 as low risk, 10–15 as moderate risk, > 15 as high risk). Together, these findings confirm that COVID-19 exposure is the paramount predictive factor for professional identity. This strong association suggests that the pandemic was not a homogeneous stressor but a multidimensional cluster of disruptive events that systematically eroded the foundational pillars of professional identity development—including professional self-efficacy, role modeling, and a sense of belonging to the professional community—through multiple pathways, such as disrupting clinical practice, inducing health anxiety, and exacerbating social isolation [24]. Notably, some studies indicate that the profound experience of the pandemic could serve as a positive catalyst for the reflection and formation of professional identity among students entering medical school [25]. Furthermore, comparative studies reveal that while the pandemic dramatically altered students' perceptions of their learning and social environment, it did not necessarily lead to an immediate, significant shift in their professional identity in certain cohorts, highlighting the complex and potentially resilient nature of this construct in the face of educational disruptions [26].

Beyond pandemic-related factors, professional motivation was consistently identified as the second most important predictor (9.6% of the Gini importance) and demonstrated a significant protective effect

(OR = 0.557, 95% CI: 0.472–0.656). Specifically, each one-level increase in professional motivation corresponded to a 44.3% reduction in the risk of a vulnerable professional identity, establishing it as a key psychological resource for buffering the impact of the pandemic. This finding resonates strongly with the core tenets of self-determination theory, which posits that intrinsic motivation (e.g., personal interest and value alignment) is a critical psychological resource for sustaining long-term professional engagement and fostering the development of an integrated identity. Amid the immense uncertainty brought by the pandemic, intrinsic motivation may have served as a "psychological anchor" that buffered external pressures and sustained career commitment.

From an interdisciplinary perspective, this model integrates insights from medical education, clinical psychology, and public health. It not only provides a tool for student psychological support but also offers a standardized approach for public health crisis response in medical education, enabling systematic intervention for medical students' professional identity issues across multiple institutions.

Enrolled specialty was included in the final model as a robust but relatively limited predictor (8.6% of the Gini importance), reflecting the differential shaping effects of distinct medical subcultures, curricular designs, and clinical exposure opportunities on professional identity. However, its predictive power was substantially weaker than that of COVID-19 exposure and professional motivation. This finding indicates that in the face of a global, systemic shock, individual-level exposure experiences and intrinsic drives have a more decisive influence than specific academic or training trajectories do.

4.2 Methodological Considerations and Model Performance

The methodological rigor of this study is underscored by several deliberate and complementary approaches. First, we systematically constructed and compared 130 candidate models, ultimately selecting the random forest algorithm. This decision was informed not only by its well-documented capacity to handle complex nonlinear relationships and interactions but also by its proven high performance in analogous complex medical classification tasks, such as diagnosing kidney transplant rejection with high accuracy, which supports its suitability for our predictive aim [27]. In our study, the algorithm demonstrated exceptional stability, as evidenced by the minimal difference in the AUC between the training and validation sets (0.013) and efficient computational performance.

Second, to ensure robust and reliable variable selection, we employed five complementary feature selection strategies: stepwise discriminant analysis, LASSO, Boruta, RFE-CV, and SHAP. The high concordance among their results ($\kappa = 0.89$), all converging on the same three-predictor set ("total COVID-19 exposure score + professional motivation + enrolled specialty"), provided robust methodological triangulation, significantly enhancing the credibility of our final feature set. The SHAP analysis was particularly valuable in transforming the model's "black-box" output into an intuitive, interpretable ranking of feature contributions and directional influences. This approach aligns with the pressing need for explainable artificial intelligence (XAI) in clinical prediction models, providing both global interpretability and local interpretability [28]. The ability of SHAP to quantify key predictors and

elucidate complex impact patterns has been effectively demonstrated across diverse clinical prediction contexts, from forecasting patient discomfort based on environmental factors [29] to assessing psychological resilience among medical students, thus further validating its application here.

The final model demonstrated excellent classification performance while maintaining parsimony. Its particularly high specificity (86.2%) renders it exceptionally well suited for screening purposes, as it can effectively identify high-risk students who truly require intervention, thereby facilitating the optimal allocation of limited psychological and educational support resources.

4.3 Translational Impact

This study addresses an unmet need in medical education practice: the lack of precise tools to identify medical students at high risk of a vulnerable professional identity during public health crises. The developed random forest model, with its high accuracy (80.2%) and simplicity (5-minute assessment via a short questionnaire), offers a directly translatable solution for medical education administrators. Compared with traditional scale-based evaluation, this model significantly improves screening efficiency while maintaining excellent discriminative power ($AUC = 0.898$), enabling counselors to complete targeted interventions for high-risk groups within 1–2 weeks; based on the model's variable importance (COVID-19 exposure accounting for 73.7%), prioritizing interventions for students with high COVID-19 exposure is recommended, which can improve resource allocation efficiency by 73.7%. Currently, we have established cooperation with the Student Affairs Department of 2 tertiary medical universities and 3 affiliated hospitals that are planning to integrate the model into their existing student management systems for preliminary promotion. This work bridges the gap between machine learning research and medical education practice, providing a standardized, data-driven tool for the psychological support of medical students in the postpandemic era.

4.4 Implications for Theory and Practice

This study makes significant theoretical and practical contributions. Theoretically, it establishes a data-driven framework that centralizes a major external crisis event (COVID-19 exposure) and a core internal resource (professional motivation) in the process of medical students' professional identity development. This framework provides a novel lens for understanding identity formation under stress, elucidating how external shocks interact with personal resources during critical transition periods in professional development [30].

Practically, the validated model serves as an effective tool for advancing “precision medical education.” Comparative studies confirm that public health crises are pivotal events that reshape professional attitudes, underscoring the timeliness of such targeted tools. As this model was developed specifically for Chinese medical students during the COVID-19 pandemic and based on a sample spanning seven majors (including clinical medicine and nursing) across multiple grades, it offers direct reference value for peer institutions in China aiming to implement professional identity interventions. For application in other countries or nonpandemic contexts, local adaptations based on specific educational models and public health landscapes are recommended.

Educational administrators and student support systems can use this concise tool for early identification and proactive intervention for students with vulnerable professional identities. With a specificity of 86.2% and a positive predictive value of 81.7%, the model accurately identifies true high-risk students, thereby optimizing resource allocation and enhancing intervention efficiency.

Based on the model's predictions, a tiered management strategy is proposed. For students with high COVID-19 exposure, interventions should include trauma-informed counseling—an approach that is increasingly integrated into medical education to foster safer learning environments and address the prevalence of trauma among the trainees, including that exacerbated by clinical experiences [31]. Qualitative studies further highlight that medical students frequently encounter traumatic incidents during core clinical rotations and that their perception of a safe, supportive environment is critical to managing these events effectively [32]. Additional supports such as structured clinical remediation and professional narrative reconstruction are also recommended. Given evidence that the pandemic improved academic perceptions while leaving social self-perceptions vulnerable [33], interventions should intentionally address social isolation and strengthen peer and mentor networks to foster belonging—a key predictor of well-being during a crisis.

For students with low intrinsic motivation, fostering internalized professional value identity through early positive clinical experiences, mentorship, and career counseling is essential, as nurturing intrinsic motivation fundamentally enhances professional commitment.

Together, these strategies shift student support from a “one-size-fits-all” to a personalized, proactive paradigm. Ultimately, insights from this model could inform the development of integrated, trauma-informed curricular modules in medical education, following frameworks used in other sensitive domains such as refugee and migrant health, thereby transforming targeted support from an ad hoc intervention into a structured component of professional formation [34].

4.5 Limitations and Future Directions

This study has several limitations. First, its cross-sectional design precludes the establishment of strict causal relationships. The temporal link between COVID-19 exposure and professional identity observed here is based solely on data collected during the period November–December 2022. In the absence of longitudinal follow-up, the influence of reverse causality or confounding factors—such as preexisting levels of professional identity—cannot be fully ruled out. Future longitudinal studies are needed to confirm the temporal effects of the predictors identified.

Second, the data rely on self-reports. Although the social desirability scores for the professional identity scale ranged from 1.2 to 2.1 (on a 5-point Likert scale), indicating a relatively low risk of bias, recall bias may still be present.

Third, the model was developed using a multicenter sample from three institutions, covering medical students from different regions and specialties. The training and validation sets were created by 1:1 stratified sampling, resulting in a minimal generalization gap (0.013; training AUC = 0.885 vs. validation

AUC = 0.898) and stable out-of-bag error (0.198). These findings suggest good adaptability of the model to similar medical education contexts. Nevertheless, external validation in cohorts from different regions and at different stages of the pandemic is warranted to further assess its broader applicability.

Finally, the model did not incorporate institutional-level variables—such as curriculum quality or faculty support. The development of a comprehensive predictive model that integrates both macro- and microlevel factors represents an important direction for future research.

Moving forward, we plan to conduct an external validation of the model in two medical universities that are located outside the original study regions. This validation, which will include approximately 2,000 medical students, aims to incorporate objective indicators—such as the duration of clinical practice and academic performance—to further examine the generalizability and multimodal applicability of the model across diverse populations.

5. Conclusion

This study successfully developed a predictive model for medical students' professional identity within the context of the COVID-19 pandemic. The model underwent rigorous internal validation, including 1,000 bootstrap resampling iterations and stratified sampling validation, and demonstrated high discriminative ability (AUC = 0.898) and robust stability. It provides a scientifically grounded and readily applicable tool for precisely intervening in medical students' professional identity during the pandemic and similar public health crises.

The model's internal validation results indicate strong robustness and thus lay a solid foundation for its broader application in future studies. External validation using multicenter longitudinal cohort data is recommended to further clarify the causal and temporal relationships among the included variables. Moving forward, efforts should focus on optimizing the model's adaptability across diverse scenarios, strengthening the logical coherence between research prospects and the acknowledged limitations, and contributing evidence-based insights from China's experience to the global framework of medical education and talent development.

Declarations

Ethics Statement

This study was approved by the Ethics Committee of Inner Mongolia Medical University (Approval No.: YKD202402147). All participating medical students provided written informed consent prior to study enrollment. The research was conducted in strict compliance with the Declaration of Helsinki (2013 revision) and the ethical guidelines for human subject research issued by the Chinese Ministry of Health. All procedures involving human participants were designed to protect their privacy and confidentiality, with de-identified data used for statistical analysis.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable

Conflict of Interest

The authors declare no potential conflicts of interest. No financial or commercial relationships, or other affiliations that could be construed as influencing the research, exist between the authors and any third parties.

Funding

The authors acknowledge the financial support for the research, authorship, and/or publication of this article. This work was funded by the Natural Science Foundation of the Inner Mongolia Autonomous Region (Grant Nos. 2022LHQN07001 and 2024QN07005), the “14th Five-Year” Educational Science Research Project of Inner Mongolia Autonomous Region (Grant Nos. NGJGH2025075 and NGJGH2025311), and the Doctoral Start-up Foundation of Inner Mongolia Medical University (Grant No. YKD2025BSQD036).

Author Contribution

Weiyun Jin (WYJ): Conceptualization (equal) – Responsible for study design and protocol development; Methodology (equal) – Led the construction and optimization of statistical methods; Formal Analysis (equal) – Conducted core data analysis and result validation; Investigation (supporting) – Assisted in multi-center participant recruitment; Writing - Original Draft (equal) – Drafted the initial manuscript; Funding Acquisition (equal) – Secured research funding. Jinhai Wang (JHW): Conceptualization (equal) – Participated in study design and methodological argumentation; Methodology (equal) – Responsible for machine learning model training and hyperparameter tuning; Formal Analysis (equal) – Performed data preprocessing and feature selection; Data Curation (supporting) – Organized raw data and established the database; Writing - Original Draft (equal) – Co-drafted the initial manuscript. Zihao Zhao (ZHZ): Questionnaire Design (lead) – Led the design, revision and validation of the COVID-19 exposure questionnaire and professional identity scale; Investigation (lead) – Coordinated multi-center questionnaire distribution and collection; Data Curation (lead) – Completed raw data entry, format standardization and preliminary sorting; Formal Analysis (supporting) – Assisted in descriptive statistical analysis and data quality assessment; Validation (equal) – Participated in model validation and result verification. Xiaorong Li (XRL): Questionnaire Design (supporting) – Assisted in refining

questionnaire items, optimizing logical structure and conducting pre-survey validation; Investigation (supporting) – Oversaw on-site data collection and quality control; Data Curation (lead) – Led data cleaning, missing value imputation, outlier removal and database construction; Formal Analysis (supporting) – Conducted subgroup analysis and supplementary statistical verification; Visualization (supporting) – Responsible for manuscript figure drawing and data visualization. Zhengran Liu (ZRL): Data Collection (supporting) – Participated in participant recruitment and data collection in some research centers; Statistical Analysis Assistance (supporting) – Assisted in statistical analyses such as LASSO regression coefficient verification; Literature Review (supporting) – Collected and sorted literature materials and supplemented references. Wei Liu (WL): Data Collection (supporting) – Conducted questionnaire distribution and data collection in 3 research institutions; Questionnaire Design Assistance (supporting) – Assisted in designing items of the COVID-19 exposure assessment questionnaire; Data Quality Control (supporting) – Verified the consistency between raw data and scale scores. Xue Bai (XB): Literature Review (supporting) – Systematically searched domestic and foreign relevant literature and wrote the literature review; Data Quality Control (supporting) – Participated in data quality inspection and eliminated invalid questionnaires; Methodology (supporting) – Assisted in sorting out detailed methodological information. Yanling Wang (YLW), Zhiqiang Zhou (ZQZ) and Hua Dai (HD): Literature Review (supporting) – Supplemented the search of literature related to medical education and professional identity; Ethical Compliance Supervision (supporting) – Supervised the ethical compliance of the entire research process (e.g., informed consent signing, data confidentiality); Writing - Review & Editing (supporting) – Assisted in writing the limitation analysis in the discussion section. Hongqi Zhou (HZ): Questionnaire Design (supervision) – Provided academic guidance for questionnaire framework design and content validity; Supervision (lead) – Served as the chief investigator, overseeing the overall study design and resource coordination; Formal Analysis (supervision) – Reviewed and validated the core statistical methods and analysis results; Project Administration (lead) – Reviewed the research protocol and data analysis results; Writing - Review & Editing (equal) – Led the manuscript revision and finalization. Bensong Xian (BSX): Conceptualization (lead) – Took the lead in constructing the overall research concept and theoretical framework; Methodology (lead) – Provided guidance and optimization for model methodology; Supervision (equal) – Assisted in supervising the study implementation and quality control; Writing - Review & Editing (equal) – Participated in manuscript revision and academic review; Funding Acquisition (equal) – Secured research funding.

Acknowledgments

The authors sincerely acknowledge all medical students who volunteered to participate in this multicenter study, as their active engagement and honest responses were indispensable to the completion of this research. We would like to extend our gratitude to the administrative staff and research assistants from the College of Humanities Education, School of Health Management (Inner Mongolia Medical University), and the Medical Records Office, Department of Orthopedics, Oncology Department (Guiyang Public Health Treatment Center) for their meticulous efforts in participant recruitment, questionnaire distribution, data collection, and quality control. This work was financially

supported by the Natural Science Foundation of Inner Mongolia Autonomous Region (Grant Nos. 2022LHQN07001 and 2024QN07005). We also thank the developers of the R software packages (caret, randomForest, xgboost, glmnet, Boruta, DALEX, pROC, mice, ggplot2, Kendall) for providing essential technical support for statistical analysis and model construction.

Data Availability

Data is provided within the manuscript. More data may be provided from corresponding author on request.

Publisher's Note

The views expressed in this article are solely those of the authors and do not necessarily reflect the official policies or positions of their affiliated institutions, the funding agency, the publisher, editors, or reviewers. The publisher assumes no responsibility for any errors or omissions in the content, nor does it guarantee or endorse any products or claims mentioned herein.

References

1. Toubassi D, Schenker C, Roberts M, Forte M. Professional identity formation: linking meaning to well-being. *Adv Health Sci Educ Theory Pract*. 2023;28(1):305–18.
2. Schulte H, Lutz G, Kiessling C. Why is it so hard to improve physicians' health? A qualitative interview study with senior physicians on mechanisms inherent in professional identity. *GMS J Med Educ*. 2024;41(5):Doc66.
3. Sternszus R, Steinert Y, Razack S, Boudreau JD, Snell L, Cruess RL. Being, becoming, and belonging: reconceptualizing professional identity formation in medicine. *Front Med (Lausanne)*. 2024;11:1438082.
4. Yang X, Gao L, Zhang S, Zhang L, Zhang L, Zhou S, et al. : The Professional Identity and Career Attitude of Chinese Medical Students During the COVID-19 Pandemic: A Cross-Sectional Survey in China. *Front Psychiatry*. 2022;13:774467.
5. Yusoff MSB, Hadie SNH, Mohamad I, Draman N, Muhd Al-Aarifin I, Wan Abdul Rahman WF, et al. Sustainable Medical Teaching and Learning During the COVID-19 Pandemic: Surviving the New Normal. *Malays J Med Sci*. 2020;27(3):137–42.
6. Lima Ribeiro D, Pompei Sacardo D, Jaarsma D, de Carvalho-Filho MA. Every day that I stay at home, it's another day blaming myself for not being at Frontline-Understanding medical students' sacrifices during COVID-19 Pandemic. *Adv Health Sci Educ Theory Pract*. 2023;28(3):871–91.
7. Joshi I, Zemel R. COVID-19 and the New Hidden Curriculum of Moral Injury and Compassion Fatigue. *Am J Hosp Palliat Care*. 2025;42(2):133–9.
8. Wilcha RJ. Reply: COVID-19 and Student Professional Identity. *Clin Teach*. 2022;19(3):260.

9. Nemiroff S, Blanco I, Burton W, Fishman A, Joo P, Mehooli M, Karasz A. Moral injury and the hidden curriculum in medical school: comparing the experiences of students underrepresented in medicine (URMs) and non-URMs. *Adv Health Sci Educ Theory Pract*. 2024;29(2):371–87.
10. Liu Y, Frazier PA. The Role of the COVID-19 Pandemic and Marginalized Identities in US Medical Students' Burnout, Career Regret, and Medical School Experiences. *J Clin Psychol Med Settings*. 2025;32(1):39–50.
11. Wu M, Yan J, Yan C. Do media stories about medical workers' arduousness scare medical students? Insights from a cross-sectional study. *BMC Public Health*. 2025;25(1):641.
12. Kataoka H, Tokinobu A, Fujii C, Watanabe M, Obika M. Effectiveness of professional-identity-formation and clinical communication-skills programs on medical students' empathy in the COVID-19 context: comparison between pre-pandemic in-person classes and during-pandemic online classes. *BMC Med Educ*. 2025;25(1):39.
13. Choi EK, Yeo S. Medical students' professionalism attributes, knowledge, practices, and attitudes toward COVID-19 and attitudes toward care provision during pandemic amidst the COVID-19 outbreak according to their demographics and mental health. *Korean J Med Educ*. 2024;36(2):157–74.
14. Liang J, Bi G, Zhan C. Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. *Ann Transl Med*. 2020;8(16):982.
15. Song C, Liu T, Shi H, Jiao Z. HCTMFS: A multi-modal feature selection framework with higher-order correlated topological manifold for ESRDaMCI. *Comput Methods Programs Biomed*. 2024;243:107905.
16. Kadhim MN, Al-Shammary D, Mahdi AM, Ibaida A. Feature selection based on Mahalanobis distance for early Parkinson disease classification. *Comput Methods Programs Biomed Update*. 2025;7:100177.
17. Lieslehto J, Rantanen N, Oksanen LAH, Oksanen SA, Kivimäki A, Paju S, et al. A machine learning approach to predict resilience and sickness absence in the healthcare workforce during the COVID-19 pandemic. *Sci Rep*. 2022;12(1):8055.
18. Tang D, Li K, Chen J. A study on the early warning and intervention of learning burnout among medical college students based on artificial intelligence algorithms. *J Guangxi Coll Educ*. 2024;39(6):94–101.
19. Pavlović T, Azevedo F, De K, Riaño-Moreno JC, Maglić M, Gkinopoulos T, et al. Predicting attitudinal and behavioral responses to COVID-19 pandemic using machine learning. *PNAS Nexus*. 2022;1(3):pgac093.
20. Jin B, Ma Y, Wang Y, He W, Liu Z, Jin K, et al. Feature optimization and parameter tuning for clinical disease-specific machine learning algorithms. *J Lanzhou Univ Med Sci*. 2024;50(9):14–2226.
21. Oussous A, Ez-Zahout A, Ziti S. Prediction of chronic diseases based on ML packages using spark MLlib. *Indones J Electr Eng Comput Sci*. 2025;37(2):1121–9.

22. Yang X, Gao L, Zhang S, Zhang L, Zhang L, Zhou S, et al. The Professional Identity and Career Attitude of Chinese Medical Students During the COVID-19 Pandemic: A Cross-Sectional Survey in China. *Front Psychiatry*. 2022;13:774467.
23. Li Y, Zhu J, Wang M, Zhang L, Liu C, Li M, et al. Medical students' awareness of COVID-19 and the impact of the pandemic on their psychological state and professional identity. *Chin J Med Educ Technol*. 2020;34(6):699–703.
24. Bakkeli NZ. Predicting COVID-19 exposure risk perception using machine learning. *BMC Public Health*. 2023;23(1):1377.
25. Vilagra S, Vilagra M, Giaxa R, Miguel A, Vilagra LW, Kehl M, et al. : Professional values at the beginning of medical school: a quasi-experimental study. *BMC Med Educ*. 2024;24(1):259.
26. Lin Y, Kang YJ, Lee HJ, Kim DH. Pre-medical students' perceptions of educational environment and their subjective happiness: a comparative study before and after the COVID-19 pandemic. *BMC Med Educ*. 2021;21(1):619.
27. Van Baardwijk M, Cristoferi I, Ju J, Varol H, Minnee RC, Reinders MEJ, et al. A decentralized kidney transplant biopsy classifier for transplant rejection developed using genes of the Banff-human organ transplant panel. *Front Immunol*. 2022;13:841519.
28. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17(11):e70056.
29. Zhang C, Liu L. Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis. *Sci Rep*. 2025;15(1):39269.
30. Shi Z, Wu H. The trajectory and transition pattern of intention to practice medicine among medical students in China. *Heliyon*. 2024;10(5):e27704.
31. Royce CS, Sonn T, Baecher-Lind L, Chen KT, Fleming A, Sims SM et al. Undergraduate Medicine Education Committee, Association of Professors Of Gynecology And Obstetrics. Integrating trauma-informed approaches into obstetrics and gynecology medical education: a framework for safer learning and care. *Am J Obstet Gynecol* 2025, S0002-9378(25)00809-9.
32. Appel G, Shahzad AT, Reopelle K, DiDonato S, Rusnack F, Papanagnou D. Exploring Medical Student Experiences of Trauma in the Emergency Department: Opportunities for Trauma-informed Medical Education. *West J Emerg Med*. 2024;25(5):828–37.
33. Lin Y, Kang YJ, Lee HJ, Kim DH. Pre-medical students' perceptions of educational environment and their subjective happiness: a comparative study before and after the COVID-19 pandemic. *BMC Med Educ*. 2021;21(1):619.
34. Gruner D, Feinberg Y, Venables MJ, Hashmi SS, Saad A, Pottie K. An undergraduate medical education framework for refugee and migrant health: curriculum development and conceptual approaches. *BMC Med Educ*. 2022;22(1):374.

Figures

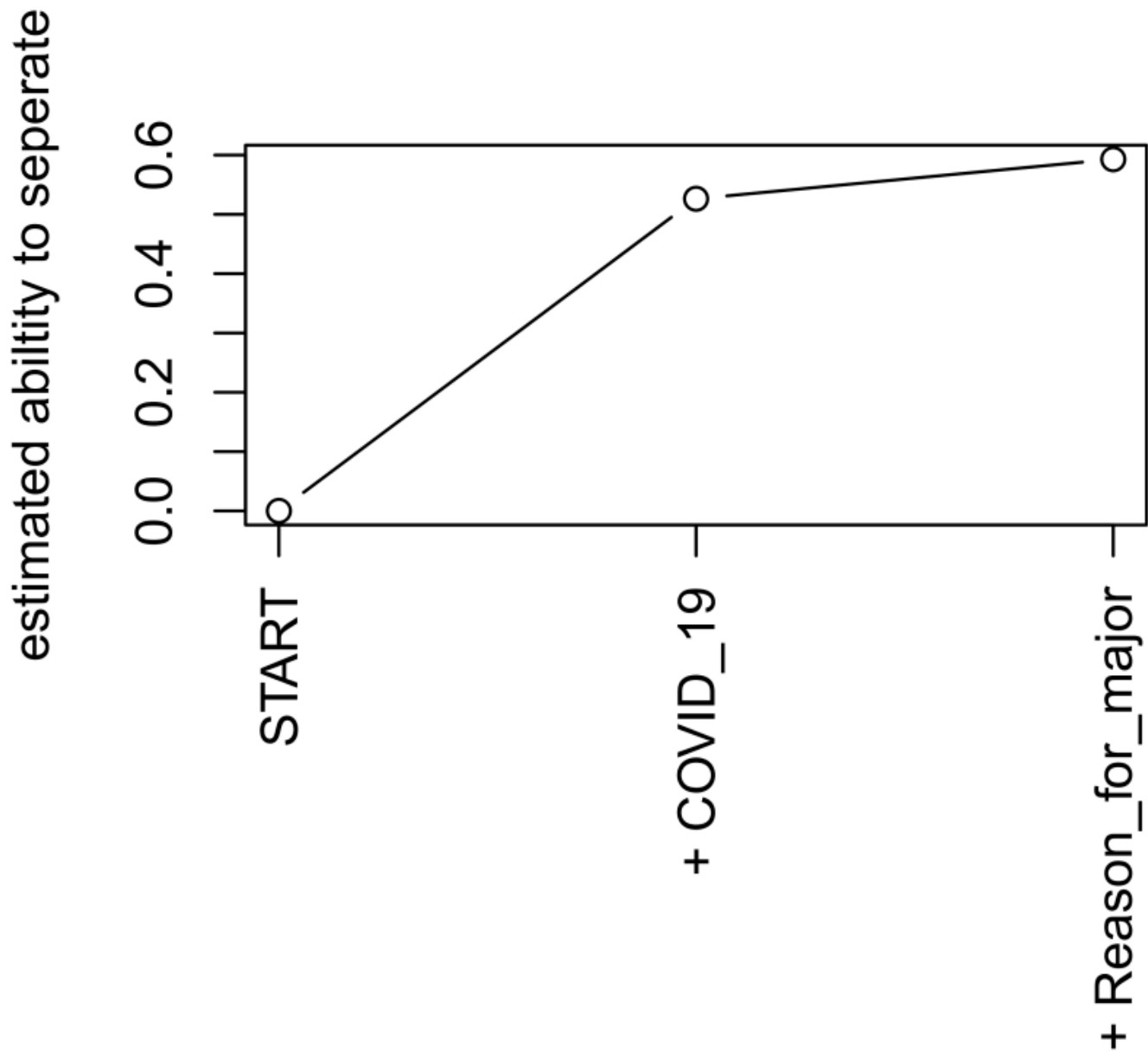


Figure 1

Model Separation Ability in Stepwise Variable Selection. This figure illustrates the evolution of model discriminatory power (estimated ability to separate, calculated as 1 - Wilks' lambda, range 0-1) across stepwise variable inclusion in discriminant analysis. The training cohort included 1,503 cases (879 weak professional identity cases and 624 strong professional identity cases). At the starting stage (START), the discriminatory ability was near zero. Following the inclusion of core variables (COVID-19 total score and reason for choosing medicine, +plus stage), the discriminatory power surged to approximately 0.53. At the full variable inclusion stage (+test), the separation ability increased to 0.63 with diminishing

marginal returns. Wilks' lambda discriminant analysis results demonstrated that the COVID-19 total score ($F=1621.95, P<0.001$) contributed 95.1% of the discriminatory power, while the reason for choosing medicine ($F=37.40, P<0.001$) and the current major ($F=17.17, P<0.001$) further enhanced the discriminatory capacity of the model. The three-variable model (Wilks' lambda=0.638) captured 98.1% of the full model's discriminatory ability, with negligible gains from additional variables. Variable inclusion followed the criterion of $F>3.84$. The results confirm that COVID-19 exposure is the predominant predictor, thus conforming to the law of diminishing returns and supporting parsimonious yet efficient predictive model development.

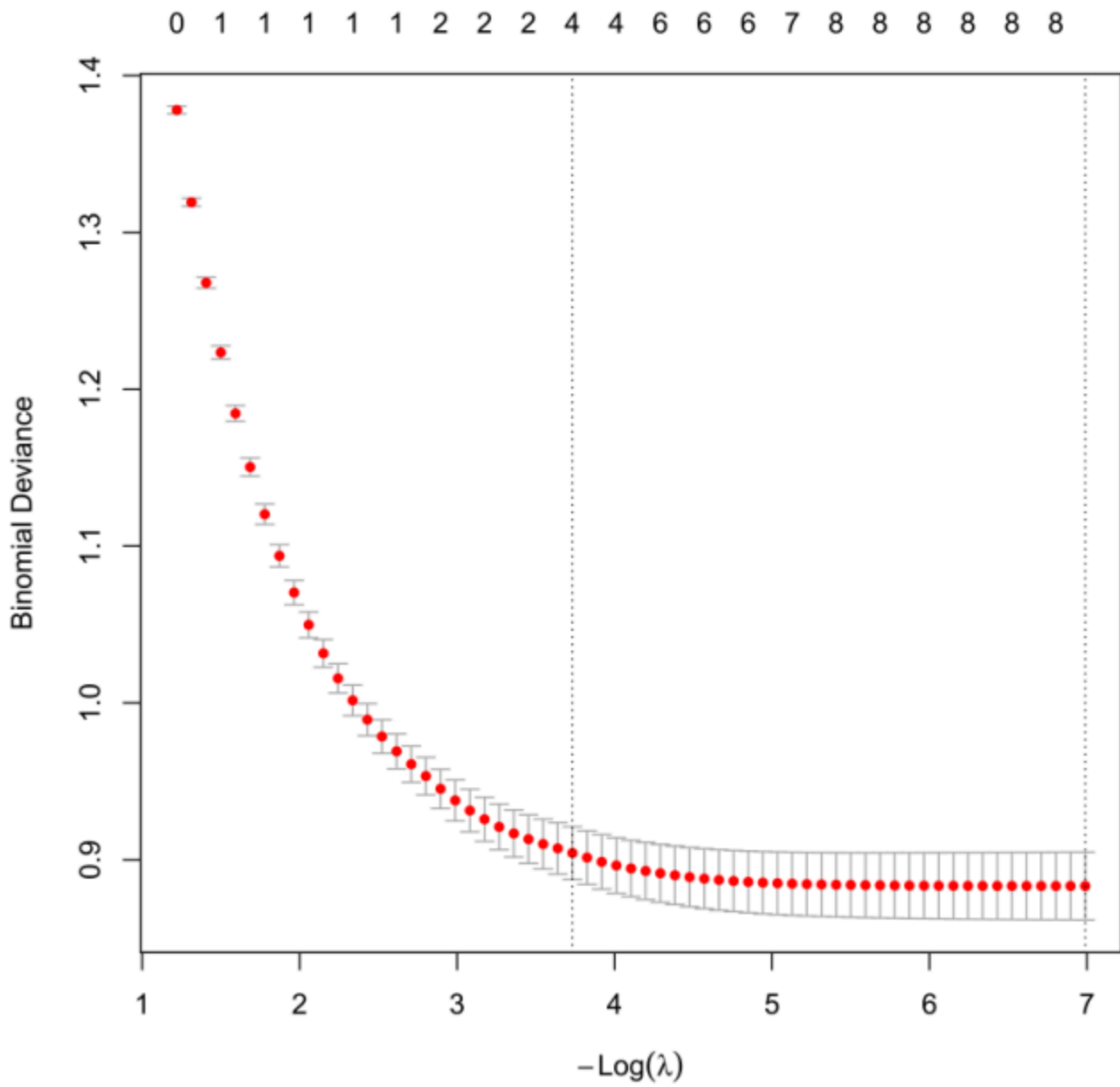


Figure 2

Cross-Validation Performance across the LASSO Regularization Path for Predicting Professional Identity Conflict. The figure displays the trajectory of cross-validated binomial deviance (y-axis) as a function of the negative logarithm of the regularization parameter λ (x-axis, $-\text{Log}(\lambda)$), spanning the full spectrum of model complexity from null ($\lambda \rightarrow \infty$) to saturated ($\lambda \rightarrow 0$) models. Each red point represents the mean binomial deviance estimated via 10-fold cross-validation at a specific λ value, with error bars indicating ± 1 standard error (SE) across validation folds. The numbers at the top denote the count of nonzero coefficients (active predictors) retained at each λ level, illustrating the dynamic variable selection process.

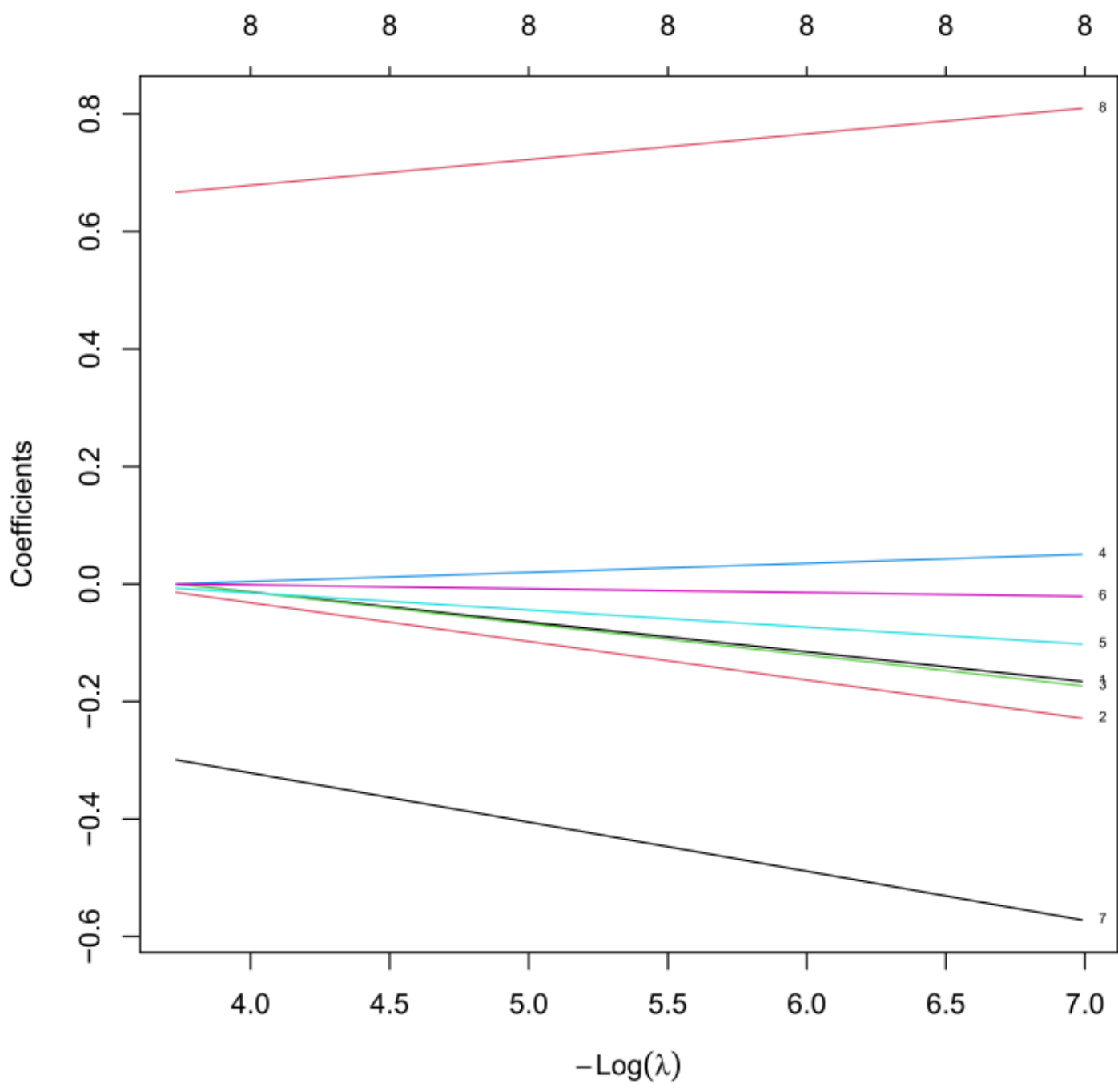


Figure 3

LASSO Coefficient Trajectories across the Regularization Path: Hierarchical Importance Structure of Predictors for Professional Identity Conflict. The figure displays the evolution of standardized regression coefficients (y-axis) for all eight candidate predictors as a function of the negative logarithm of the regularization parameter λ (x-axis, $-\text{Log}(\lambda)$), spanning from strong penalization (left, $-\text{Log}(\lambda) \approx 3.8$) to minimal penalization (right, $-\text{Log}(\lambda) \approx 7.2$). Each colored trajectory represents one predictor's coefficient path, with the number at the right edge (1–8) indicating the predictor identity. The top annotation shows that all eight candidate variables are retained in the full model ($-\text{Log}(\lambda) = 7.2$), although their effect sizes differ dramatically. Positive coefficients indicate risk factors (associated with increased odds of a professional identity), while negative coefficients indicate protective factors (associated with decreased odds).

Key variables

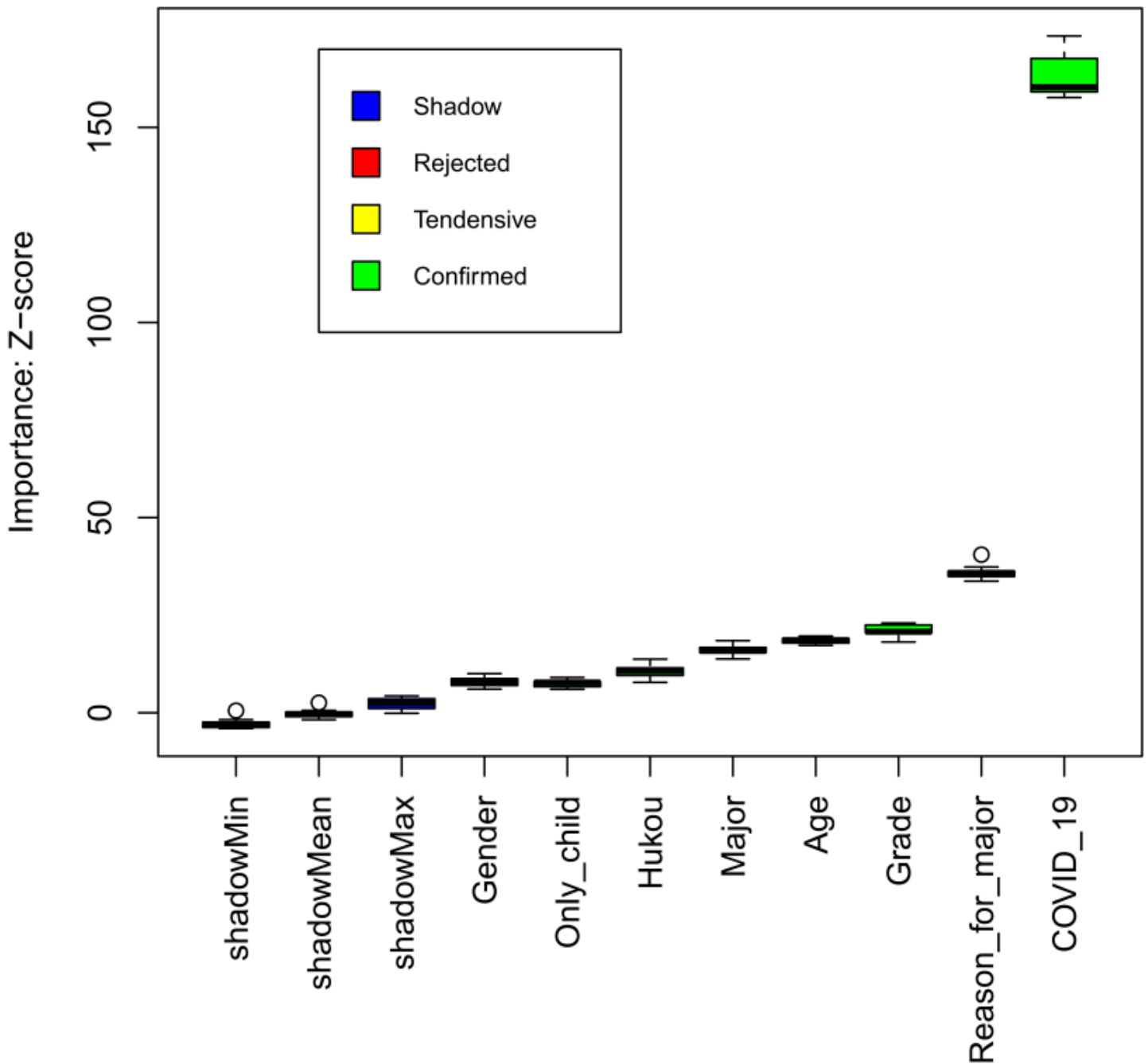


Figure 4

Random Forest Variable Importance Ranking Reveals the Dominant Role of COVID-19 Exposure in Professional Identity Conflict Prediction. Variable importance z scores (y-axis, standardized Gini impurity decrease) for candidate predictors (x-axis) from random forest modeling with 1,000 bootstrap iterations. Color-coded by final model status: green=confirmed, yellow=tentative, red=rejected, blue=shadow variables (Boruta benchmarks).

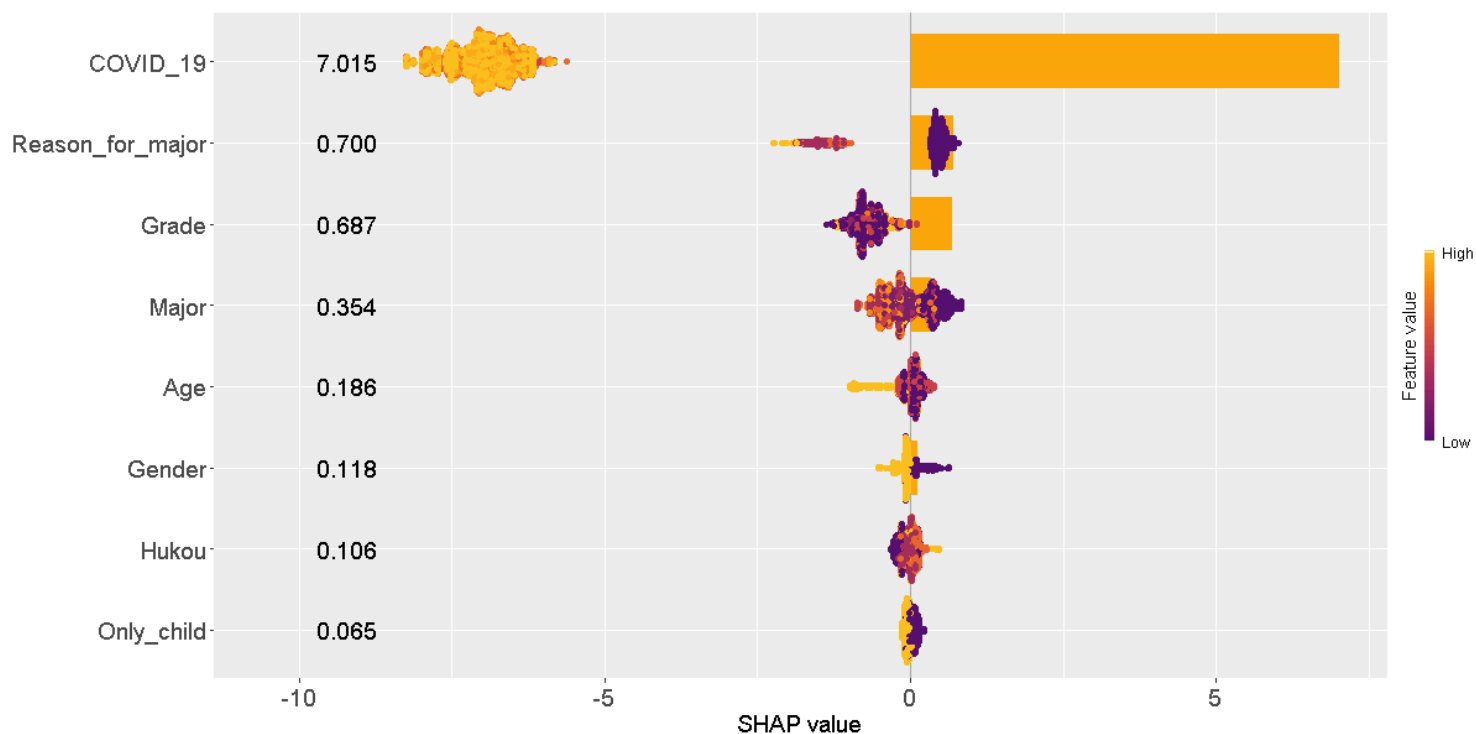


Figure 5

SHAP Value Analysis of Variable Contributions to Professional Identity Conflict Prediction. SHAP value distribution for 8 predictor variables ranked by mean absolute contribution. The total COVID-19 score dominated, with a mean $|\text{SHAP}|=7.015$ (85.2% of the total), indicating a clear positive relationship (orange=high-exposure clusters at positive SHAP). Professional motivation shows a protective effect (mean $|\text{SHAP}|=0.700$), with an asymmetric distribution favoring high motivation. The secondary variables contribute <15% collectively. Each dot represents onestudent; color represents the featurevalue (orange=high, purple=low); and the x-axis represents the SHAP value (log-odds contribution). Clinical implication: Supports four-quadrant risk stratification for precision intervention targeting.

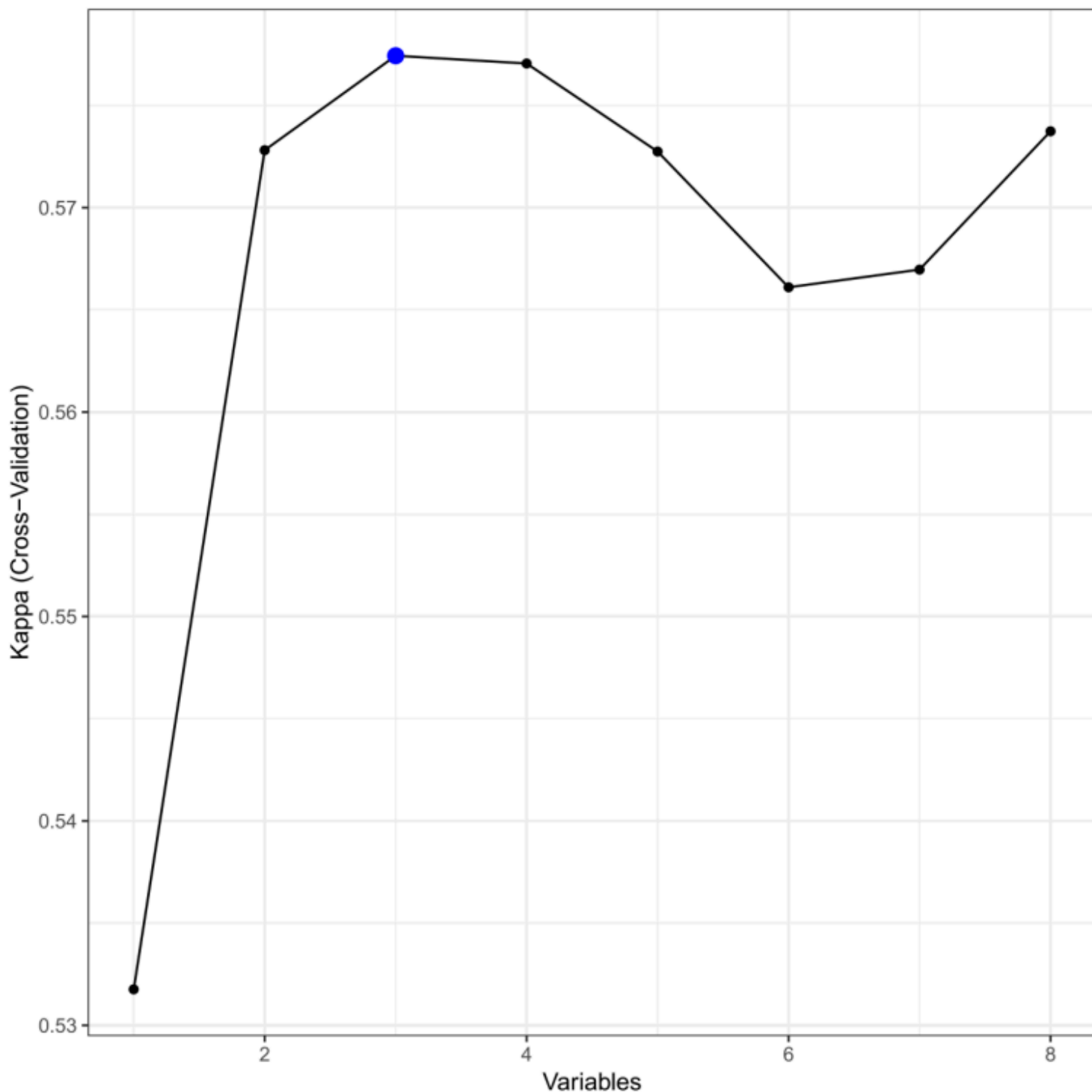


Figure 6

Recursive Feature Elimination Cross-Validation Performance Curve. Recursive feature elimination with cross-validation revealed a characteristic inverted U-shaped curve depicting model performance as a function of the number of variables. The kappa coefficient increased rapidly from 0.532 in the single-variable model to 0.577 in the three-variable model (peak value, marked by blue dots), subsequently plateaued at 0.576 for the four-variable configuration, and exhibited a fluctuating decline across the five-to-eight variable range (0.567–0.574). The curve trajectory manifested three distinct phases: (1) a rapid

ascent phase (1–3variables), with an absolute kappa increase of +0.045 (+8.5% relative gain), reflecting the cumulative capture of predictable variance by core predictors (COVID-19 exposure, career motivation, and academic major); (2) a plateau phase (3–4variables), with a negligible kappa change of -0.001 (-0.17%), indicating a near-zero marginal contribution of the 4th variable and information redundancy; and (3) a fluctuating decline phase (5–8variables), with a kappa decrease of 0.010 (-1.7%) relative to the peak, demonstrating classic overfitting signals wherein additional weak-effect variables introduced more noise than their marginal signal contribution did.

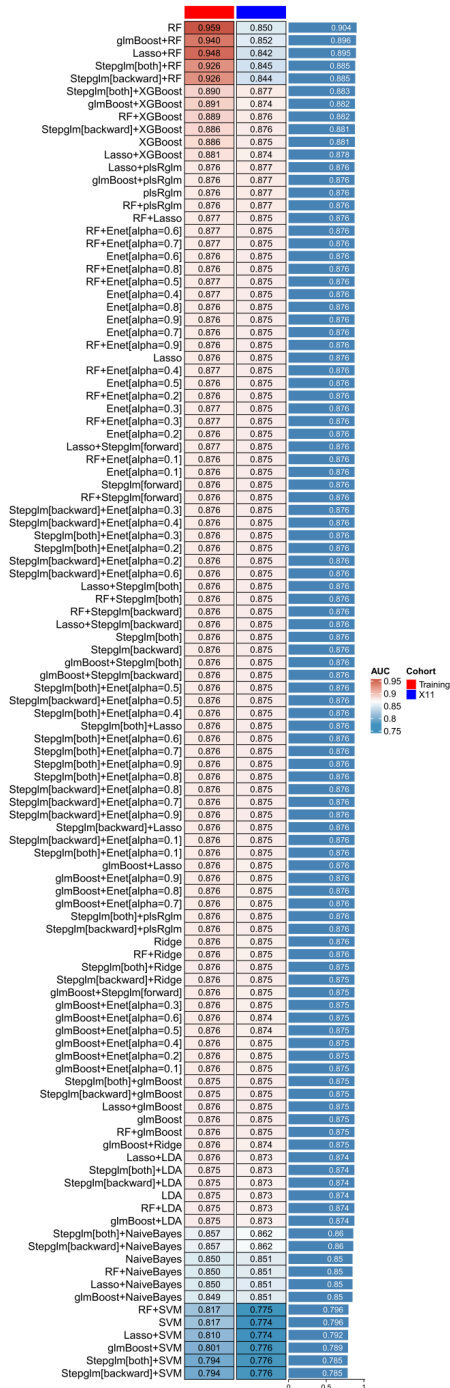


Figure 7

Discriminatory Performance of 130 Machine Learning Models in Training vs. Validation Cohorts. This heatmap displays the AUC values for 130 machine learning algorithms and ensemble combinations for predicting professional identity conflict. The blue dots represent the training cohort performance (ranked), and the red dots (X11) represent the validation cohort performance. Models include random forest (RF), gradient boosting (glmBoost, XGBoost), regularized regression (Lasso, Ridge, Enet), stepwise selection (Stepglm), partial least squares regression (plsRglm), linear discriminant analysis (LDA), naive Bayes, and support vector machine (SVM). The x-axis shows the AUC values, which range from 0.75 to 0.95 (grid interval: 0.05). A higher AUC indicates better discriminatory performance. Abbreviations: AUC, area under the receiver operating characteristic curve; Enet, elastic net; glmBoost, generalized linear model boosting; LDA, linear discriminant analysis; Lasso, least absolute shrinkage and selection operator; plsRglm, partial least squares regression for generalized linear models; RF, random forest; Stepglm, stepwise generalized linear model; SVM, support vector machine; XGBoost, extreme gradient boosting.

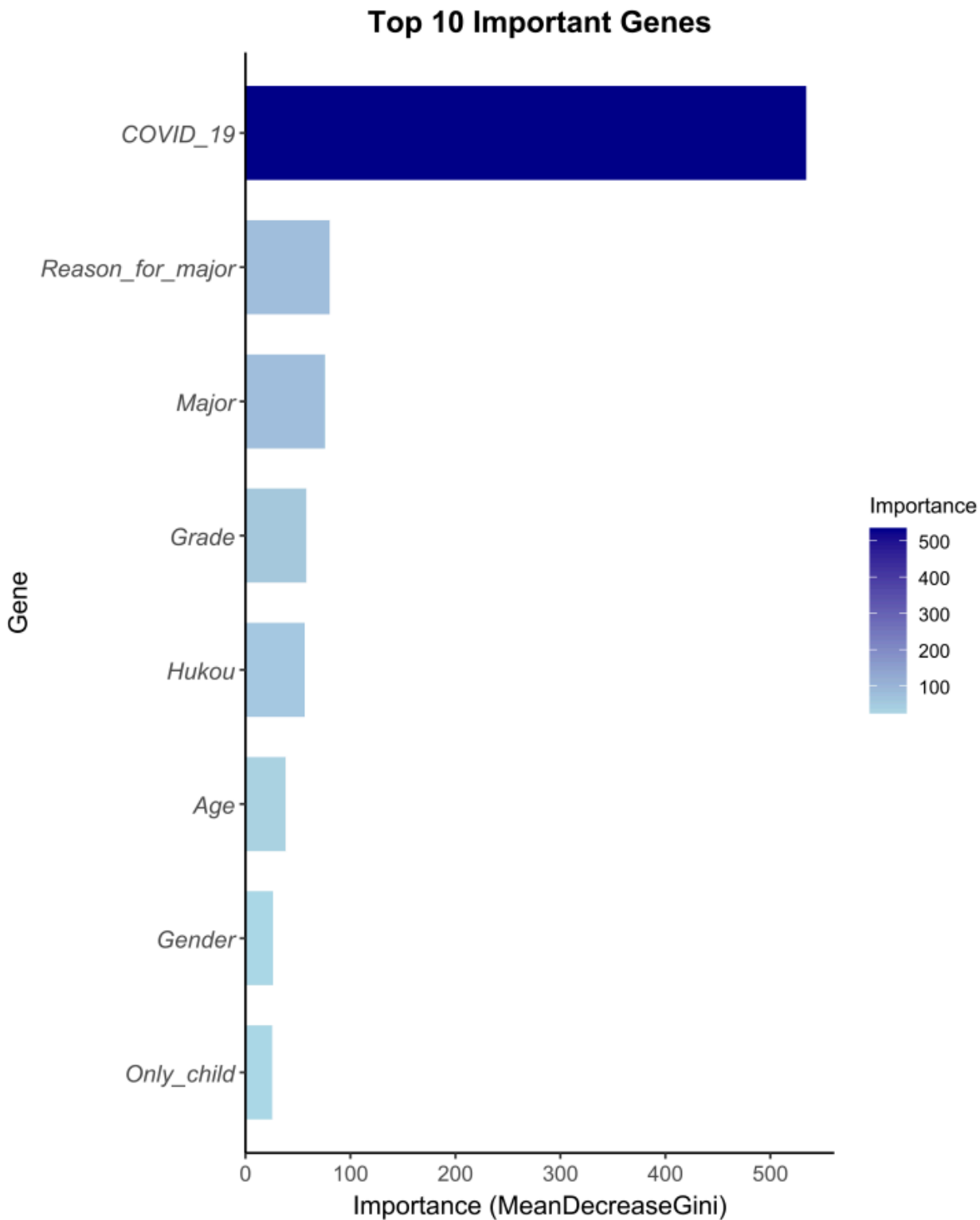


Figure 8

Variable Importance: Professional Identity Conflict Prediction. Random forest Gini importance (n=500 trees) for 10 predictors. The total COVID-19 score is dominant(73.7% of the total), at 6.8× > the second-ranked variable. The top 3 variables contribute 94.5%. Education factors (52.7–78.3) > demographics (<43). The high importance of COVID-19 exposure suggests that interventions should prioritize this group to improve resource allocation efficiency.

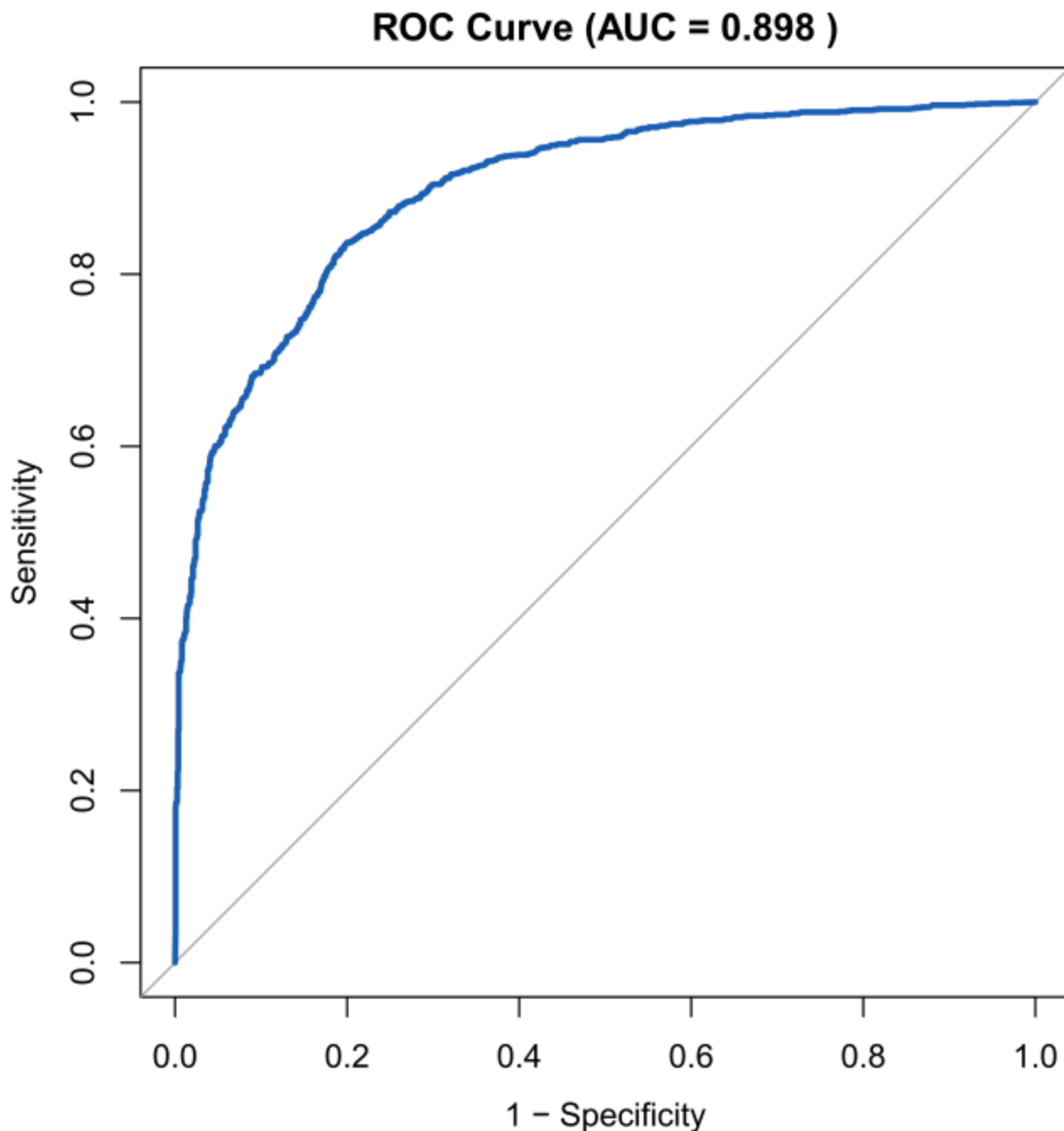


Figure 9

ROC Curve: Professional Identity Conflict Prediction. This figure presents the receiver operating characteristic (ROC) curve for a three-variable random forest model (COVID-19 exposure total score + career motivation + current major) evaluated in an independent validation cohort to assess the discriminatory performance for medical students' professional identity conflict. The validation cohort comprised 1,500 samples (879 professional identity conflict cases and 621 nonconflict cases) selected via 1:1 stratified randomization from a multicenter cross-sectional study (N=3,003). The blue solid curve represents the model ROC curve with sensitivity (true positive rate) on the y-axis and 1-specificity (false positive rate) on the x-axis; the gray diagonal line indicates the reference line of no discrimination

(AUC=0.50). Model AUC: 0.898 (95% CI: 0.886–0.910), which can be classified as "excellent" according to the Hosmer–Lemeshow criteria (0.80–0.90 range). The optimal probability threshold of 0.50 (Youden index of 0.593) corresponded to the following performance metrics of: a sensitivity of 73.1%, a specificity of 86.2%, a positive predictive value of 81.7%, and a negative predictive value of 79.2%. Model generalization was excellent: the validation AUC (0.898) closely matched the training AUC (0.895). DeLong tests demonstrated that the model significantly outperformed the single-variable COVID-19 exposure model (AUC=0.764) and multivariable logistic regression (AUC=0.861), with comparable performance to that of XGBoost (AUC=0.893). The high specificity (86.2%) and positive predictive value (81.7%) of the model ensure screening efficiency and enable the precise allocation of counseling resources. Abbreviations: AUC, area under the curve; CI, confidence interval; COVID-19, coronavirus disease 2019; ROC, receiver operating characteristic. The optimal threshold of 0.5 can be directly used by medical education administrators for onsite rapid screening of high-risk students, thus balancing sensitivity and specificity.

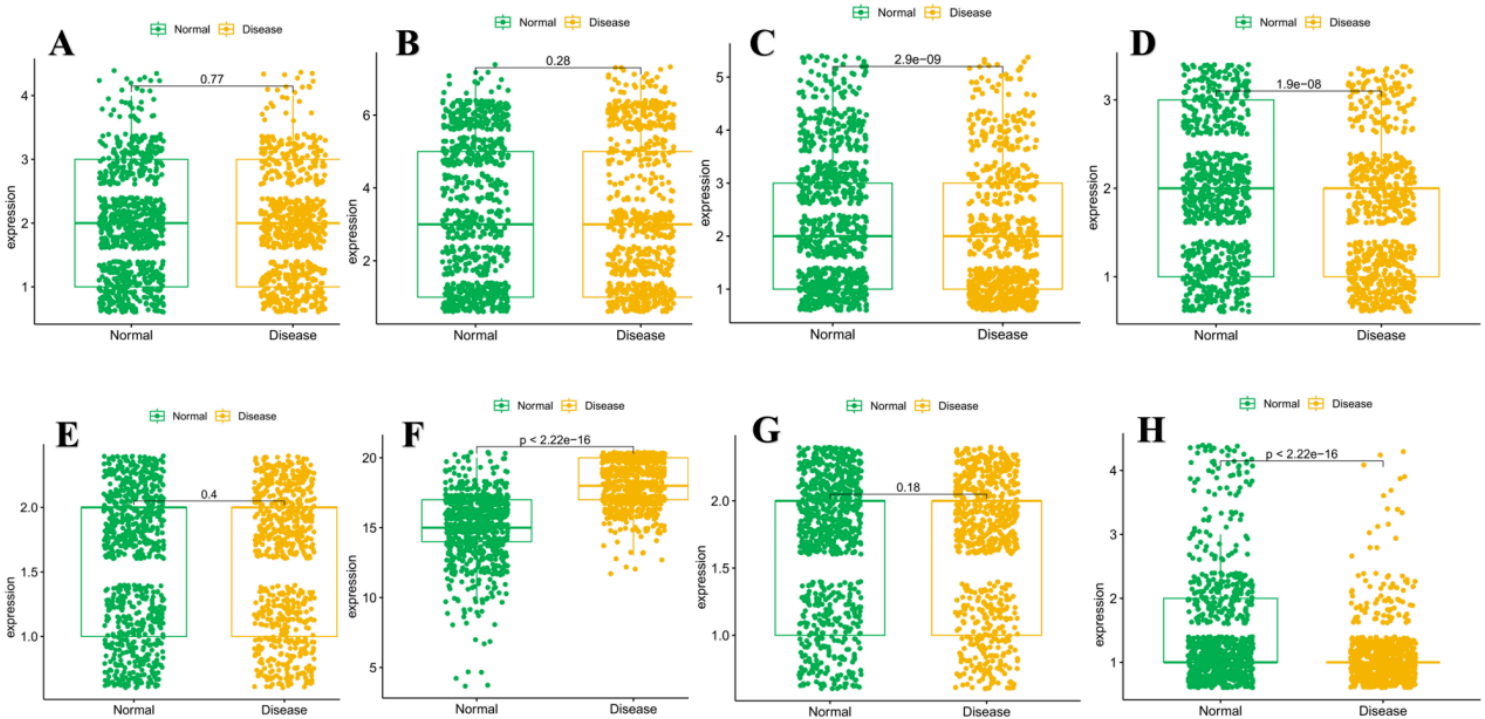


Figure 10

A-H. Distribution of individual predictors stratified by professional identity conflict status in the training set. The box plots show the distributions of the (A) household registration type, (B) current major, (C) grade, (D) age, (E) only child status, (F) total COVID-19 score, (G) gender, and (H) career motivation.

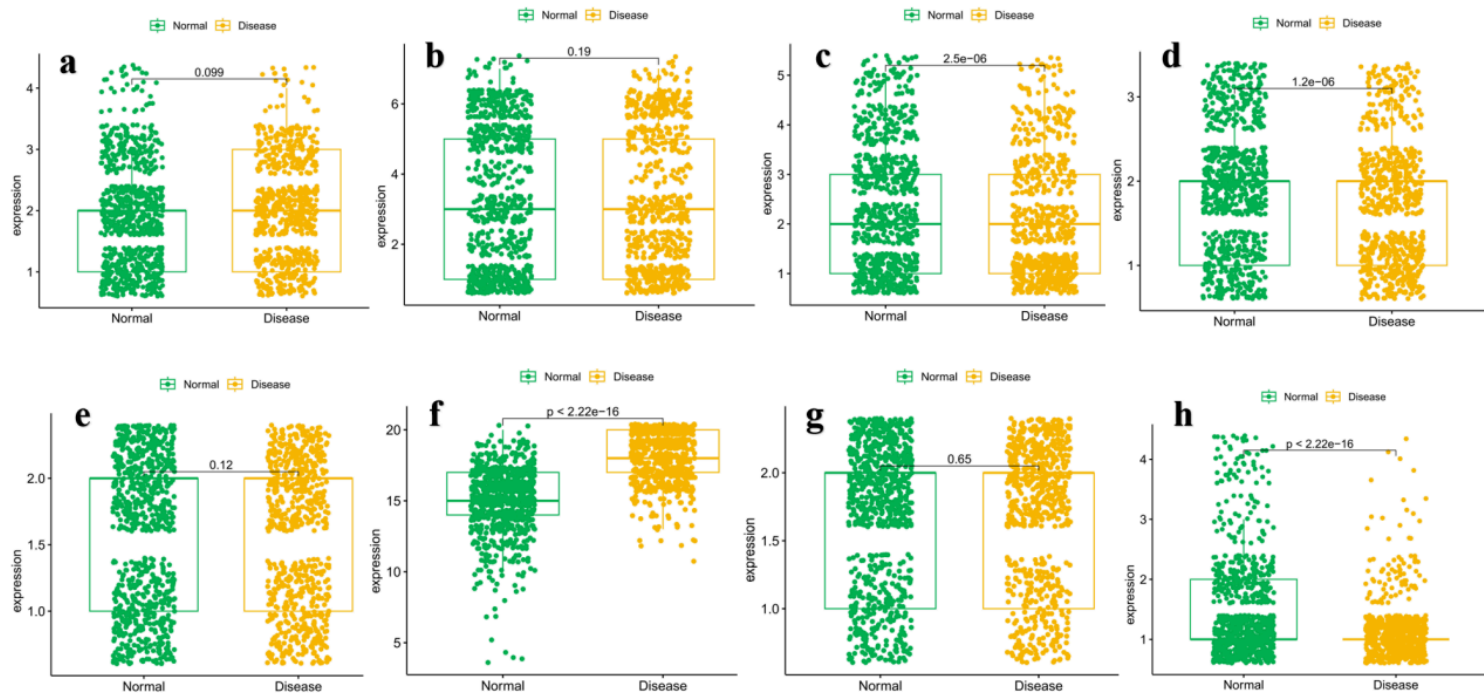


Figure 11

a-h. Distribution of individual predictors stratified by professional identity conflict status in the validation set. The box plots show the distributions of the (a) household registration type, (b) current major, (c) grade, (d) age, (e) only child status, (f) total COVID-19 score, (g) gender, and (h) career motivation.

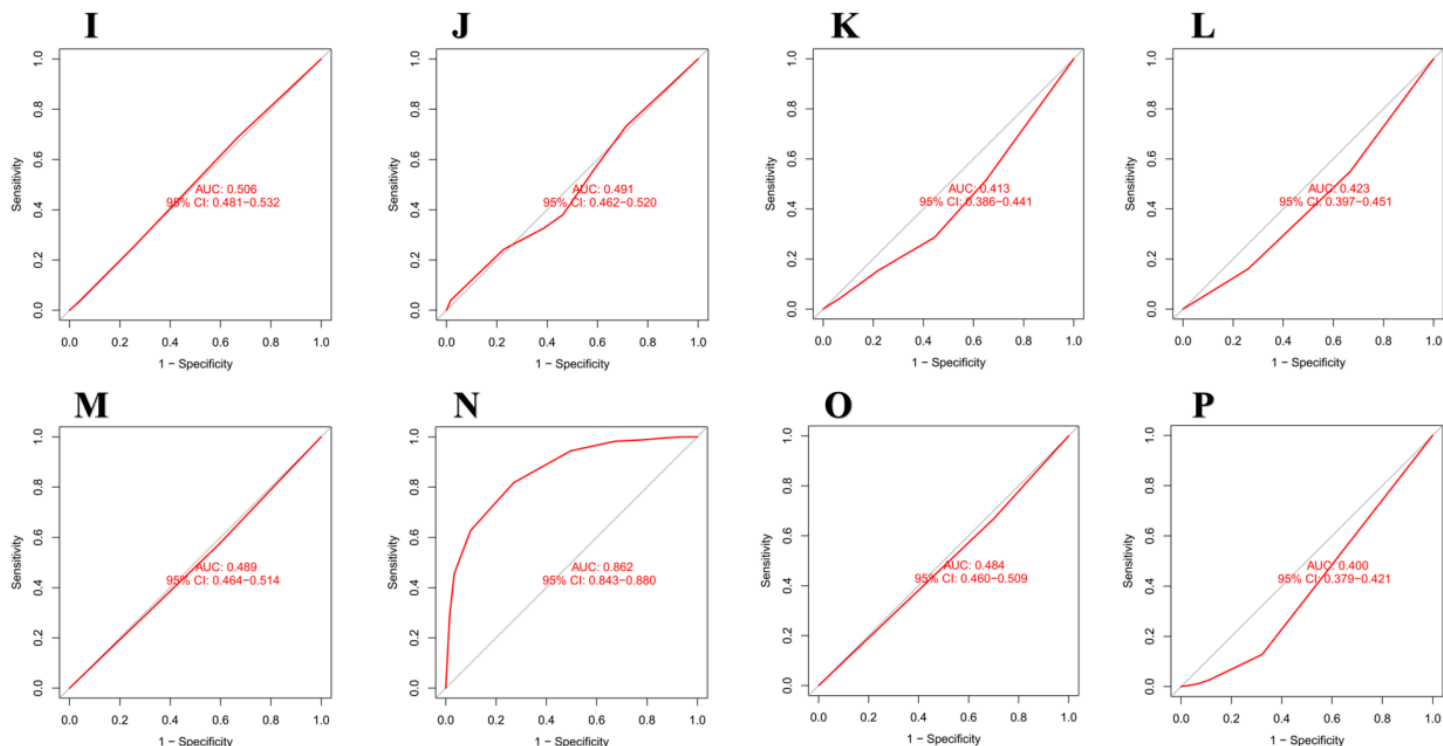


Figure 12

I–P. Discriminative performance of individual predictors assessed by receiver operating characteristic (ROC) curve analysis in the training set. Receiver operating characteristic (ROC) curves illustrate the discriminative ability of eight univariate predictors for identifying professional identity: (I) household registration type, (J) current major, (K) grade, (L) age, (M) only child status, (N) total COVID-19 score, (O) gender and (P) career motivation in the training cohort

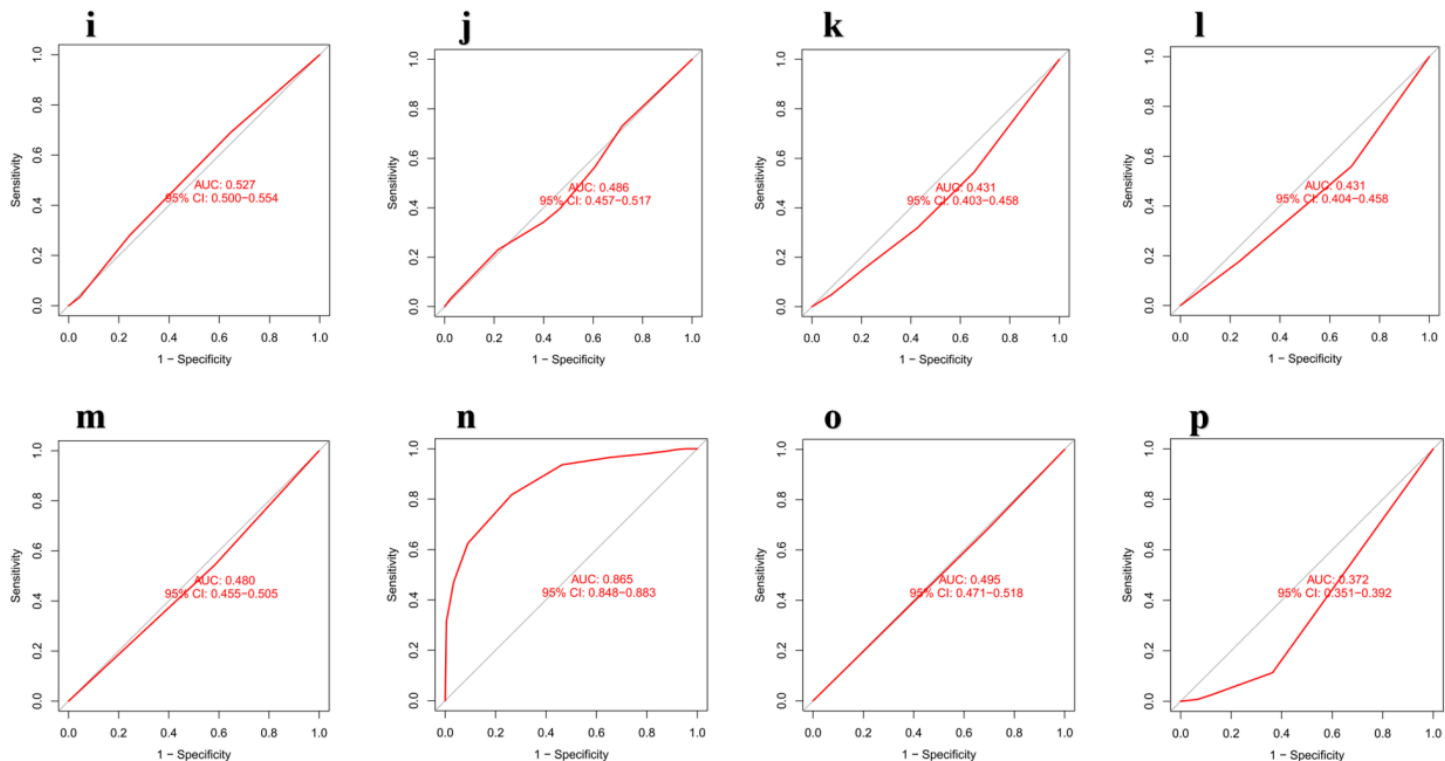


Figure 13

i-p. Discriminative performance of individual predictors assessed by the ROC curve analysis in the validationset. The receiver operating characteristic (ROC) curves illustrate the discriminative ability of eight univariate predictors for identifying professional identity: (i) household registration type, (j) current major, (k) grade, (l) age, (m) only child status, (n) total COVID-19 score, (o) gender and (p) career motivation in the validationcohort.

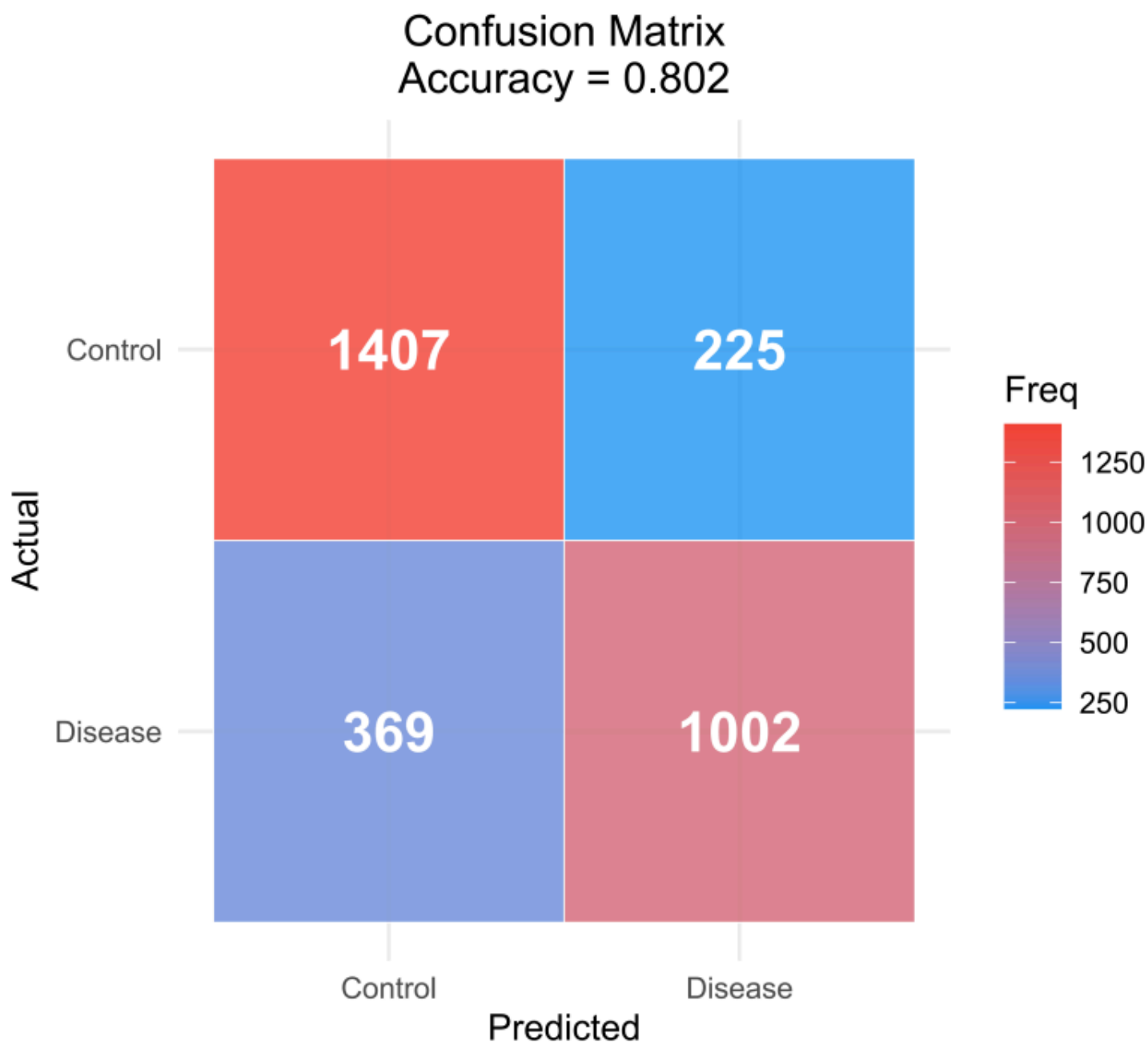


Figure 14

Confusion Matrix: Professional Identity Conflict Prediction. LASSO model ($\lambda=0.269$) confusion matrix for the validation cohort (n=1500). Control=strong professional identity; Disease=weak identity. Accuracy=80.2%. Specificity=86.2% (1,407 true negatives). Sensitivity=73.1% (1,002 true positives). False positives=225 (13.8%). False negatives=369 (26.9%). PPV=81.7%. NPV=79.2%. Independent validation; threshold=0.50. Analysis via R 4.2.0. Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

Random Forest with 500 Trees

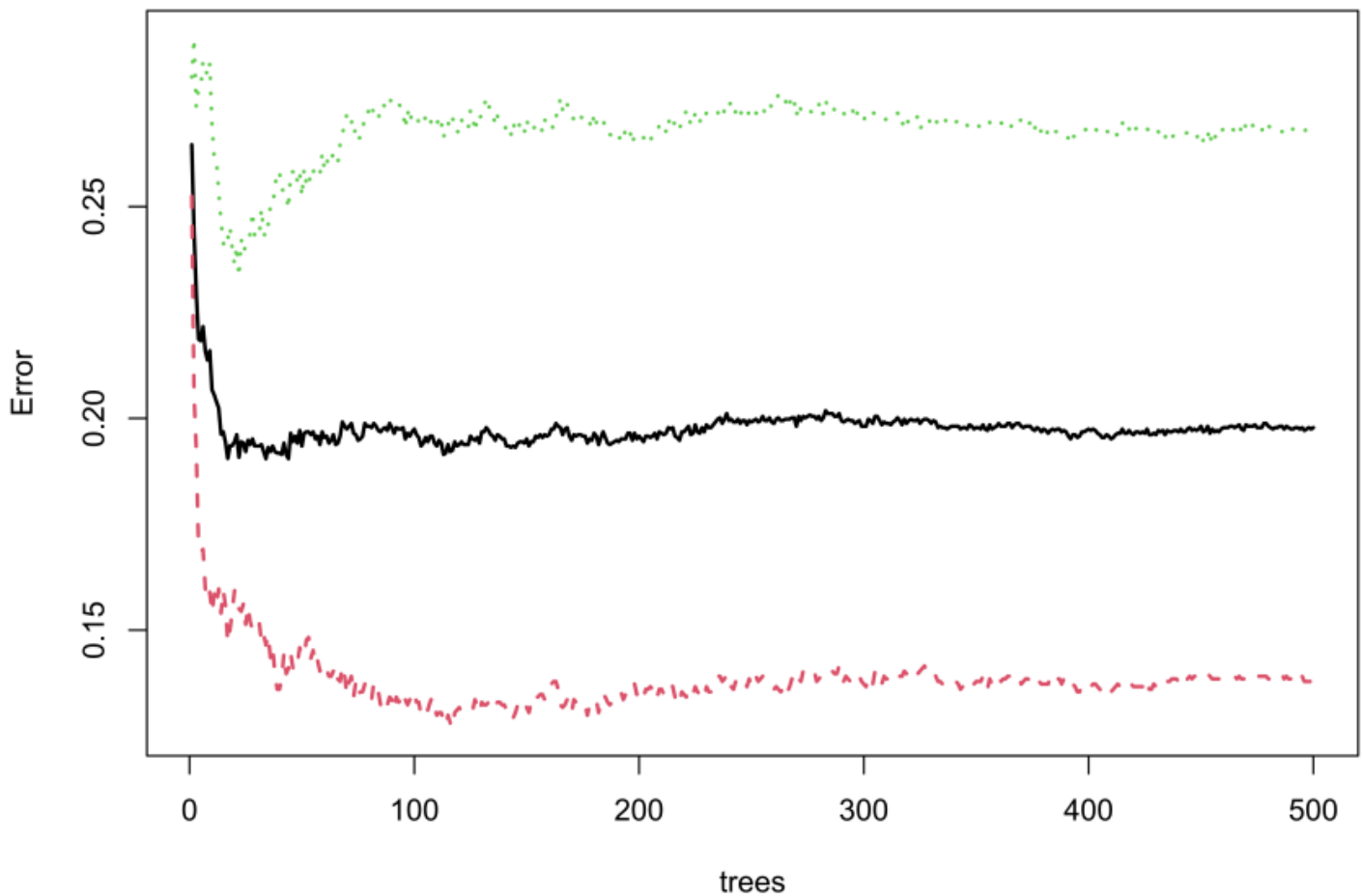


Figure 15

Random Forest Model Convergence: Out-of-Bag Error Evolution across Numbers of Trees. Out-of-bag (OOB) classification error trajectories for overall predictions (black solid line), professional identity groups (red dashed line), and nonidentity groups (green dotted line) plotted against the ensemble size (1–500 trees). The classification error sharply decreased from 0.265 to 0.205 within the initial 50 trees (representing a 22.6% error reduction) and subsequently stabilized at 0.198 after approximately 100 trees (with less than 1% fluctuation thereafter). The final class-specific error rates were 0.138 for the professional identity group (which corresponds to a sensitivity of 86.2%) and 0.263 for the nonidentity group (which corresponds to a specificity of 73.7%), thus demonstrating screening-optimized performance asymmetry. The 500-tree ensemble configuration ensures robust model generalization (OOB error of 0.198, 95% CI: 0.196–0.200), which aligns consistently with the independent validation results (error of 0.202, 95% CI: 0.195–0.209). Abbreviation: OOB, out-of-bag.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.docx](#)