

Identifying Mild Cognitive Impairment Using Decision Tree–Based Machine Learning with Physical, Functional, and Psychosocial Measures in Community-Dwelling Older Adults: Evidence from the Northern Japanese ORANGE Registry

Ayuto Kodama

Akita University

Takako Ohnuma

Akita University

Kana Sasaki

Akita University

Kaoru Sugawara

Akita University

Nobuhiro Fujiyama

Akita University

Youko Umetsu

City Hall of Yokote

Tsuyoshi Ono

Omori Hospital

Hidetaka Ota

`hidetaka-ota@med.akita-u.ac.jp`

Akita University

Research Article

Keywords: mild cognitive impairment, machine learning, random forest, community health check-up, depression, social frailty, older adults

Posted Date: March 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8729100/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

Mild cognitive impairment (MCI) is common in later life and represents a key target for early identification and prevention. Scalable, non-imaging approaches using routinely collected community health data may support risk stratification and guide follow-up assessments.

Methods

We analyzed community health check-up data from Akita Prefecture, Japan. The outcome was binary MCI classification (0 = non-MCI; 1 = MCI). Candidate predictors included demographics, medical history, physical function, and psychosocial measures. Data were split into training (70%) and test (30%) sets using stratification. We trained a decision tree, random forest, and gradient-boosted trees with five-fold cross-validation and hyperparameter tuning. Model discrimination and classification metrics were evaluated on the independent test set. Permutation importance was computed for the best-performing model, and a shallow decision tree was derived using the top-ranked predictors for interpretability.

Results

The analytic sample included 2,650 participants (non-MCI: $n = 1,893$; MCI: $n = 757$). On the test set, the random forest model achieved the highest ROC AUC (0.719). At a 0.5 threshold, accuracy was 0.753, with sensitivity 0.254 and specificity 0.952. Using the Youden threshold (~ 0.256) increased sensitivity to 0.794 while reducing specificity to 0.537. Permutation importance ranked GDS-15 score, osteoporosis, social frailty, and living alone among the top predictors.

Conclusions

A random forest model demonstrated moderate discrimination for classifying MCI using routinely collected community health variables. The choice of operating threshold had a substantial impact on the sensitivity–specificity trade-off, underscoring the importance of clearly defining intended use and decision thresholds. External validation and prospective evaluation are required before clinical deployment.

Introduction

Mild cognitive impairment (MCI) is an intermediate clinical state between normal cognitive aging and dementia, associated with an elevated risk of progression and functional decline [1]. Dementia prevention has become a major public health priority, as emphasized in recent Lancet Commission reports that advocate for multidomain risk-reduction strategies [2–3]. In parallel, diagnostic frameworks

increasingly emphasize biological and clinical staging across the cognitive decline continuum [4]. Despite these advances, population-level screening remains challenging. The U.S. Preventive Services Task Force has noted uncertainty regarding the net benefit of universal cognitive screening in asymptomatic older adults, reinforcing the need for pragmatic, scalable approaches to identify individuals who may benefit from further assessment [5]. Community-based programs often collect demographic information, comorbidity profiles, physical function data, and psychosocial measures. Leveraging these routinely collected data for MCI risk stratification could facilitate targeted follow-up. Mood and social vulnerability may be particularly relevant: depression is common among individuals with MCI and may co-occur with or worsen cognitive symptoms [6]. Broader indicators of biological vulnerability—such as low bone density—have also been associated with incident dementia in population studies [7]. Lifestyle and vascular risk modification can influence cognitive outcomes. For example, the SPRINT MIND trial found that intensive blood pressure control reduced the risk of MCI, supporting the plausibility of prevention pathways even in later life [8]. Collectively, these findings motivate practical, community-oriented prediction approaches that integrate multiple domains and clarify how decision thresholds affect the trade-off between missed cases and false positives. Machine learning (ML) methods may capture nonlinear relationships and interactions within such multidomain data [9]. However, robust reporting and evaluation are essential to avoid overly optimistic claims and to clarify the intended use of these tools. In recent years, reporting guidelines for prediction modeling and AI-based decision support have been significantly expanded [10–14]. In this study, we aimed to develop and internally evaluate decision tree–based models (single tree and ensemble models) to classify MCI among community-dwelling older adults using routinely collected demographic, medical, physical, and psychosocial data. We also sought to identify influential predictors and to construct an interpretable, shallow decision tree for use in clinical communication.

Methods

Study Design and Participants

This cross-sectional modeling study was conducted in Akita Prefecture, Japan, between 2018 and 2025. Participants were community-dwelling adults aged 65 years or older, capable of walking independently and living at home without personal assistance. Exclusion criteria included major depression, severe hearing or visual impairment, stroke, intellectual disability, certification for long-term care need under the Japanese public system, and inability to complete cognitive tests. The study was approved by the Ethics Committee of the Faculty of Medicine, Akita University (Approval No. 273) and conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants.

Outcome

The outcome was binary MCI classification (0 = non-MCI, 1 = MCI). Following Petersen's criteria [15], MCI was defined as cognitive impairment without loss of independence in daily activities, operationalized

using the NCGG-FAT cognitive test. Impairment was defined when scores on cognitive subtests were more than 1.5 standard deviations below age- and education-adjusted population means derived from Japanese community-dwelling cohorts [16].

Candidate predictors

Candidate predictors were restricted to non-cognitive variables collected at the same health check-up visit to minimize outcome leakage; no NCGG-FAT scores or other cognitive test variables were used as predictors. Demographic variables included age, sex, and education level. Medical history was recorded for hypertension, stroke, cardiac disease, diabetes, hyperlipidemia, osteoporosis, respiratory disease, osteoarthritis (OA), fracture history after age 60, malignant neoplasm, Parkinson's disease (PD), Alzheimer's disease (AD), depression, and rheumatoid arthritis (RA). Depressive symptoms were assessed using the 15-item Geriatric Depression Scale (GDS-15; score range 0–15, with higher scores indicating more symptoms). Physical frailty was defined using the Fried frailty phenotype [17] and its Japanese NCGG-SGS operationalization [18]. Social frailty was assessed using the Makizako Social Frailty Index [19] and categorized as robust (0), prefrail (1), or frail (2–5).

Data Splitting and Preprocessing

Data were stratified by outcome and randomly split into training (70%) and test (30%) sets. Preprocessing included median imputation for missing numeric values, removal of zero-variance predictors, and z-score normalization of numeric variables. All preprocessing steps were fit using the training data only and then applied to the held-out test set. Although tree-based models are largely scale-invariant, we applied z-score normalization to maintain a consistent preprocessing pipeline across models and to facilitate future extensions to non-tree-based approaches.

Model Development

Three models were developed: (i) a decision tree (CART), (ii) a random forest, and (iii) gradient-boosted trees (XGBoost). Hyperparameters were tuned using stratified five-fold cross-validation within the training set, with model selection based primarily on ROC AUC.

To support reproducibility, the final tuned hyperparameters, preprocessing details, and software information are provided in Additional file 1: Table S1. The analysis code is available from the corresponding author on reasonable request.

Performance Evaluation

Final models were retrained on the full training set and evaluated on the held-out test set. Performance metrics included ROC AUC, accuracy, sensitivity, specificity, precision, recall, and F1 score, with MCI (1)

defined as the positive class. Approximate 95% confidence intervals (CIs) for ROC AUC were computed using the method of Hanley and McNeil [20]. For classification metrics at prespecified thresholds, 95% CIs were computed using Wilson score intervals.

Interpretation and Reporting

Permutation importance was calculated for the best-performing model. To enhance interpretability, a shallow decision tree was constructed using the top-ranked predictors. Reporting followed TRIPOD + AI [13]. Risk of bias was assessed using PROBAST + AI [12], and considerations for early-stage clinical evaluation were informed by DECIDE-AI [10].

Results

Participants

The dataset included 2,650 community-dwelling older adults (non-MCI n = 1,893; MCI n = 757).

Data split and outcome prevalence

The data were stratified by outcome and split into training (70%) and test (30%) sets. This yielded 1,854 participants in the training set (non-MCI n = 1,325; MCI n = 529) and 796 in the test set (non-MCI n = 568; MCI n = 228). The overall MCI prevalence was 28.6% (757/2,650).

Model discrimination and classification performance

On the independent test set (Table 1), the random forest achieved the highest discrimination (ROC AUC 0.719; 95% CI 0.678–0.761). At the default 0.5 probability threshold, it showed high specificity (0.952) but low sensitivity (0.254), reflecting conservative classification under class imbalance. XGBoost and the single decision tree yielded lower ROC AUC (0.649; 95% CI 0.605–0.692 and 0.641; 95% CI 0.597–0.685, respectively) with different sensitivity–specificity trade-offs. Precision and F1 scores followed a similar pattern, with random forest providing the highest precision (0.682) but reduced recall.

Receiver operating characteristic and precision–recall curves (random forest)

The random forest ROC curve (Fig. 1) demonstrated moderate separation between MCI and non-MCI across thresholds. Because MCI prevalence was < 30%, the precision–recall curve (Fig. 2) was additionally examined to contextualize performance in the setting of class imbalance. At higher-recall operating points, precision decreased, illustrating the expected trade-off for screening-oriented thresholds.

Calibration of predicted probabilities (random forest)

Calibration was assessed visually using the calibration plot (Fig. 3). Predicted probabilities showed reasonable agreement with observed outcome proportions in the mid-range, while variability increased toward the extremes, where fewer observations contributed to each bin. Quantitative calibration metrics (e.g., Brier score and calibration intercept/slope) are important for probability-based deployment and should be prioritized in external validation and prospective evaluation.

Table 1
Model performance on the independent test set.

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	ROC_AUC
RandomForest	0.753	0.254	0.952	0.682	0.254	0.371	0.719
XGBoost	0.700	0.399	0.820	0.472	0.399	0.432	0.649
DecisionTree	0.730	0.228	0.931	0.571	0.228	0.326	0.641

Permutation importance (mean decrease in ROC AUC after permuting each predictor) was highest for GDS-15 score (0.0061), followed by osteoporosis (0.0015) and social frailty (0.0013) (Table 2). Living alone (0.0010), malignant neoplasm (0.0009), and fracture history since age 60 (0.0008) also contributed. Several predictors exhibited near-zero or negative importance, which may occur when predictors are weakly informative, correlated with stronger predictors, or when importance estimates vary due to sampling noise in the test set.

Table 2
Permutation importance (top 10 predictors) for the random forest model.

Variable_EN	Importance
GDS-15 score	0.006
Osteoporosis	0.001
Social frailty	0.001
Living alone (yes/no)	0.001
Malignant neoplasm	0.001
History of fracture (age ≥ 60)	0.000
Alzheimer's disease	0.000
Parkinson's disease	-0.000
Current smoker (yes/no)	-0.000
≥ 5% weight loss in 2 years	-0.000

Operating point analysis and confusion matrix (random forest)

Thresholding substantially changed clinical behavior. At the default threshold of 0.5, the random forest prioritized specificity 0.952 (95% CI 0.932–0.967) with sensitivity 0.254 (95% CI 0.202–0.315). In the test set, this corresponded to TN = 541 and FP = 27 among non-MCI cases (n = 568), and TP = 58 and FN = 170 among MCI cases (n = 228). At this threshold, PPV was 0.682 (95% CI 0.577–0.772) and NPV was 0.761 (95% CI 0.728–0.791). Using the Youden-derived threshold (~ 0.256), sensitivity increased to 0.794 (95% CI 0.737–0.841) with specificity 0.537 (95% CI 0.496–0.578). This corresponded to TN = 305 and FP = 263 among non-MCI cases, and TP = 181 and FN = 47 among MCI cases (Fig. 4). At the Youden threshold, PPV was 0.408 (95% CI 0.363–0.454) and NPV was 0.866 (95% CI 0.827–0.898), illustrating that a higher-sensitivity screening threshold increased false positives but reduced missed MCI cases.

Discussion

Using routinely collected community health check-up data, we compared a single decision tree with two ensemble-based models to classify mild cognitive impairment (MCI) among community-dwelling older adults in Akita, Japan. Among the three approaches, the random forest model demonstrated the highest discriminatory performance (ROC AUC = 0.719) on the test set. Permutation-based variable importance identified depressive symptoms (GDS-15), social frailty, living arrangements, and osteoporosis as key contributors, suggesting that psychosocial and frailty-related factors are important alongside traditional comorbidity markers. These results support the feasibility of non-imaging, community-friendly models for MCI risk stratification, although performance remains modest and must be interpreted according to intended use—e.g., screening versus diagnostic support [5, 13]. Model interpretation indicated that psychosocial vulnerability and multimorbidity both played roles in differentiating cases. GDS-15 and social frailty indicators, including living alone, consistently ranked among the most important features. This supports a multidomain framework in which affective symptoms, social support, and physiological vulnerability co-occur with cognitive decline in older adults. Importantly, these variables are potentially modifiable through community-based preventive strategies—such as mental health services, social engagement programs, and bone health interventions—particularly when the model is used for prioritizing follow-up rather than standalone diagnosis.

Our threshold analysis further highlights the importance of operating point selection. At the default 0.5 threshold, the model favored specificity at the cost of sensitivity. Using the Youden index-derived threshold improved sensitivity to 0.794 but reduced specificity to 0.537. Threshold selection should reflect the clinical context, resource availability, and acceptable trade-offs. Decision curve analysis can help quantify net benefit across different thresholds by considering the relative consequences of false negatives and false positives. Calibration is equally important: even well-discriminating models can yield poorly calibrated probabilities, which may mislead downstream decisions. Recalibration may therefore be necessary before deployment. Class imbalance requires reporting metrics beyond accuracy. In our

data, where MCI cases were less common, the default threshold produced high specificity but low sensitivity. Sensitivity, specificity, precision, and confusion matrices at clinically meaningful thresholds provide a more informative picture, especially in scenarios where the cost of missed cases may be higher than false alarms. To support transparency and clinical communication, we supplemented ensemble performance results with interpretability outputs. Permutation importance and a shallow surrogate tree highlighted key variables driving predictions. These tools are valuable for hypothesis generation and explanation but should not be misinterpreted as causal evidence. For example, permutation importance can be unstable with correlated predictors, and surrogate trees offer only an approximation of ensemble behavior.

Our findings align with broader trends in non-imaging ML applications for MCI prediction: modest discrimination, variable performance by threshold, and the importance of domain integration. These models require external validation to assess generalizability and ongoing monitoring to account for changing population characteristics over time. Staged evaluation, from internal validation to prospective testing, is essential before implementation.

We followed TRIPOD + AI reporting guidelines [13] and applied PROBAST + AI [12] to assess potential sources of bias and applicability. Future work should include validation in other municipalities, calibration assessment, and evaluation of real-world implementation impact (e.g., uptake of follow-up assessments, clinical efficiency, and unintended consequences), following early-stage evaluation frameworks such as DECIDE-AI [10].

Several strengths and limitations should be noted. Strengths include the use of routinely collected health check-up data, inclusion of both simple and ensemble models, held-out test set validation, and interpretability analysis. Limitations include the cross-sectional design, possible outcome misclassification depending on the operational definition of MCI, and limited generalizability due to the single-region sample. Threshold and interpretability analyses may be optimistic and require further validation. Correlated predictors could also influence variable importance rankings. External validation and prospective impact assessment remain necessary before clinical deployment. Going forward, to eliminate the single-region limitation of this analysis and evaluate transferability, external validation will be conducted using independent cohorts collected outside Akita Prefecture. Specifically, during FY2026–FY2027, we will assemble a validation dataset from community-based health-check programs and/or ORANGE Registry sites in other Japanese prefectures (e.g., additional municipalities in the Tohoku region and collaborating sites nationwide) where the same non-cognitive predictors can be obtained using comparable protocols.

Conclusion

In this cross-sectional study of community-dwelling older adults in Akita, Japan, a random forest model using routinely collected health check-up data achieved modest discrimination for classifying mild cognitive impairment (MCI). The most influential predictors included depressive symptoms, social frailty

indicators, living arrangement, and osteoporosis. The choice of operating threshold substantially influenced the sensitivity–specificity balance, highlighting the need to predefine intended use and to align thresholds with clinical capacity and consequences.

Before implementation, the model requires external validation in independent populations, calibration assessment and adjustment if necessary, and prospective evaluation. Future studies should investigate whether model-informed screening or triage improves timely follow-up and outcomes without imposing undue burden or harm.

Abbreviations

AUC, area under the curve; CART, classification and regression tree; CONSORT-AI, Consolidated Standards of Reporting Trials—AI extension; DECIDE-AI, Developmental and Exploratory Clinical Investigations of Decision support systems driven by Artificial Intelligence; GDS-15, 15-item Geriatric Depression Scale; MCI, mild cognitive impairment; ML, machine learning; PROBAST+AI, Prediction model Risk Of Bias Assessment Tool—AI extension; ROC, receiver operating characteristic; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials—AI extension; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis—AI extension.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of the Faculty of Medicine, Akita University (Approval No. 273) and was performed in accordance with the Declaration of Helsinki. Informed consent was obtained from the participants.

Acknowledgements

The authors thank all the participants in this survey. Additionally, we would also like to thank all staff at Yokote Hall and Omori Hospital who provided assistance in performing the assessments.

Authors' contributions

Conceived the study and participated in the study design: A.K. Recruited and collected data: A.K. , T.O. , Y.U., K.S., N.F., T. Oand H.O. Analyzed data: A.K. All authors participated in interpretation of the results. A.K. drafted the manuscript, and all authors contributed to critical review and revision of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study contain potentially identifiable information and are not publicly available. De-identified data are available from the corresponding author on reasonable request, subject to ethics approval and data-sharing agreements.

Clinical trial number

Not applicable.

References

1. Petersen RC, Lopez O, Armstrong MJ, et al. Mild cognitive impairment. *Nat Rev Dis Primers*. 2021;7(1):33. 10.1038/s41572-021-00269-y.
2. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2020;396(10248):413–46. 10.1016/S0140-6736(20)30367-6.
3. Livingston G, Sommerlad A, Orgeta V, et al. Dementia prevention, intervention, and care. *Lancet*. 2017;390(10113):2673–734. 10.1016/S0140-6736(17)31363-6.
4. Jack CR Jr, Bennett DA, Blennow K, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535–62. 10.1016/j.jalz.2018.02.018.
5. US Preventive Services Task Force. Screening for cognitive impairment in older adults: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2020;323(8):757–63. 10.1001/jama.2020.0435.
6. Ismail Z, Elbayoumi H, Fischer CE, et al. Prevalence of Depression in Patients With Mild Cognitive Impairment: A Systematic Review and Meta-analysis. *JAMA Psychiatry*. 2017;74(1):58–67. 10.1001/jamapsychiatry.2016.3162.
7. Xiao T, Zhang Y, Chen X, et al. Association of Bone Mineral Density and Dementia: The Rotterdam Study. *Neurology*. 2023;100(20):e2125–33. 10.1212/WNL.0000000000207220.
8. Williamson JD, Pajewski NM, Auchus AP, et al. Effect of Intensive vs Standard Blood Pressure Control on Probable Dementia: A Randomized Clinical Trial (SPRINT MIND). *JAMA*. 2019;321(6):553–61. 10.1001/jama.2018.21442.
9. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med*. 2019;380(14):1347–58. 10.1056/NEJMr1814259.
10. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ*. 2022;377:e070904. 10.1136/bmj-2021-070904.
11. SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–63. 10.1038/s41591-020-1037-7.

12. Moons KGM, Damen JAA, Kaul T, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. 10.1136/bmj-2024-082505.
13. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. 10.1136/bmj-2023-078378.
14. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–74. 10.1038/s41591-020-1034-x.
15. Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Intern Med*. 2004;256:183–94. 10.1111/j.1365-2796.2004.01388.x.
16. Inoue M, Jinbo D, Nakamura Y, Taniguchi M, Urakami K. Development and evaluation of a computerized test battery for Alzheimer’s disease screening in community-based settings. *Am J Alzheimers Dis Other Dement*. 2009;24:129–35. 10.1177/1533317508330222.
17. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. *J Gerontol Biol Sci Med Sci*. 2001;56(3):M146–56.
18. Satake S, Shimada H, Yamada M, Kim H, Yoshida H, Gondo Y, et al. Prevalence of frailty among community-dwellers and outpatients in Japan as defined by the Japanese version of the Cardiovascular Health Study criteria. *Geriatr Gerontol Int*. 2017;17:2629–34.
19. Makizako H, Shimada H, Tsutsumimoto K, Lee S, Doi T, Nakakubo S, et al. Social Frailty in Community-Dwelling Older Adults as a Risk Factor for Disability. *J Am Med Dir Assoc*. 2015;16:e10037–11. <https://doi.org/10.1016/j.jamda.2015.08.023>.
20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36. 10.1148/radiology.143.1.7063747.

Figures

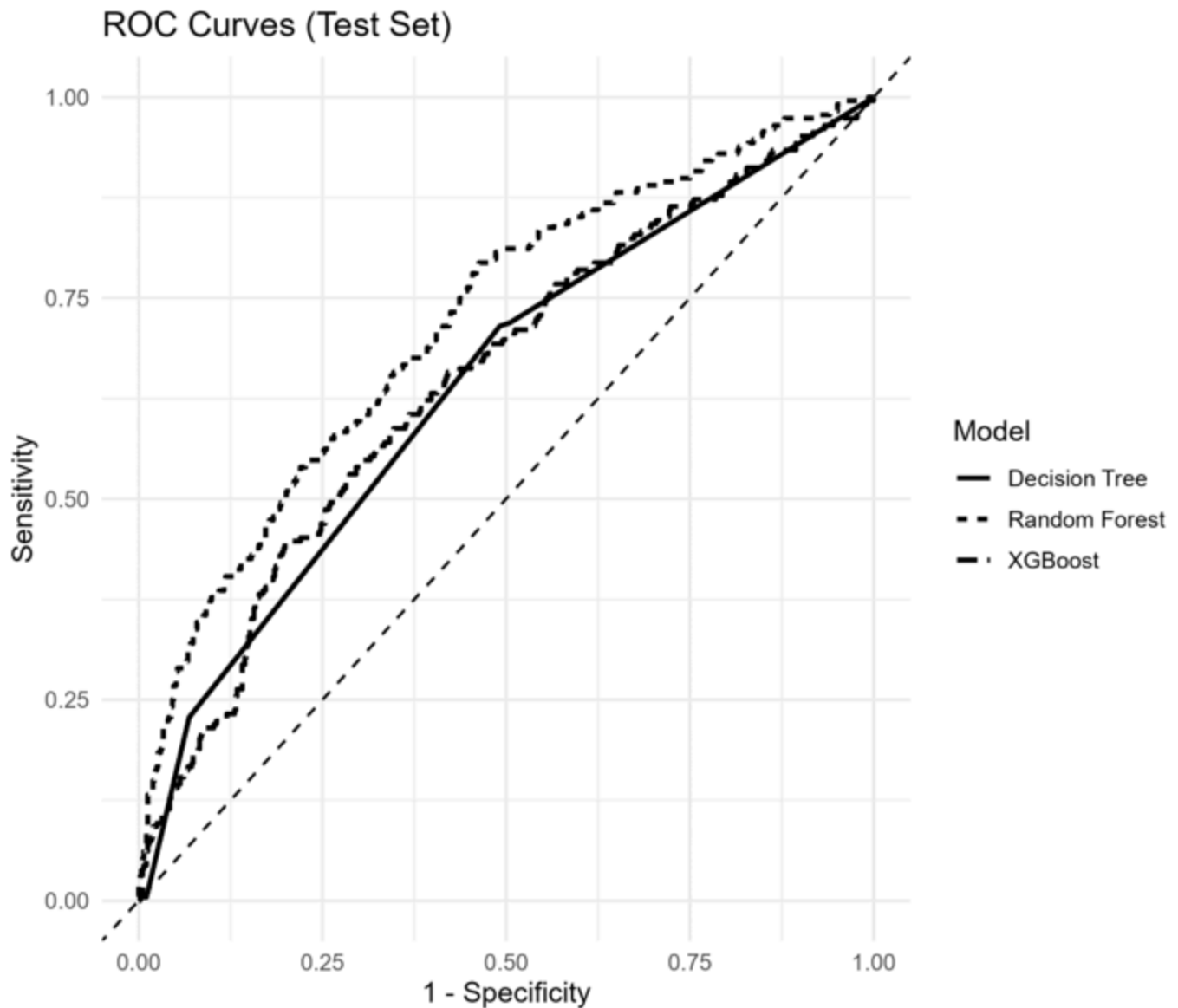


Figure 1

The optimal right ventricular outflow tract (RVOT) conduit in the Ross procedure remains controversial. This study addresses outcomes of decellularized (DH) and cryopreserved homograft (CH) for RVOT reconstruction.

Methods: Patients aged 8 years or older who underwent the Ross procedure using DH or CH were reviewed. Conduit durability and pulmonary valve function were compared between both groups

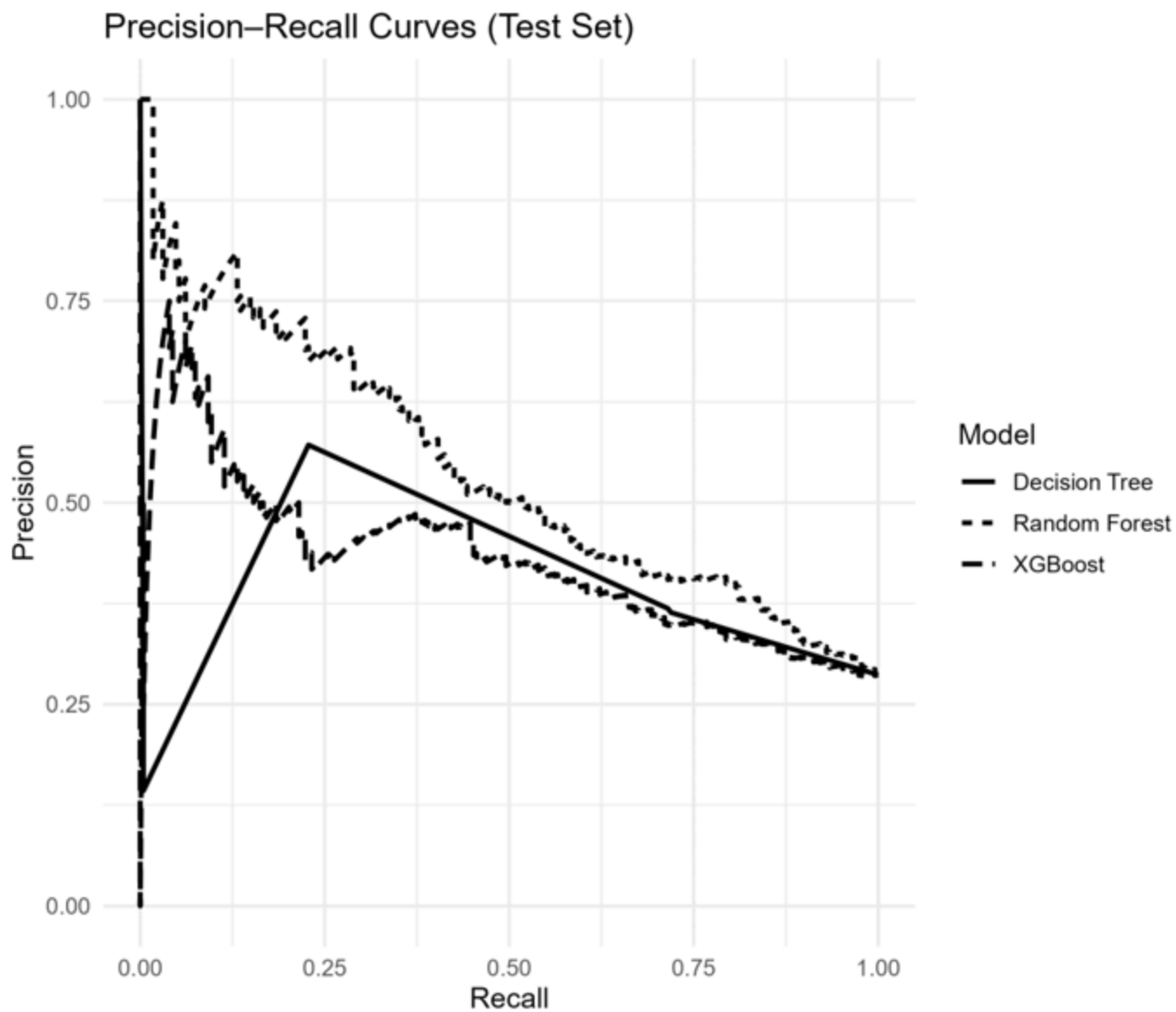


Figure 2

Precision–recall curves on the test set (three models).

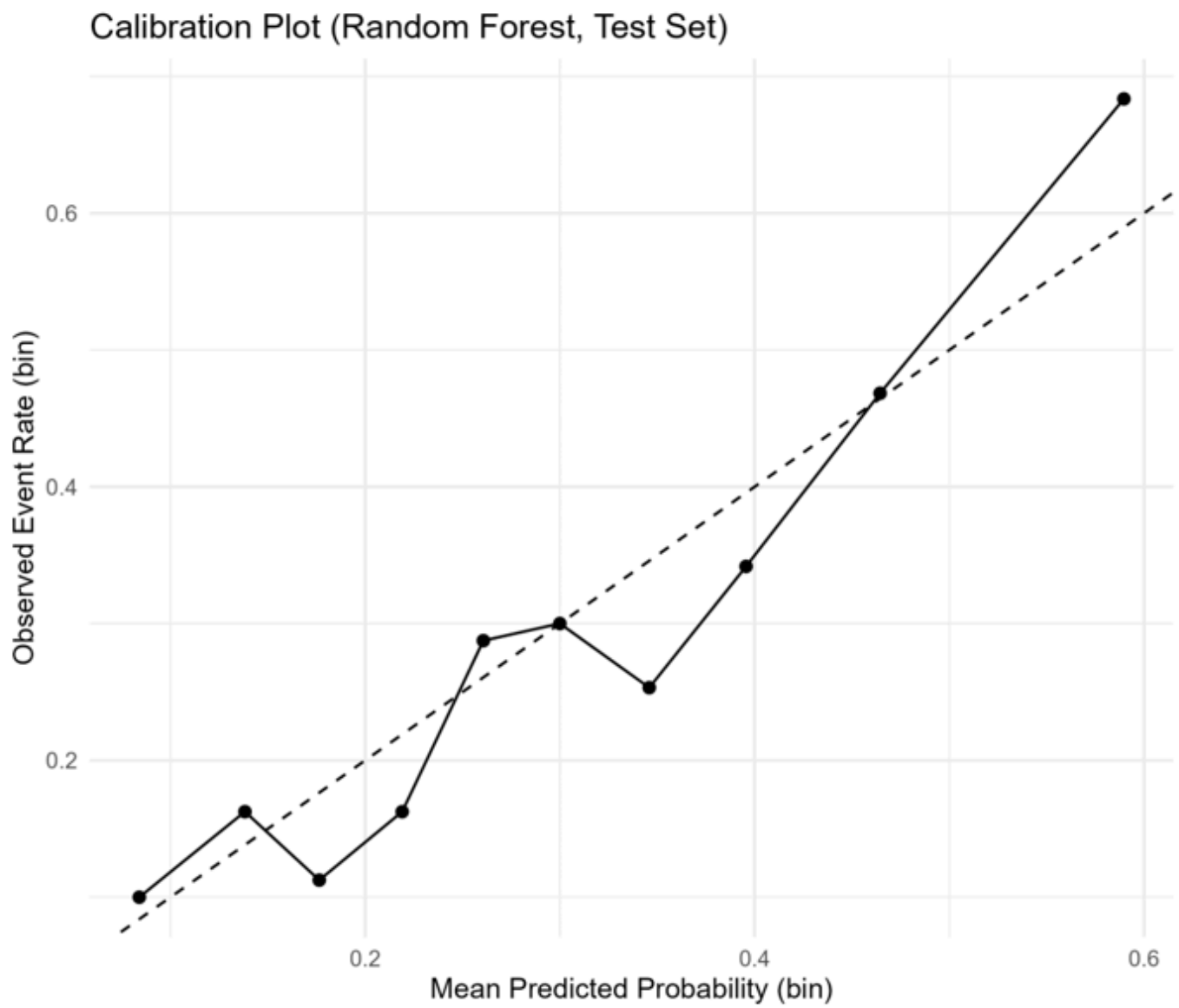


Figure 3

Calibration plot for the random forest model on the test set.

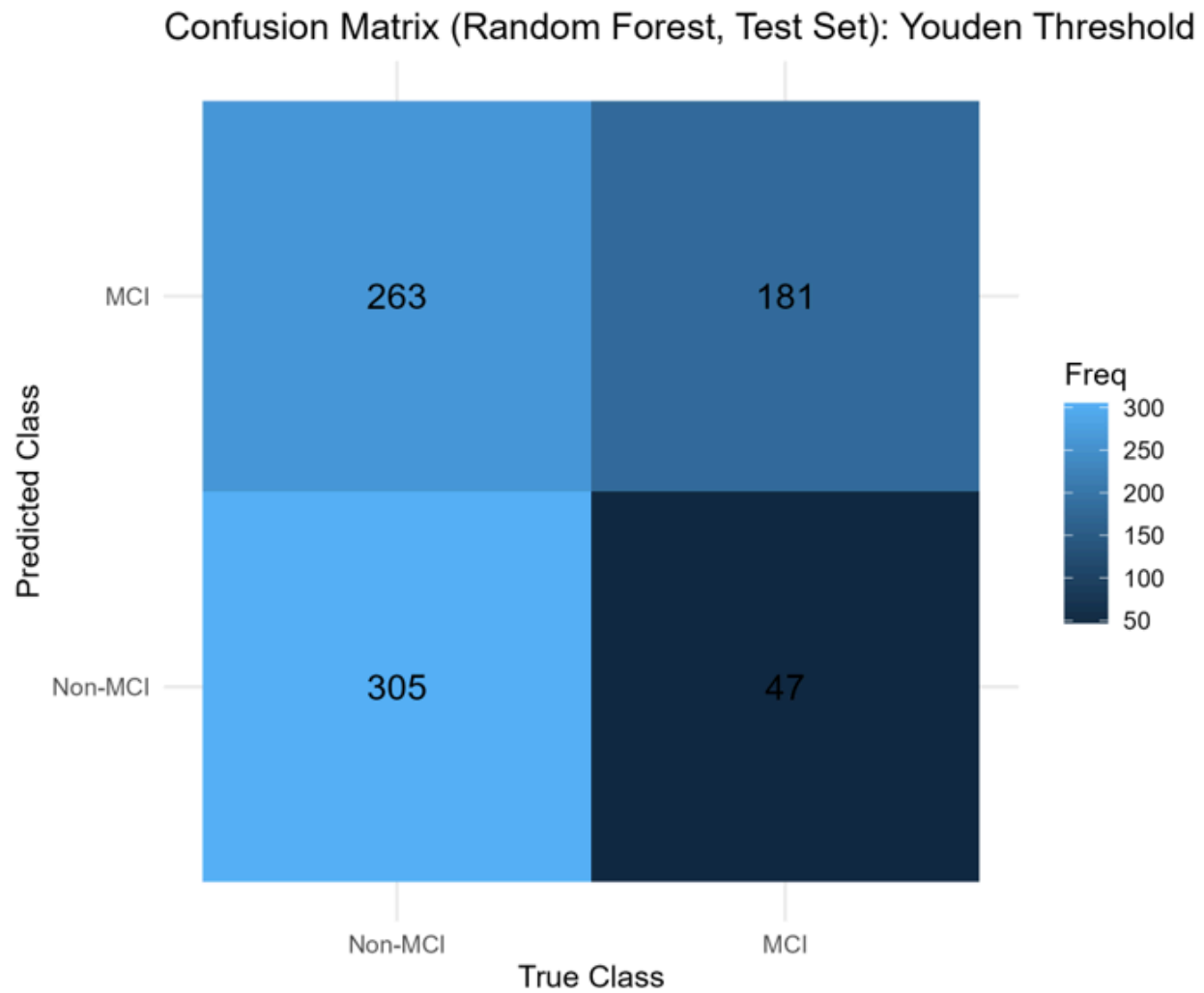


Figure 4

Confusion matrix for the random forest model at the Youden threshold on the test set.

Supplementary Files

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

- [Additionalfile2TableS1updated.docx](#)