

Tianyu Liu¹, Huitong Ding^{2,3}, Lindsay A. Farrer^{1,3,4,5,6,7,8}, Phillip H. Hwang^{1,2,3}

¹ Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA; ² Department of Anatomy & Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA;

³ The Framingham Heart Study, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; ⁴ Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA; ⁵ Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; ⁶ Biomedical Genetics, Department of Medicine, Boston University Medical School, Boston, MA, USA; ⁷ Alzheimer's Disease Research Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA, Boston, MA, USA; ⁸ Department of Neurology and Ophthalmology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA

Introduction

- Dementia poses a growing public health burden, underscoring the need for accessible and scalable approaches to early risk prediction.
- Current biomarker studies rely on specialized assays that are not routinely available in primary care, whereas blood measures obtained from complete blood count with differentials during standard clinical visits may offer a low-cost and widely deployable alternative.
- We aim to apply survival random forest models to predict time to incident dementia using common blood measures in the Framingham Heart Study (FHS).

Methods

1. Study Population

- Dementia-free adults aged over 65 years from the Generation 2 cohort who attended Exam 9 between 2011 and 2014.
- All participants had complete information on 14 blood-based measurements and were followed for incident all-cause dementia through December 2023.

2. Blood measurements predictors: Aspartate aminotransferase (AST), alanine aminotransferase (ALT), white blood cell count, red blood cell count, hemoglobin, hematocrit, platelet count, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count, glucose, and creatinine.

3. Outcome: Incident all-cause dementia based on expert consensus using DSM-IV criteria

4. Statistical Analysis

- We conducted the survival random forest using 5-fold cross-validation and performed a grid search to optimize key hyperparameters.
- We adjusted for age at baseline and sex as additional features.
- Time dependent discrimination was evaluated using the area under the Receiver Operating Characteristic curve (AUC), and the evaluation times were defined as the 20th–80th percentiles of follow-up time.

Conclusions

- Although survival random forest models demonstrated only modest discriminative performance, these findings highlight the potential value of scalable and noninvasive blood-based measures for population-level dementia risk assessment.
- Future work should evaluate alternative machine learning approaches, incorporate additional clinical blood biomarkers, and leverage larger and more diverse cohorts to improve predictive accuracy for early screening and primary prevention.

Results

Table 1. Descriptive analysis of demographic characteristics and blood measures of the overall sample at Exam 9 (2011-2014)

	Overall Sample ^a (N=1605)		Overall Sample ^a (N=1605)
Sex		Lymphocyte count (thousands/uL)	1.6 (1.0)
Male	708 (44%)	Monocyte count (thousands/uL)	0.5 (0.2)
Female	897 (56%)	Eosinophil count (thousands/uL)	0.2 (0.1)
Age at Exam 9 (Years)	73.6 (6.3)	Basophil count (thousands/uL)	0.0 (0.1)
White blood cell count (thousands/uL)	6.1 (1.8)	AST (U/L)	22.8 (7.8)
Red blood cell count (millions/uL)	4.3 (0.4)	ALT (U/L)	20.8 (10.1)
Hemoglobin (g/dL)	13.6 (1.3)	Glucose (mg/dL)	104.3 (20.4)
Hematocrit (%)	39.9 (3.8)	Creatinine (mg/dL)	1.0 (0.3)
Platelet count (thousands/uL)	232.1 (62.0)	Incident All-Cause Dementia	94 (6%)
Neutrophil count (thousands/uL)	3.8 (1.3)	Time to Dementia Diagnosis (days)	3160.2 (885.1)

Abbreviations: dL, deciliters; g, grams; mg, milligrams; U, unit; uL, microliters

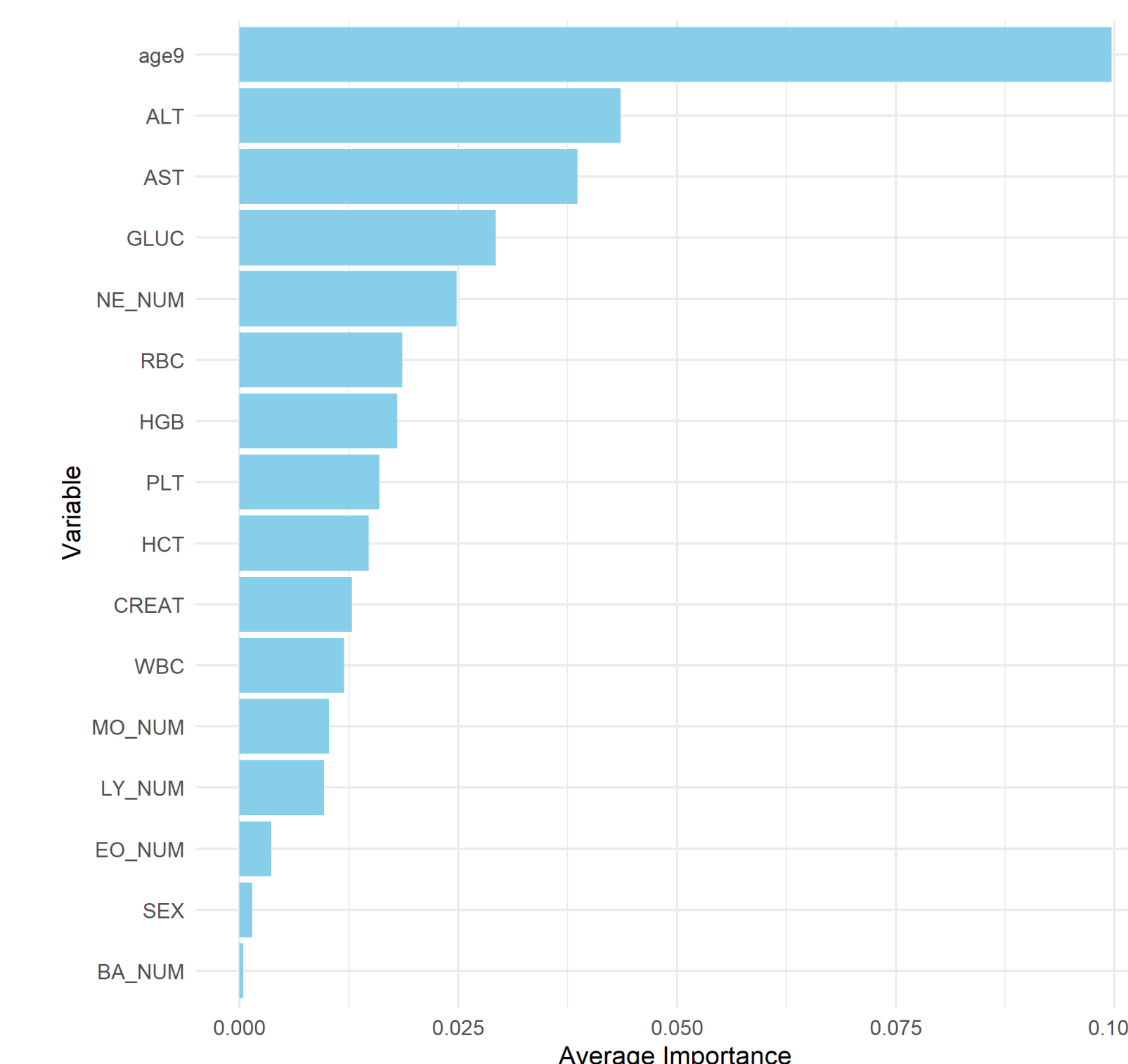
^a Continuous variables: mean (standard deviation); categorical variables: number (percentage)

Table 2. The 5-fold-average AUC values of each time point for the survival random forest prediction model

	Time point						
Percentile ^a	20 th	30 th	40 th	50 th	60 th	70 th	80 th
AUC	0.659	0.657	0.656	0.640	0.634	0.615	0.603

^aPercentile refers to continuous follow up time in the study categorized into percentile-based groups (i.e., 20th percentile is the first 20th percentile of follow up time in the sample, with each proceeding percentile group representing later follow up time)

Figure 1. Average variable importance across five folds in the survival random forest model



- Baseline age exhibited the greatest feature importance, followed by the four most influential blood biomarkers:
- AST, ALT, glucose, and neutrophil count.**

Abbreviations: age9, Age at Exam 9; AST, Aspartate aminotransferase; ALT, Alanine aminotransferase; GLUC, Glucose; NE_NUM, Neutrophil count; RBC, Red blood cell count; HGB, Hemoglobin; PLT, Platelet count; HCT, Hematocrit; CREAT, Creatinine; WBC, White blood cell count; MO_NUM, Monocyte count; LY_NUM, Lymphocyte count; EO_NUM, Eosinophil count; BA_NUM, Basophil count

Contact Information Tianyu Liu, MS: tianyul@bu.edu