

Research Background

- ❖ Dementia is rising rapidly worldwide, primarily due to an aging global population, creating a substantial social and healthcare burden.
- ❖ Early diagnosis of dementia enables early intervention, slowing disease progression.
- ❖ Electronic Health Records (EHRs) contain rich, routinely collected clinical data that offer a valuable resource for developing machine learning models to predict dementia risk.

Objective

This study aims to leverage advanced machine learning to develop dementia risk prediction models, utilizing longitudinal EHR data.

Challenges

- ❖ Many key variables in EHR have a significant proportion of missing values.
- ❖ Data is collected from multiple domains, leading to integration challenges.
- ❖ EHR often appears in unstructured formats, complicating data pre-processing and modeling.

UMass Memorial Cohort

- ❖ UMass Memorial Health System is located in central Massachusetts (Worcester County).
- ❖ EHR data is collected from 2017 to 2024.
- ❖ The cohort includes over 3.5 million patients.

Characteristics

Demographic Information	Cases (n=5,622)	Controls (n=44,976)
Age, mean (SD)	82(8)	82(8)
Female (%)	3,383(60.2)	27,026(60.1)
White (%)	4,926(87.6)	39,414(87.6)
Black or African American (%)	133(2.44)	1,010(2.2)
Asian (%)	91(1.6)	718(1.6)
American Indian or Alaska Native (%)	12(0.21)	75(0.16)
Not specify (%)	457(8.12)	3,755(8.34)

Research Overview

(a) Cohort Selection

UMass Memorial

Record from 2017 to 2024
>3.5m patients
>154m lab results
>47m diagnosis records
>48m medication records

Select patients & Study period

Index date definitions:
Cases: the first dementia diagnosis date.
Controls: the latest visit date by 12/31/2024.

(b) Patient Selection

UMass cohort until 12/31/2024 (n=3,489,127)

Record from 2017/10/01 to 2024/12/31 (n=1,286,062)

All patients whose index ages older than 65 (n=283,900)

Patients whose record year before index date is longer than 1 years (n=192,863)

Matching 8x controls based on age, sex, and race (n=50,598)

Dementia cases (n=5,622) Controls (n=44,976)

(c) Dementia Risk Prediction

Longitudinal EHR Data

Machine learning models

Dementia risks

Model 1: Comorbidities
Model 2: Model 1 + lab tests
Model 3: Model 2 + lab tests frequencies

Results

(a) XGBoost ROC curve: Model 1 (AUC = 0.56), Model 2 (AUC = 0.68), Model 3 (AUC = 0.80)

(b) Random forest ROC curve: Model 1 (AUC = 0.58), Model 2 (AUC = 0.66), Model 3 (AUC = 0.75)

(c) Logistic regression ROC curve: Model 1 (AUC = 0.58), Model 2 (AUC = 0.63), Model 3 (AUC = 0.71)

(d) SVM ROC curve: Model 1 (AUC = 0.57), Model 2 (AUC = 0.66), Model 3 (AUC = 0.74)

Figure 2: ROC curves using different classifiers : (a) XGBoost, (b) Random forest, (c) Logistic regression, (d) SVM; Model configurations: Model 1 (comorbidities), Model 2 (Model 1 + lab test results), and Model 3 (Model 2 + lab test frequencies).

Key findings:

- ❖ XGBoost achieved the highest performance across all models.
- ❖ Adding number of lab measures strongly enhanced prediction model performances.

Interpretation:

- ❖ Monitoring the frequency of lab tests may reveal underlying comorbidities, thereby aiding in dementia risk prediction.

Important Features

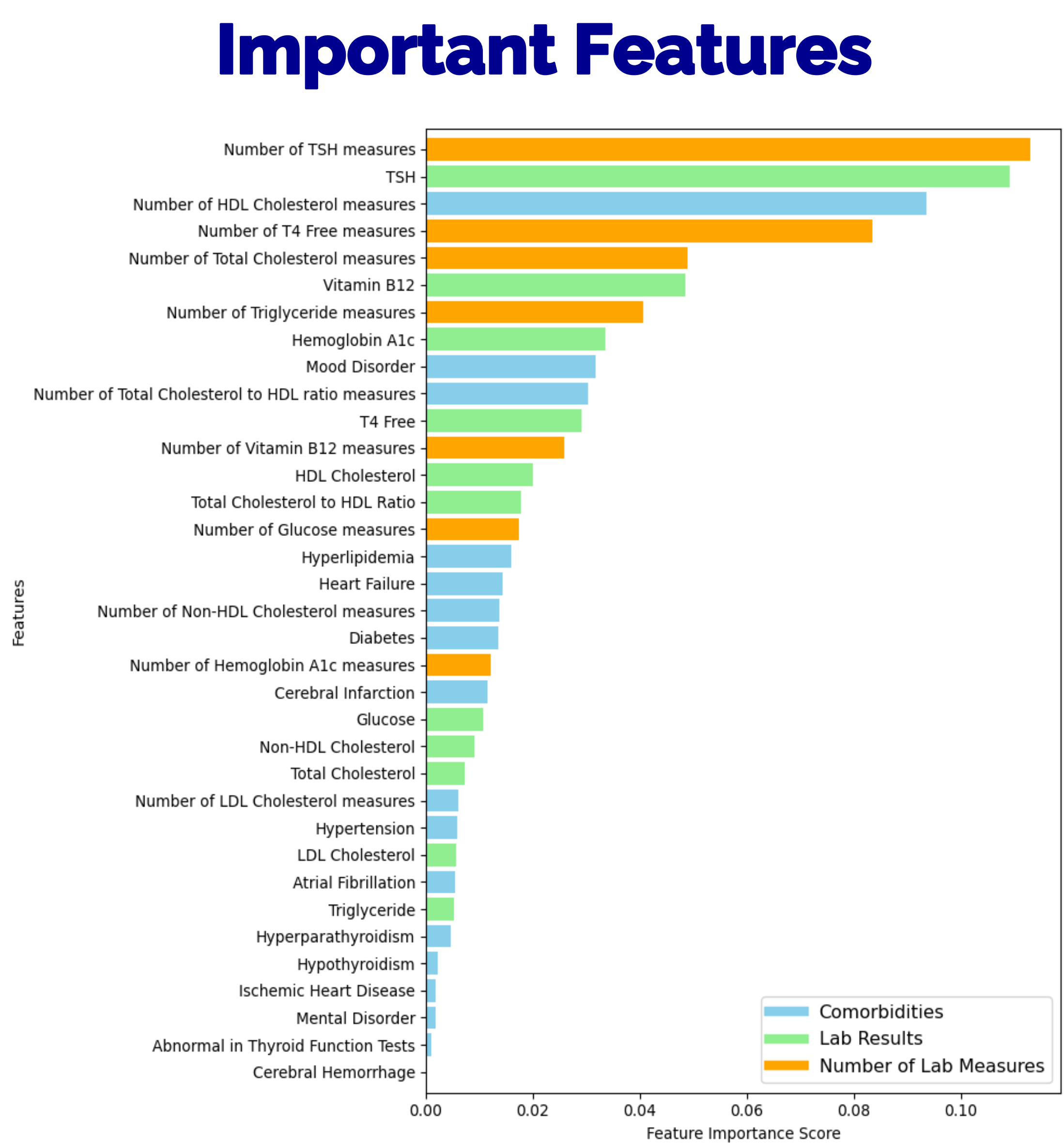


Figure 4: Feature importance generated by XGBoost: Feature importance of comorbidities are shaded in blue, lab tests are shaded in green, number of lab measures are shaded in orange.

Stratify Analysis

AUC by XGBoost	Model 1	Model 2	Model 3
Male	0.57	0.70	0.80
Female	0.56	0.69	0.80
White	0.56	0.70	0.81
Non-white	0.60	0.70	0.77

Conclusion

- ❖ Machine learning models that integrate comorbid conditions, longitudinal laboratory tests, and medication usage patterns show their potential in predicting dementia risk.
- ❖ The findings highlight the potential of routinely collected EHR data as a scalable resource for identifying individuals at elevated risk of dementia.