N3C Data Analytics for Attribute Importance and Prediction Tasks

Mirna Elizondo, Chloe Jones, Jelena Tešić
Department of Computer Science, College of Science and Engineering, Texas State University

DataLab12.github.io

Center for Analytics & Data Science

Abstract

We developed a robust method to summarize high-dimensional N3C EHR data, focusing on comorbidities of Diabetes and Sleep Apnea. Our evaluation of six models across four labeling scenarios, along with **Hamming-distance-based clustering**, enhances feature aggregation. Our **stacked ensemble model** achieved impressive results, with a precision of 0.92 and a recall of 0.93 for Diabetes predictions. While the accuracy for Sleep Apnea predictions is weaker, the insights remain valuable. Overall, our findings demonstrate that N3C EHR data **effectively predict diabetes and its comorbidity with Sleep Apnea**, providing a scalable solution for clinical decision support.

Motivation

The National Clinical Cohort Collaborative (N3C) is a major electronic health record (EHR) resource with over **20 million records** and faces challenges such as **imbalanced attributes, missing data, and inconsistent concept IDs**.

Figure 1: N3C EHRs show that a significant number of female (left) and male (right) patients have Sleep Apnea AND Diabetes.

Diabetes and sleep apnea share overlapping risk factors (e.g., obesity, inflammation) and form a complex, cyclical relationship, and are heavily underdiagnosed in the US population.

Label	SA	D	SAD	SAORD
Females	201,272	380,838	74,522	507,317
Males	201,258	380,556	74,515	507,224
Diag. Total	402,530	761,394	149,037	1,014,541
No Diag. Total	4,136,608	3,737,217	4,364,040	3,509,967

Table 1: N3C Patients Identified as 'Not Present' or 'Present' for Sleep Apnea (SA), Diabetes (D), Both (SAD), Either (SAORD).

Objectives

Attribute space reduction: Aggregate correlated EHR concepts to shrink the high-dimensional attribute space.

Attribute importance: Identify key attributes driving Sleep Apnea–Diabetes overlap and union.

Predictive modeling: Compare models for single and combined conditions to assess comorbidity prediction.

Bias and gender differences: Examine gender-specific patterns and potential bias in Sleep Apnea and Diabetes care.

TEXAS STATE UNIVERSITY
MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

Data

Data frame	Concept ID #	Concept Name #	Total Patients
Devices	6,462	5,034	178,602
Measures	26,122	25,839	251,832
Observations	14,094	13,925	237,155
Procedures	57,102	56,469	242,632
Conditions	54,600	54,585	249,545
Drugs	37,242	36,948	245,362
Visits	57	56	177,057

Table2: N3C Concept ID and Concept Name Counts by Data Frame After The Preprocessing Module

Target Label	Con.	Dev.	Drug	Meas.	Fem	Male
SA	11	69	29	2	55	56
D	46	62	38	21	91	76
SAD	4	55	46	4	66	51
SAORD	44	61	42	20	91	76

Table 3: Count of N3C Concepts selected by Random Forest by data frame, gender, and Sleep Apnea (SA), Diabetes (D), Both (SAD), Either (SAORD).

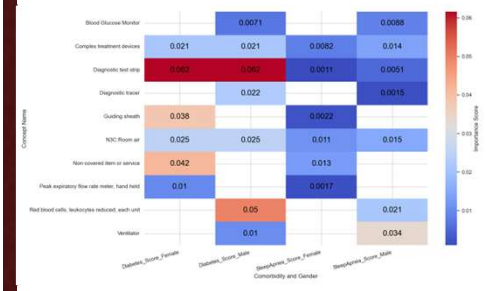

Figure 2: Correlation Heatmap of N3C Attribute Importance Scores By Gender for Sleep Apnea (SA), Diabetes (D), Both (SAD), Either (SAORD).

Methodology

The methodology comprises three key processing modules: preprocessing, attribute selection, and predictive modeling. Each module addresses a specific aspect of data preparation and analysis to support accurate modeling of electronic health records (EHRs).

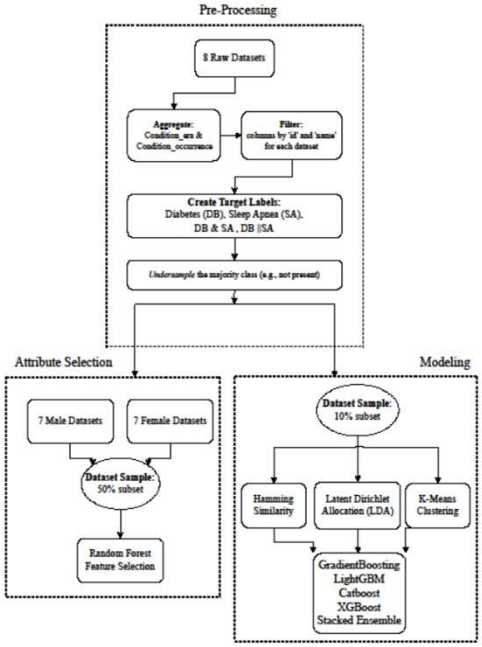

Figure 3: Flowchart of the proposed methodology: preprocessing is needed for both attribute selection and modeling tasks.

Findings

Target	Best Model	F1 Score
SA	ST-LDA / GB-LDA / LGBM-LDA / RF-LDA / XGB-LDA	0.71
D	ST-LDA	0.89
SAD	LGBM-Hamming	0.76
SAORD	ST-LDA	0.86

Table 4: F1-scores for CatBoost and Stacked Ensemble models across LDA, Hamming similarity, and K-means techniques for targets SA, D, SAD, and SAORD.

- Across four label groups (D, SA, SAD, SAORD), attribute importance revealed **gender-specific treatment pattern**s and **links between the comorbid conditions.**
- **Sleep Apnea:** both genders depended on diagnostic and treatment tools but followed **different management** strategies.
- **Diabetes:** males emphasized **respiratory devices**, whereas females focused more on **blood-related measurements**.
- **Sleep Apnea and Diabetes (SAD): CPAP**-related concepts were more important for males, while trauma-related concepts (**skull fractures**) were more important for females.
- **Sleep Apnea OR Diabetes:** device-based concepts dominated for both genders; specific devices differed between males and females.

Summary and Future Work

This study demonstrates that data-driven approaches can significantly enhance targeted healthcare strategies by identifying gender-specific treatment needs and the critical intersections between comorbid conditions. The Stacked ensemble model (ST) effectively highlighted these insights. **Future work** includes expanding the method to identify misdiagnoses and missed diagnoses in EHRs.

Experimental Results

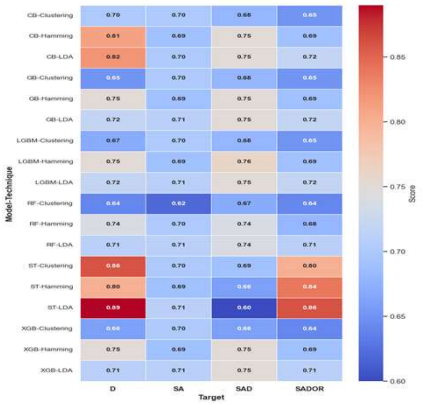

Figure 4: Distribution of F1-scores for the positive class for CatBoost and the Stacked Ensemble across three techniques—LDA (top), Hamming similarity (middle), and K-means clustering (bottom). Results are shown for the target labels Diabetes (D), Sleep Apnea (SA), and SAD

Acknowledgments The analyses were conducted with data accessed through the NCATS N3C Data Enclave, NCATS Contract No. 75N95023D00001 and Axle Informatics Subcontract: NCATS-P00438-B . The research was supported in part by NSF Expand AI #2334268.