N3C Data Analytics for Attribute Importance and Prediction Tasks

***Abstract*—This paper introduces a novel method for summarizing noisy tabular electronic health record (EHR) data characterized by large volumes of instances and interrelated attributes. We propose solutions for two key challenges: determining attribute importance and conducting predictive modeling under fixed capacity constraints, such as those imposed by the N3C Palantir platform. Our study targets two comorbidities, Diabetes and Sleep Apnea—demonstrating considerable overlap in their records. We evaluate six models across four label scenarios: the two comorbidities separately, their intersection, and their union. To address these, we introduce Hamming distance-based clustering as a method for attribute aggregation. Analysis of the four labels reveals gender-related concept attributes. Among the models tested, the stacked ensemble model, using Hamming distance clustering, outperforms others, achieving the highest precision (0.92) and recall (0.93) for Diabetes and strong results for the Sleep Apnea OR Diabetes diagnosis (precision 0.87, recall 0.90). The diagnoses for Sleep Apnea alone and Sleep Apnea AND Diabetes exhibit much lower precision and recall. The data-driven finding is that the N3C EHR contains enough information to effectively predict Diabetes and Diabetes AND Sleep Apnea diagnoses, but not to predict the other two. Next, our results underscore the effectiveness of combining Hamming distance clustering with stacked ensemble learning for handling noisy EHR data and predicting comorbidities. In summary, the proposed approach provides a scalable framework for enhancing clinical decision-making and identifying complex health patterns, potentially improving patient outcomes.**

***Index Terms*—gradient boosting, predictive modeling, noisy data, comorbidity**

I. Introduction

The intersection of computer science and healthcare has led to substantial advancements, driven by the proliferation of extensive electronic health records (EHRs) and medical data. The National Clinical Cohort Collaborative (N3C) is a national resource of real-world data researchers use to speed medical research. N3C systematically collects data composed of electronic health records (EHR) from different institutions and harmonizes this data into the Enclave. The Enclave offers space for collaborative studies and can be connected with external data sets, creating a multi-modal picture of health outcomes [1]. The N3C contains many EHRs (20+ million), so modeling the N3C EHRs as tabular data presents several challenges. These include highly imbalanced attributes, hundreds of thousands of characteristics, missing values, and variations in Concept IDs used for the same concept across different N3C data frames.

The modeling of the N3C EHRs as tabular data presented several challenges in the past: the highly imbalanced attributes, the number of characteristics in hundreds of thousands, missing values, and variations in Concept IDs used for the same concept across different systems resulted in dismal predictive performance [2]. Recently, NCATS has piloted the use of the N3C infrastructure for additional health conditions [1], as many patient records show a diagnosis of both Sleep Apnea and Diabetes (SAD). Figure 1 illustrates the N3C record distribution for diabetes and sleep apnea and their intersection for males and females. Almost identical distribution of the diagnoses among sexes will help us discover the bias in disease treatment, if any.

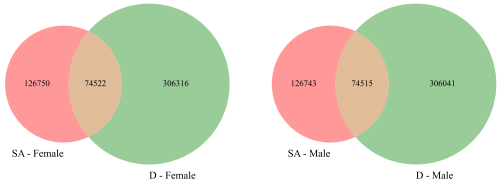

Fig. 1. N3C EHRs show that a significant number of female (left) and male (right) patients have Sleep Apnea AND Diabetes.

The *motivation* for this research is to compare and contrast the modeling for both conditions at their intersection (SAD) and union (SAORD) with the individual modeling Sleep Apnea (SA) or individually modeling Diabetes (D) to evaluate how effectively the analytics supported by the N3C EHR perform when the boundaries between comorbidities are unclear. Our proposed approach develops a predictive healthcare analytics pipeline. This study specifically focuses on the N3C platform and its associated data, which presents several challenges due to its scale. The data consists of millions of patients, several hundred thousand attributes, and billions of entries, significantly limiting data processing.

First, we propose supervised and unsupervised learning to combine the highly correlated concepts and reduce the number of unique concepts (labels) for the patient population. Second, we introduce the stacked boosting ensemble model to minimize the predictor list further and identify significant attributes for millions of patients in the N3C database whose EHRs contain diagnoses of Diabetes (D) and Sleep Apnea (SA) and their intersection and union. The contributions of this paper are (1) an attribute space reduction method for EHR synchronization, (2) an attribute importance method for analyzing comorbidity patterns, and (3) predictive modeling employed for different conditions and the prediction of both or either condition. This paper is organized as follows: Section II covers the related work, Section III presents the methodology, Section IV, and Section V outlines the proof of concept and the findings for Kaggle and N3C datasets. Conclusions, future work, and acknowledgments are provided in Section VI.

II. Related Work

Recent advancements in data cleaning and integration methodologies have contributed significantly to analyzing electronic health records (EHRs). Deep clustering algorithms such as Structural Deep Clustering Network (SDCN), Efficient Deep Embedding and Spectral Clustering (EDESC), and Self-supervised Heterogeneous Graph Pooling (SHGP) demonstrated the ability to enhance clustering and integration of heterogeneous healthcare data sources [3]. Clustering methodologies address data quality issues, offering insights into resolving discrepancies and standardizing biomedical metadata [4]. A significant problem in utilizing EHR-derived data is the potential for amplifying societal biases [5], so it is crucial to implement strategies to mitigate bias and ensure fairness in healthcare analytics.

The 2021 comprehensive reviews of data cleaning methods in healthcare underscore the importance of maintaining high data quality for reliable analysis and decision-making [6]. In parallel, clustering approaches for grouping patient records in EHR systems compared traditional algorithms with newer techniques tailored for large and complex healthcare datasets, suggesting that spectral clustering techniques could improve data organization, patient stratification, and healthcare decision-making processes in modern healthcare systems. [7]. Feature engineering and selection are critical to building effective predictive models from EHR data. Given the high dimensionality of EHR data, identifying relevant features is essential. SNOMED serves as the standard coding system. If mapping from another system is unclear, we map the concept to multiple SNOMED concepts. This inclusion approach results in duplication and limited enhancement for predictive models—an issue also noted in electronic health records (EHR) with the large number of ”concepts” related to specific conditions [8]. Note that codes often represent various aspects of the same condition rather than unrelated ones, and advanced machine learning techniques effectively accurately code and classify SNOMED CT concepts [9].

Machine learning models applied to EHRs have shown promise in early prediction of clinical deterioration, facilitating timely interventions, and improving patient care [10].

It has been seen in health-related studies that a wide range of methods are frequently applied; this domain uniquely utilizes classical machine learning models and predominantly favors pure undersampling techniques (with a notable exception in software), reflecting the high proportion of work in this area (34.3%) [11].

In contrast, deep learning models, like the Bidirectional Gated Recurrent Unit (GRU), predict future disease diagnoses by leveraging embeddings created through Word2Vec from structured medical vocabularies such as SNOMED CT. These embeddings incorporate longitudinal medical data and enable the integration of new features, such as binned observation values and social determinants of health, enhancing predictive accuracy [9].

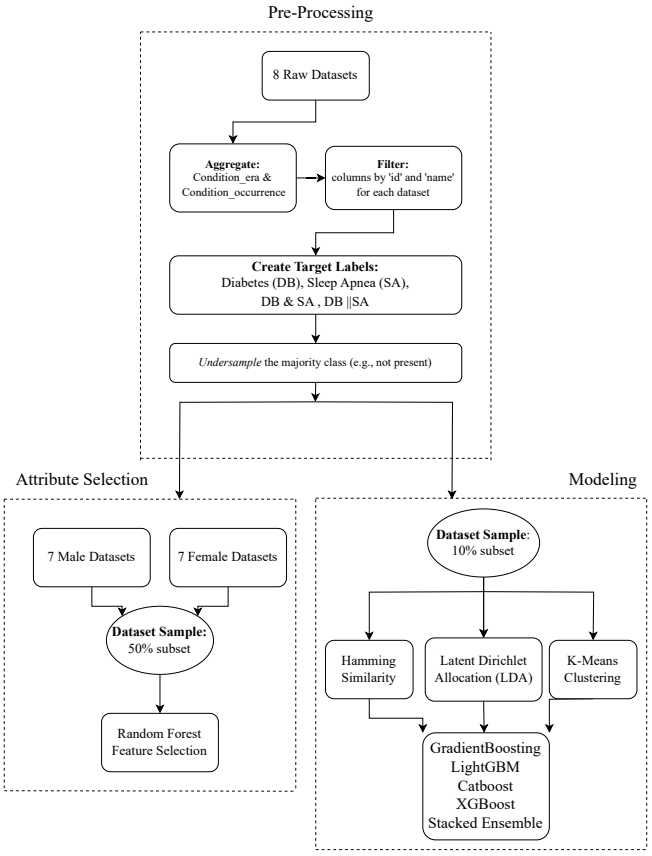

Fig. 2. Flowchart of the proposed methodology: preprocessing is needed for both attribute selection and modeling tasks.

III. 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).

The preprocessing module receives one aggregated data frame from Kaggle and eight raw EHR data frames from N3C, which include records for conditions, observations, visits, procedures, drugs, and measurements. It then aggregates the data, filters for relevant columns, creates target labels, addresses class imbalance, handles missing values, and normalizes the data. The attribute selection module samples records as necessary to account for processing constraints. Next, it applies the random forest-embedded feature selection method for male and female subjects separately to identify the most impactful attributes and mitigate bias. The modeling module also samples the data if needed to evaluate the predictive power of Random Forest (RF), Gradient Boosting (GB), LightGBM (LGBM), XGBoost (XGB), and CatBoost (CB), as well as a sequential model utilizing dense layers and ReLU activation functions from TensorFlow (TF) based on the preprocessed electronic health records (EHRs). We propose the Stacked ensemble model (ST) be used to combine the predictions from these gradient-boosting algorithms, aiming to achieve enhanced accuracy and robustness in the final predictions.

A. Preprocessing Module

TABLE I
MOTIVATION ORIGINATES FROM THE UNIQUE CHARACTERISTICS OF N3C DATA CATALOG EHRS, INCLUDING THE NUMBER OF CONCEPT IDS, CONCEPT NAMES, INSTANCES (ROWS), AND RAW DATASET COLUMNS (E.G., 'DATA_PARTNER_ID,' 'CONDITION_START_DATE,' 'CONDITION_END_DATE').

Source	Concept ID #	Concept Name #	Cols.	Rows
device_exposure	6,504	5,061	19	571,107,525
measurements	26,932	26,668	30	16,054,603,701
observations	14,098	13,935	25	3,228,355,577
procedures	57,819	57,259	19	1,248,124,769
condition_era	53,075	53,204	8	1,175,596,792
condition_occurrence	53,020	53,148	21	3,183,174,915
drug_era	37,406	37,111	9	1,138,041,266
visit_occurrence	57	11	23	1,786,570,960

Table I outlines the unique attributes of the N3C data, which include 22,649,720 patients and 33 significant visit records, as described in the N3C Palantir Enclave Data Catalog: Level 2 De-identified (DE-ID) Harmonized Data. The dataset contains numerous columns, but only the 'concept_id' and 'concept_name' are of primary significance. Notably, multiple 'concept_id' values often correspond to the same 'concept_name,' making it difficult to classify thousands of concepts into distinct, non-overlapping categories without expert medical input [2]. Therefore, this study uses the 'concept_name' as a unifying identifier to standardize entries across data frames in English.

Next, the electronic records face class imbalance challenges in target labels (condition by condition), as demonstrated in the N3C data Figure 1 for Sleep Apnea (SA) and Diabetes (D). We employ two primary techniques to address the class imbalance in our dataset: SMOTE (Synthetic Minority Oversampling Technique) and undersampling. These methods are designed to balance the class distribution and mitigate the potential for model bias toward the majority class (condition present). SMOTE generates synthetic samples for the minority class through interpolation between existing minority class instances. This technique effectively increases the number of minority class samples by creating new, synthetic examples that combine existing ones. In contrast to SMOTE, undersampling involves reducing the number of samples from the majority class to match the size of the minority class. This method aims to achieve a balanced class distribution by removing excess majority class instances. The undersampling fraction was calculated as followsle_fraction=minority class count/majority class count. Finally, the module handles missing values as null values by excluding them from the datasets. We used the MinMaxScaler to normalize the data by scaling feature values to the range [0, 1].

B. Attribute Selection

In this study, we employ a comparative analysis of three distinct techniques—Linear Discriminant Analysis (LDA), Hamming similarity, and K-Means clustering—each utilized separately to assess their effectiveness in topic modeling and reducing the number of concept codes in electronic health record (EHR) data. By evaluating these methods individually, we aim to uncover their strengths and weaknesses in handling the complexities of EHR data, especially concerning the relationships between conditions.

Hamming similarity groups similar conditions as a preprocessing step, while linear discriminant analysis (LDA) and K-Means clustering are used for dimensionality reduction and clustering. This study presents a novel method for analyzing Electronic Health Records (EHR) by comparing LDA for topic modeling, K-Means for clustering, and Hamming similarity for measuring distances between binary vectors to reduce the attribute space. Although previous research has explored combining LDA and K-Means for document clustering [12, 13], the use of Hamming similarity remains under-explored. Hamming similarity is compelling for binary-encoded EHR data, enhancing clustering performance by accurately capturing data point similarities [14]. The study also focuses on attribute space reduction by employing techniques such as Linear Discriminant Analysis (LDA), Hamming Similarity (Hamming), and K-means clustering (Clustering) to streamline datasets and improve model performance. LDA enhances class separability, while Hamming Similarity and K-Means Clustering group similar attributes to aid pattern recognition and predictive precision. For instance, consider two patients: Patient A and Patient B, in our dataset Patient A has health records for conditions C and D, where condition C is marked as 'Present' and condition D as 'Not Present.' In contrast, Patient B has condition records for conditions C and D, with condition D marked as 'Present' and condition C as 'Not Present.' Both patients are subsequently grouped into the same standard cluster. If two patients have records marked as 'Present' for both conditions C and D, they would be grouped into the same standard cluster based on the condition profiles for C and D. By using methods such as Hamming Similarity, Linear Discriminant Analysis (LDA), or K-Means Clustering.

This clustering approach groups similar condition profiles based on shared characteristics without considering individual patient-level differences in their records. The long-tail data source aggregation combines diverse datasets—clinical notes, lab results, and demographic information—into multiple sets for individual analysis. Each dataset is standardized to ensure consistency, and techniques like data imputation are applied to address missing values. Normalization allows for a more nuanced examination of the relationships between conditions, enabling targeted insights while maintaining the integrity of each separate dataset.

C. Attribute Importance Modeling

RF is an ensemble learning method that builds multiple decision trees and aggregates their predictions to enhance precision and robustness. Compared to individual decision trees, it effectively handles high-dimensional data and reduces overfitting. Gradient Boosting builds a series of weak learners to progressively improve the model by minimizing the loss

function gradient, which is known for its simplicity and robustness to overfitting. XGB builds on this with optimizations for speed and efficiency, making it suitable for large datasets [15]. LGBM further improves training efficiency with its gradient-based approach for tree splitting and support for categorical attributes [16]. CB addresses categorical attributes using ordered boosting and oblivious trees, reducing preprocessing needs [17]. The Stacked Ensemble model (ST) combines XGB, LGBM, and CB outputs in a two-stage process, followed by integration with a random forest model [18]. This systematic comparison not only emphasizes the distinct contributions of each technique but also evaluates them against various machine learning models, including ensemble methods and a stacked model. Ultimately, it enhances the interpretability and usability of EHR data for clinical decision-making and research.

D. Predictive Modeling

For the predictive modeling task, we adapt the Random Forest (RF), Gradient Boosting (GB), LightGBM (LGBM), XGBoost (XGB), and CatBoost (CB) individually and propose the StackedGBM (ST) model approach to combine multiple classifiers for enhanced predictive performance. These methods were chosen based on their proven effectiveness in handling complex datasets[2]. The models were trained and evaluated using different grouping techniques, such as Clustering, Hamming, and LDA, to account for underlying data distributions and noise. Leveraging the complete feature set ensured that potentially informative variables were not omitted, balancing the trade-off between model complexity and predictive power. Machine Learning methods use the full feature set after aggregation as it maximizes the model's ability to capture all available information from the data.

TABLE II
KAGGLE DATASET DIABETIC PATIENT COUNTS BY RACE AND GENDER [19].

Race→ Gender↓	Caucasian	African American	Hispanic	Other	Asian	**Total**
Female	40,822	11,728	1,092	748	318	**54,708**
Male	37,548	7,482	945	757	323	**47,055**
Total	78,370	19,210	2,037	1,505	641	**101,763**

IV. PROOF OF CONCEPT USING KAGGLE DIABETIC PATIENT HOSPITAL READMISSION DATASET

This smaller-scale study uses the Kaggle Diabetic Patient dataset [19] to evaluate data processing models and feature selection algorithms. The data contains the hospital admissions records for diabetic patients across 130 US hospitals over nine years. This dataset has 49 independent variables (patient demographics, admission details, and various medical and medication-related factors) plus one dependent variable indicating readmission within 30 days.

Table II outlines the breakdown of the admitted 101,763 diabetic patients concerning gender and race. The Kaggle dataset attributes and patient distribution are similar to N3C data [19] when we compare it to previous work [2]. We use the Kaggle data to determine which of the proposed approaches on predictive modeling, class imbalance, and attribute selections are justified to be deployed on the N3C Palantir Enclave regarding regional scalability and effectiveness.

TABLE III
INDICES USED IN TABLES AND FIGURES FOR LABELS, TECHNIQUES, AND MACHINE LEARNING MODELS.

Label	**Label Index**
Diabetes	D
Sleep Apnea	SA
Sleep Apnea and Diabetes	SAD
Sleep Apnea or Diabetes	SADOR
Technique	**Technique Index**
K-Means Clustering	Clustering
Hamming Similarity	Hamming
Linear Discriminant Analysis	LDA
Machine Learning Models	**Machine Learning Models Index**
Random Forest	RF
Gradient Boosting	GB
LightGBM	LGBM
XGBoost	XGB
CatBoost	CB
StackedGBM	ST
TensorFlow	T4

A. Experiment 1: Predictive Modeling

Table IV presents the modeling analysis results for all Kaggle patients, highlighting the performance across several key metrics.

TABLE IV
MODELING RESULTS FOR ALL PATIENTS (DIABETIC AND NON-DIABETIC) IN THE KAGGLE DATASET

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
RF	0.0114	**0.9018**	0.5667	0.0057
GB	0.0198	0.9015	0.4545	0.0101
XGB	0.0360	0.9003	0.3636	0.0189
LGBM	0.0146	0.9016	0.4783	0.0074
T4 (0.3)	**0.0440**	0.9003	0.3876	**0.0233**
T4 (0.5)	0.0074	**0.9018**	0.5789	0.0037
T4 (0.7)	0.0007	0.9017	**1.0000**	0.0003
T4 (0.9)	0.0000	0.9017	0.0000	0.0000

The analysis indicates that while the T4 models (specifically hidden layers 0.3 and 0.7 of the total number of input features) show some interesting results in terms of precision and recall, they suffer from poor balance between the two metrics, leading to low F1 scores. Additionally, traditional machine learning models (RF, GB, XGB, and LGBM, index in III) display more stable performance but still struggle with recall and F1 score, likely due to class imbalance or the challenge of predicting complicated conditions. Results suggest that further model tuning and consideration of class balancing techniques (such as SMOTE or undersampling) could be beneficial to improve overall performance. This is supported by recent research on class imbalance in cyberbullying detection, where resampling techniques improved classifier performance depending on the dataset size, imbalance ratio, and classifier used. [20]

TABLE V
SAMPLING EXPERIMENTS WITH THE KAGGLE DATASET BY MODEL: RF, GB, LGBM, XGB, AND CB, INDEX IN TABLE III.

Sampling Technique	Accuracy	Precision	Recall	F1 Score
Undersampled-RF	0.6140	0.8384	0.6140	0.6856
SMOTE-RF	0.8825	0.8215	0.8825	0.8394
Undersampled-GB	0.6383	0.8399	0.6383	0.7049
SMOTE-GB	0.8497	0.8126	0.8497	0.8292
Undersampled-LGBM	0.6383	0.8399	0.6383	0.7049
SMOTE-LGBM	**0.8887**	**0.8472**	**0.8887**	0.8412
Undersampled-XGB	0.6125	0.8384	0.6125	0.6844
SMOTE-XGB	0.8863	0.8319	0.8863	0.8407
Undersampled-CB	0.6172	0.8402	0.6172	0.6882
SMOTE-CB	0.8881	0.8440	0.8881	**0.8430**

B. Experiment 2: Class Imbalance Mitigation

In the Kaggle Dataset, there were three readmitted labels, *NO*, > 30', and < 30, but we converted the problem into a binary classification. The number of patients in the 'Not Present' class: 54864 samples (*NO* and > 30 combined), and in the 'Present' class: 11357 samples (< 30). We applied two sampling techniques to address the class imbalance: SMOTE and undersampling. Table V summarizes the results of our data sampling experiments. The goal was to evaluate which class balancing technique to use in the N3C Palantir Enclave, considering both scalability and effectiveness. We tested RF, GB, LGBM, XGB and CB models in this experiment, refer to Table V. Although SMOTE generally enhances performance metrics such as accuracy, precision, recall, and F1 score compared to undersampling, using SMOTE in the N3C Palantir Enclave environment led to crashes due to the large scale of the majority class. Thus, we selected the undersampling technique as it balances the trade-off between precision and computational efficiency.

TABLE VI
TOP 10 KAGGLE DATASET ATTRIBUTES BY MODEL: RF, GB, XGB, LGBM, AND CB, INDEX IN TABLE III.

Feature Name	RF	GB	XGB	LGBM	CB
Number Labs	0.0838	0.0722	0.0039	28592.7186	5.6019
Number In	0.0474	0.2124	0.0359	26762.5682	22.2504
Number Meds	0.0766	0.0686	0.0038	21755.7628	5.9754
Time	0.0598	0.0462	0.0045	14836.6168	5.5140
Age	0.0493	0.0316	0.0047	11270.5395	4.8927
Number Diag.	0.0401	0.0341	0.0053	10715.9261	3.8374
Disch. Home	0.0069	0.0255	0.0151	9239.5975	3.7517
Number Proc.	0.0395	0.0203	0.0041	7443.1871	2.4396
Number Out	0.0189	0.0192	0.0034	4167.5924	1.1496
Female	0.0167	0.0032	0.0036	3126.6036	0.3914

C. Experiment 3: Attribute Selection

Given the scale of the N3C dataset, we focused on selecting simpler algorithms that required less memory and processing power so they could scale better for the 22+ million patients. The attribute selection process identified key attributes that significantly enhanced model performance. The process filtered the data by gender to create two subsets (Male and Female) sampled to include 50% of N3C patients, 1,938,941 males, and 2,262,832 females, due to processing limits on the N3C Palantir Enclave. The attribute selection methods employed for this experiment are RF, GB, XGB, LGBM, and CB. The RF algorithm is used as the baseline for attribute selection. RF has been the fastest modeling algorithm on the N3C Enclave platform, requiring less memory and processing power than other machine learning methods [2]. Secondly, RF's inherent ability to handle large datasets and its embedded feature importance mechanism makes it ideal for selecting relevant attributes in high-dimensional datasets, such as the N3C Palantir Enclave data. Furthermore, Table VI shows that RF consistently identifies vital attributes (e.g., Number of Lab Procedures, Number of Inpatients, Number of Medications, Time in Hospital) across different models, reinforcing its reliability in feature selection. Using RF as the primary method for feature selection, we ensure that the process remains scalable and efficient while focusing on the most essential attributes.

V. PROOF OF CONCEPT USING N3C DATA BENCHMARK

N3C data The N3C patient enclave [1] leverages a standardized 'concept_id' system, designed to follow the Observational Medical Outcomes Partnership (OMOP) standard data model [21], to harmonize electronic health record (EHR) data from diverse sources across the United States. Researchers utilize this system to aggregate and analyze EHR data, extract cohorts based on specific criteria, and transform data into a uniform format compatible with OMOP. The EHR data often suffers from biases such as underrepresenting certain demographic groups, misclassification of conditions, inconsistencies in data entry practices, class imbalance, gender disparities, and performance bottlenecks associated with more complex algorithms [2].

While the N3C data harmonization process provides valuable insights, it has challenges. These include biases inherent in EHR data, such as underrepresenting certain demographic groups, misclassifying conditions, and inconsistencies in data entry practices. Additional challenges have been found in N3C processing, such as class imbalance, gender disparities, and performance bottlenecks associated with more complex algorithms [2]. Therefore, addressing these challenges is crucial for enhancing the precision of data analyses and addressing complex needs for different conditions or genders, such as Sleep Apnea (SA) and Diabetes (D).

TABLE VII
AVERAGE N3C PATIENTS IDENTIFIED AS 'NOT PRESENT' OR 'PRESENT' FOR SA, D, SAD, AND SAORD COUNTS, INDEX IN TABLE III.

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

N3C preprocessing: Relevant data frames (conditions, observations, devices, drugs, visits, procedures, and measurements) were selected from the N3C Palantir Enclave based on insights gained from the Kaggle dataset analysis (Section IV) and as

seen illustrated in [2]. A 50% sample of the data was used for feature selection, while a 10% sample was used for predictive modeling with an 80:20 train-test split. The distribution of patients in the 10% sample, categorized by gender, is presented in Table VII. To evaluate potential gender disparities and mitigate bias in feature importance, we split the dataset by gender (male and female), excluding other recorded gender categories (e.g., Unknown, Other), and calculated feature importance scores for each subgroup. The preprocessing module intakes eight raw EHR datasets of 18,075,334 EHR patient records. First, it aggregates the 'condition_era' and 'condition_occurrence' data frame on 'person_id,' 'condition_concept_name' and 'condition_concept_id' columns for each to include all conditions for each patient. Then, the module filters relevant columns for the remaining six datasets, including the 'person_id,' 'concept_name,' and 'concept_id' columns. To create a unique set of 'concept_name,' we removed duplicate entries from the datasets; detailed information can be seen in Table IX

TABLE VIII
DISTRIBUTION OF N3C PATIENTS FOR EACH DATA FRAME BY GENDER

Data frame	Total Male Patients	Total Female Patients
Devices	2,262,374	2,262,146
Measures	2,261,595	2,262,737
Observations	2,263,757	2,264,519
Procedures	2,261,084	2,262,161
Conditions	2,262,009	2,263,128
Drugs	2,261,767	2,262,303
Visits	2,261,244	2,261,873

Target Label Development In Table VII, there were 761,394 Diabetes (D) patients and 402,530 Sleep Apnea (SA) patients. There were 761,394 patients with both (SAD) and one million patients with either (SAORD). Thus, we had one million patients who have Sleep Apnea or Diabetes out of twenty-two million patients. To create the target labels, we utilized ten OMOP concepts for Sleep Apnea: 406 'complicated' and 126 'uncomplicated' OMOP concepts, and the intersection of patients who experienced the comorbidity the number of patients per label can be seen in Figure 1. After sampling, filtering, and label creation, the final number of patients is 18,075,334.

High overlap between diagnoses highlights the variations between physician documentation and system categorization in EHRs. For example, the Sleep Apnea label encompasses nine conditions, including 'Obstructive Sleep Apnea of adult,' 'Sleep apnea,' 'Obstructive Sleep Apnea (adult),' and 'Obstructive Sleep Apnea syndrome,' among others. Some of these conditions overlap, such as 'Obstructive Sleep Apnea of adult' and 'Obstructive Sleep Apnea (adult),' which may represent similar or identical clinical manifestations. The diabetes (D) label accounts for complicated and uncomplicated cases and encompasses 536 conditions. Table VII outlines details on patient counts for each label. Sleep Apnea is a well-documented comorbidity with Diabetes (D) and presents a compelling case for analysis due to its prevalence and impact on respiratory health. [22]. The conditions selected for both target labels were identified using N3C Concept Sets, which facilitate the classification and organization of relevant conditions for analysis.

TABLE IX
N3C CONCEPT ID AND CONCEPT NAME COUNTS BY DATA FRAME AFTER THE PREPROCESSING MODULE

Data frame	ConceptId #	ConceptName #	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

This work analyzes the differences in attribute importance for the following FOUR labels: (1) Sleep Apnea (SA), (2) Diabetes (D), (3) Diabetes AND Sleep Apnea (SAD), and (4) Sleep Apnea OR Diabetes (SAORD) to mitigate bias and ensure a comprehensive understanding of the factors influencing these conditions.

Attribute Summary The presence of concept names such as 'Obstructive Sleep Apnea syndrome' and 'Obstructive Sleep Apnea syndrome' underscores a significant challenge in medical data analysis, particularly in distinguishing between variations in terminology, potential duplicate entries, or inconsistencies in data labeling, which can lead to misinterpretation or redundancy in the study. To mitigate confusion encountered by machine learning models, we employed clustering and grouping techniques to address concept ambiguity. The 'conditions' and 'measurements' sources contribute four selected unique features each, whereas the 'devices' and 'drugs' sources present a broader range with 55 and 46 unique features, respectively. Thus, the 'devices' and 'drugs' data frames provide a set of attributes crucial for predictive analysis. The gender-based split shows that females are associated with 66 features greater than zero, compared to 51 for males. The difference suggests that features related to females are more prevalent or considered more relevant in the dataset. An importance threshold of 0.01 was selected as it offers a practical compromise. Table X summarizes the number of attributes selected per source and by gender, focusing on the most impactful characteristics without causing system performance issues. For example, lowering the threshold for the Sleep Apnea attributes further to 0.001 resulted in 113 features, and a threshold of 0.0001 increased the number to 121 features across both genders. For devices related to treatment for both Diabetes (D) and Sleep Apnea, the importance scores varied significantly between genders. For instance, the device "SURECHEK BLOOD GLUCOSE MONITOR" showed a higher importance score for males (0.0123) compared to females (0.0019). In contrast , the device "N3C: Room air" had a higher importance score for females (0.0377) than males (0.0116). This finding might indicate that the treatment differences are prevalent to genders. Similarly, the "Ventilator" also displayed a higher importance score for females (0.0068) compared to males (0.0041), indicating that this device may be used more frequently or be more critical for females in the management of their conditions. Further-

more, the ”BD PEN NEEDLE MINI 31GX3/16” showed a higher importance score for males (0.0128), while its female counterpart, the ”BD PEN NEEDLE SHORT 31GX5/16”, had a much lower score (0.0025). The variations in importance scores highlight how different devices may play varied roles in treating conditions across genders.

TABLE X
COUNT OF N3C CONCEPTS SELECTED BY RANDOM FOREST BY DATA FRAME, GENDER, AND SA, D, SAD, AND SAORD. INDEX IN TABLE III

Target Label	Source				Gender	
	Conditions	Devices	Drugs	Measurements	Female	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

A. N3C Attribute Selection By Gender

Table V illustrates the sampling experiments adopted for each data frame, and Table VIII shows the distribution of genders for each data frame in the study.

This understanding informs future decisions regarding data preprocessing and model training strategies, predicting the preprocessed to address class imbalance by downsampling. Various classifiers in predictive modeling for Sleep Apnea, including GB, XGB, LGBM, and CB, were trained using the cross-validated baseline hyperparameters from the Kaggle dataset. Their predictions were combined into a meta-dataset, which was then used to train a RandomForestClassifier meta-model. Model performance was evaluated using recall, accuracy, and F1 score metrics.

The data was then split into three subsets: female, male, and a combined set. The female and male subsets were sampled at 50% for attribute scoring, while the combined set was sampled at 10% for modeling. Our custom environment in the Enclave—configured with one core and 29,696 MiB memory for the driver and four executor instances with 2.4 cores and 15,360 MiB memory each—proved insufficient for the algorithms selected, given the scale of the data. These memory bottlenecks resulted in crashes, and without being able to increase the number of executors or cores, we were limited to which algorithms to employ. For this reason, we employed both supervised and unsupervised learning to merge highly correlated concepts and reduce the number of unique ideas. Finally, gradient boosting models and a stacked ensemble were used to evaluate predictive performance for each target label.

B. N3C Random Forest Attribute Scores

From our attribute selection and sampling method results using the Kaggle Dataset, we selected Random Forest and Gradient Boosting models to show the performance of the three attribute space reduction techniques. The attribute scores for various medical concepts were analyzed across different targets, specifically Diabetes (D), Sleep Apnea (SA), both (SAD), and either (SAORD) seen in Table X, XII, XIII, and XIV respectively.

TABLE XI
TOP 10 N3C CONCEPTS SELECTED BY GENDER FOR **D**, INDEX IN TABLE III.

FEMALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Red blood cells, leukocytes reduced, each unit	0.0793
Devices	Catheter, transluminal angioplasty, non-laser ...	0.0671
Measure	Oxygen saturation in Venous blood	0.0518
Devices	Non-covered item or service	0.0506
Conditions	Carotid artery obstruction	0.0451
Devices	CVS TEST STRIP	0.0434
Devices	Catheter, infusion, inserted peripherally ...	0.0413
Devices	Treatment devices, design, and construction ...	0.0388
Devices	Introducer/sheath, guiding, fixed-curve, other than peel-away	0.0383
Devices	N3C:Room air	0.0378
MALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	High flow oxygen nasal cannula	0.0954
Devices	CVS TEST STRIP	0.0798
Devices	Catheter, infusion, inserted peripherally ...	0.0610
Conditions	Renal disorder due to type 2 diabetes mellitus	0.0504
Drugs	3 ML insulin lispro 100 UNT/ML Pen Injector	0.0493
Devices	Technetium tc-99m tetrofosmin, diagnostic, per study dose	0.0379
Devices	Basic nasal oxygen cannula	0.0368
Conditions	End-stage renal disease	0.0347
Conditions	Chronic kidney disease due to type 2 diabetes 0.0329	
Devices	Non-covered item or service	0.0326

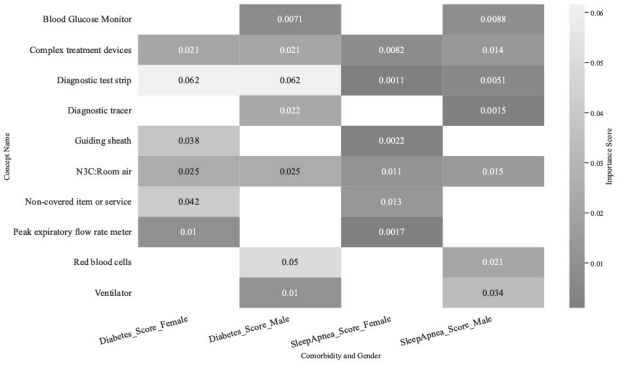

Fig. 3. Correlation Heatmap of N3C Attribute Importance Scores By Gender for D and SA, index in Table III.

C. N3C Target Connection Analysis

Identifying the most critical attributes for each predictive label, disaggregated by gender, provides valuable insights into comorbidity patterns. By examining shared attributes among Sleep Apnea (SA), Diabetes (D), both (SAD), and either (SAORD), we uncovered relationships between these conditions. Although predicting sleep apnea has been proven to be complex, leveraging knowledge of comorbidities may enhance our understanding of this condition. Table XI summarizes impact scores for attributes for Diabetes, Table XII for Sleep Apnea, and Table XIII for both (SAD), and Table XIV for either condition (SAORD). The concept of ”Alginate or other fiber gelling dressing, wound cover, sterile, pad size 16 sq. in. or less, each dressing” has importance scores of 0.0051

TABLE XII
TOP 10 N3C CONCEPTS SELECTED BY GENDER FOR **SA**, INDEX IN TABLE III.

FEMALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Technetium tc-99m macro aggregated albumin, diagnostic, per study dose	0.0458
Devices	Cervical, collar, semi-rigid, thermoplastic ...	0.0381
Devices	Oral thermometer, reusable, any type, each	0.0206
Devices	TRUE TEST GLUCOSE TEST STRIPS	0.0199
Devices	Patient programmer, neurostimulator	0.0191
Drugs	benzocaine 200 MG/ML Mucosal Spray	0.0132
Devices	Non-covered item or service	0.0127
Drugs	prednisone 20 MG Oral Tablet [Deltasone]	0.0120
Devices	ACCU-CHEK AVIVA PLUS TEST STRP	0.0115
Devices	N3C:Room air	0.0113
MALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Red blood cells, blood product	0.0524
Devices	Continuous positive airway pressure ...	0.0474
Devices	Low osmolar contrast material, 300-399 mg/ml iodine concentration, per ml	0.0181
Devices	Walking boot, pneumatic and vacuum, with or without joints ...	0.0179
Devices	Lead, pacemaker, other than transvenous vdd single pass	0.0178
Devices	Transfusion of Convalescent Plasma (Nonautologous) ...	0.0078
Devices	Injection, octafluoropropane microspheres, per ml	0.0077
Drugs	24 HR metoprolol succinate 200 MG Extended Release Oral Tablet	0.0069
Drugs	1000 ML glucose 100 MG/ML Injection	0.0066
Drugs	calcium levulinate	0.0063
Devices	CVS TEST STRIP	0.0051

TABLE XIII
TOP 10 N3C CONCEPTS SELECTED BY GENDER FOR **SAD**, INDEX IN TABLE III.

MALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Introducer/sheath, guiding, intracelectrophysiological logical	0.0922
Devices	Treatment devices, design, and construction	0.0520
Devices	Introducer/sheath, other than guiding, other than intracelectrophysiological logical	0.0357
Devices	Water seal drainage container and tubing for use with implanted chest tube	0.0348
Devices	Technetium tc-99m tetrofosmin, diagnostic, per study dose	0.0253
Devices	Miscellaneous supply or accessory ...	0.0239
Devices	Catheter, transluminal angioplasty, non-laser	0.0235
Devices	Continuing positive airway pressure unit	0.0216
Drugs	empagliflozin 10 MG Oral Tablet [Jardiance]	0.0195
FEMALE CONCEPTS		
Source	**Concept Name**	**Importance**
Conditions	Closed fracture of skull	0.0476
Measures	Klebsiella pneumoniae+Klebsiella variicola+Klebsiella quasipneumoniae DNA [NCncRange]	0.0442
Devices	Continuous positive airway pressure/Bilevel positive airway pressure mask	0.0420
Devices	Injection, perfluoro lipid microspheres, per ml	0.0266
Devices	Puraply am, per square centimeter	0.0244
Drugs	levothyroxine sodium 0.15 MG Oral Tablet	0.0241
Devices	Oxygen mask	0.0238
Devices	Adapter/extension, pacing lead or neurostimulator lead (implantable)	0.0230
Devices	Foot drop splint, recumbent positioning ...	0.0227
Devices	Non-coring needle or stylet with or without catheter	0.0225

for Diabetes and 0.0230 for Sleep Apnea in males, and "Closure device, vascular (implantable/insertable)" concept showed importance scores of 0.0031 and 0.0166 for Diabetes and Sleep Apnea in males, respectively, and 0.0112 and 0.0166 in females. These attributes were impactful for both outcomes. The concept "CVS TEST STRIP" is 0.0011 for Sleep Apnea, 0.0616 for Diabetes in females, 0.0051 for Sleep Apnea, and 0.0616 for Diabetes in males. The device "CVS TEST STRIP" is essential for monitoring glucose levels in Diabetes, and it might be necessary for Sleep Apnea in comprehensive patient management [23].

The "Introducer/sheath, guiding, intracelectrophysiological, fixed-curve, other than peel-away" concept has scores of 0.0022 for Sleep Apnea and 0.0383 for Diabetes in females. The objective of a sheath in medical procedures is to provide a protective and supportive conduit for the introduction and manipulation of other medical devices, such as catheters or wires [24]. The specification is that it is used in an 'intracelectrophysiological' test to detect the heart's electrical activity. Additionally," shows score also s of 0.0113 for Sleep Apnea, 0.0247 for Diabetes in females, and 0.0151 and 0.0247 for males. For a comprehensive view of these comparisons, refer to Table XI for the top 10 Diabetes attributes to Table XII for the top 10 Sleep Apnea attributes. The impact of concepts such as "Non-covered item or service," "Peak expiratory flow rate meter, handheld," "Red blood cells, leukocytes reduced, each unit," and "SURECHEK BLOOD GLUCOSE MONITOR" is different across the genders as seen in Table XI. Issues related to the costs and availability of medical devices, which vary by gender, can impact healthcare accessibility and management. The differences in device importance are influenced by clinical management protocols tailored to gender-related needs [25]. For a visual representation of these findings, refer to Figures 3. These figures illustrate the scores for concepts common to multiple conditions (SA and D), with the X-axis representing the concept names and the Y-axis showing gender categories (Female and Male). The color intensity within each cell indicates the score magnitude, facilitating the identification of disparities or similarities between genders across conditions. Table XI for the top 10 Diabetes attributes, and Tables XII for the top 10 Sleep Apnea attributes. Next, we compare and contrast the overall findings.

Shared Medical Concepts: Certain medical concepts are essential across multiple conditions, suggesting common underlying factors or overlapping pathophysiological mechanisms. For example, the 'Introducer/sheath, guiding, intracelectrophysiological' concept scored 0.0383 when scoring against the Diabetic label, see Table XI, and 0.0922 for both conditions (SAD), see Table XIII. In our analysis of female attribute importance for SAD, the feature 'Closed fracture of a skull'

was identified as significant, with an importance score of 0.0476. Thus, the 'Closed fracture of a skull' might reflect a shared or indirect factor that becomes relevant only when a patient has both SADs, highlighting potential interactions or risk patterns specific to patients with SADs. Next, the feature 'N3C: Room air' no longer appears in the top 10 features when analyzing the combined conditions of SAD. The 'N3C: Room air' may be relevant for distinguishing SAD on its own for female patients in Table XI and Table XII.

TABLE XIV
TOP 10 N3C CONCEPTS BY GENDER FOR SAORD, INDEX IN TABLE III.

MALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Anchor/screw for opposing bone-to-bone or soft tissue-to-bone	0.1721
Devices	Guidewire	0.0709
Conditions	Chronic kidney disease due to hypertension	0.0629
Devices	Tracheostomy mask, oxygen	0.0506
Devices	Ventilator	0.0465
Devices	Technetium tc-99m sestamibi, diagnostic, per study dose	0.0428
Conditions	Obstructive sleep apnea syndrome	0.0335
Conditions	Diabetes mellitus	0.0306
Devices	ONE TOUCH ULTRA TEST STRIPS	0.0290
Conditions	Inguinal hernia	0.0290
FEMALE CONCEPTS		
Source	**Concept Name**	**Importance**
Devices	Guidewire	0.0719
Devices	Treatment devices, design, and construction; complex (irregular blocks, special shields, compensators, wedges, molds or casts)	0.0687
Devices	Non-covered item or service	0.0671
Conditions	Atrial septal defect	0.0639
Devices	Surgical supply; miscellaneous	0.0584
Conditions	Chronic kidney disease due to type 2 diabetes mellitus	0.0523
Devices	Basic nasal oxygen cannula	0.0476
Devices	Face tent oxygen delivery device	0.0375
Devices	Red blood cells, leukocytes reduced, each unit	0.0357
Drugs	bumetanide 2 MG Oral Tablet	0.0348

Gender-Specific Differences: The scores varied significantly between males and females for the same medical concept and condition, highlighting the role of gender in disease manifestation and treatment efficacy. For example, "Closure device, vascular" shows different scores for Diabetes and Sleep Apnea across genders, as seen in Figure 3, which suggests that vascular complications might present differently or have other clinical implications in males versus females. The observed gender-specific differences highlight the importance of incorporating gender-specific data into predictive models.

Condition-Specific Significance: Data confirms that blood glucose monitoring is directly related to Diabetes management, as we hypothesized. Concept scores like "CVS TEST STRIP" have drastically higher scores for Diabetes compared to Sleep Apnea, both in males and females.

Potential Overlaps and Comorbidities: "N3C: Room air" has relevance for both Sleep Apnea and Diabetes, which might be due to the impact of Sleep Apnea on metabolic health and vice versa. The relevance of specific attributes for multiple conditions points to the potential for the joint analysis of comorbidities; refer to Table XIII.

D. N3C Target Diagnosis Prediction

The hyperparameters for the models used in the ST are the baseline cross-validated values from the gradient boosting models (GB, LGBM, XGB, CB). Specifically, GB has a learning rate of 0.2 and a maximum depth of 5, with 100 estimators; the LGBM has a learning rate of 0.2, with 50 levels and 50 estimators; XGB has a learning rate of 0.2, a maximum depth of 200, and a minimum child weight of 1; and CB has a learning rate of 0.05 and 200 iterations.

Table XV summarizes the performance of six models for three aggregation techniques for four different outcomes in terms of precision, recall, and accuracy. The ST model with Hamming distance clustering achieves the highest overall modeling score for the Diabetes (D) model, with 0.924 precision and 0.935 recall. We conclude that we can effectively model Diabetes target labels from N3C data and propose the ST with Hamming distance clustering for predicting Diabetes patients. Another strong target label is Sleep Apnea AND Diabetes (SAORD). Our proposed ST with the Hamming technique achieves 0.852 precision and 0.874 recall, while the ST-LDA combo achieves slightly better 0.875 precision and 0.900 recall. The Stacked model (ST) consistently demonstrates the highest average metrics across precision, recall, and accuracy for most conditions, indicating its effectiveness in combining the strengths of the underlying algorithms. Clustering as a technique consistently underperforms where Hamming and LDA are comparable for all models and labels. The distribution of F1 scores is illustrated in Figure 4.

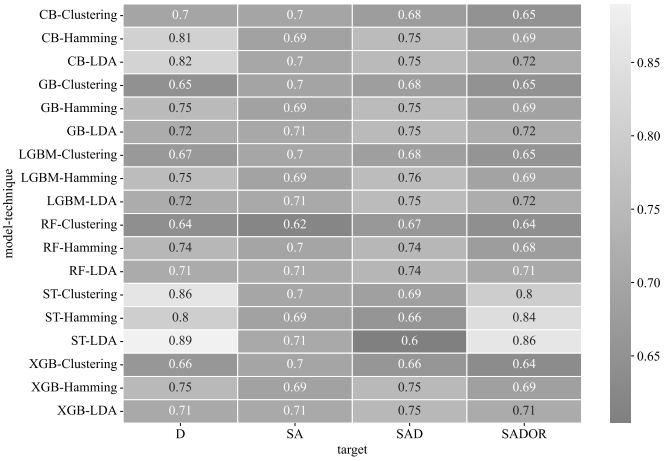

Fig. 4. Distribution of F1-Scores for the positive class using CatBoost and Stacked Ensemble model by Technique ((top) LDA, (middle) Hamming Similarity (bottom) KMeans Clustering) Results are displayed for the target labels Diabetes (D), Sleep Apnea (SA), and SAD.

VI. CONCLUSION AND FUTURE WORK

The paper focuses on analyzing the attribute importance for Diabetes (D), Sleep Apnea (SA), both conditions (SAD), or either condition (SAORD). Using smaller clea aner the Kaggle, hospital readmission dataset[19] we were able to make some

TABLE XV
N3C DISTRIBUTION OF AVERAGE PRECISION, RECALL, AND ACCURACY OVER TARGET LABELS (SA, D, SAD, SAORD) FOR THE POSITIVE CLASS 'PRESENT' USING MODELS (RF, GB, XGB, LGBM, CB, ST) BY TECHNIQUES (LDA, HAMMING AND CLUSTERING). INDEX MEANINGS IN TABLE III

Technique	Precision					Recall					Accuracy				
Clustering	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean
RF	0.6713	0.6286	0.6864	0.6601	0.6616	0.6985	0.6360	0.7033	0.6772	0.6788	0.5975	0.6081	0.6374	0.6115	0.6136
GB	0.6773	0.6467	0.6917	0.6658	0.6704	0.7082	0.6619	0.7025	0.6836	0.6891	0.5978	0.7396	0.6590	0.6234	0.6550
LGBM	0.6779	0.6486	0.6925	0.6663	0.6713	0.6957	0.6586	0.6982	0.6810	0.6834	0.6452	**0.7559**	0.6727	0.6323	0.6765
XGB	0.6763	0.6482	0.6735	0.6526	0.6627	0.6948	0.6611	0.6845	0.6693	0.6774	0.6363	0.7472	0.6378	0.6066	0.6570
CB	0.7211	0.6483	0.6915	0.6655	0.6816	0.7557	0.6611	0.7027	0.6829	0.7006	0.6501	0.7482	0.6573	0.6239	0.6699
ST	0.8731	0.6483	0.6788	0.8213	0.7554	0.9041	0.4814	0.6794	0.8608	0.7314	0.8229	0.6583	0.7063	0.7507	0.7346
Hamming	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean
RF	0.7438	0.7012	0.7427	0.6928	0.7201	0.7388	0.7024	0.7482	0.7050	0.7236	0.7495	0.6964	0.7418	0.6576	0.7114
GB	0.7549	0.7105	0.7536	0.7044	0.7309	0.7524	0.7072	0.7539	0.7149	0.7321	0.7548	0.6799	0.7507	0.6742	0.7149
LGBM	0.7572	0.7115	**0.7554**	0.7063	0.7326	0.7530	0.7063	**0.7591**	0.7141	0.7331	0.7608	0.6863	0.7530	0.6826	0.7207
XGB	0.7513	0.7104	0.7501	0.7020	0.7285	0.7485	0.7053	0.7504	0.7141	0.7296	0.7526	0.6819	0.7461	0.6697	0.7126
CB	0.8169	0.7095	0.7535	0.7042	0.7460	0.8304	0.7056	0.7568	0.7131	0.7515	0.7978	0.6800	0.7514	0.6777	0.7267
ST	**0.9236**	0.7102	0.6703	0.8523	**0.7554**	**0.9351**	**0.7332**	0.6727	0.8738	0.7314	0.7733	0.7048	0.6501	0.8210	0.7346
LDA	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean	D	SA	SAD	SAORD	Mean
RF	0.7127	0.7102	0.7427	0.7154	0.7203	0.7159	0.7070	0.7377	0.7232	0.7210	0.7199	0.7235	0.7482	0.6985	0.7225
GB	0.7187	0.7143	0.7536	0.7273	0.7285	0.7272	0.7051	0.7508	0.7396	0.7307	0.7160	0.7112	0.7539	0.7044	0.7214
LGBM	0.7192	0.7160	**0.7554**	0.7277	0.7296	0.7296	0.7058	0.7512	0.7393	0.7315	0.7138	0.7158	**0.7591**	0.7066	0.7238
XGB	0.7128	0.7151	0.7501	0.7238	0.7255	0.7250	0.7047	0.7476	0.7354	0.7282	0.7033	0.7162	0.7504	0.7031	0.7183
CB	0.8224	0.7140	0.7535	0.7239	0.7535	0.8425	0.7070	0.7494	0.7353	0.7586	0.8004	0.7071	0.7568	0.7021	0.7416
ST	0.8996	**0.7162**	0.6135	**0.8749**	0.7761	0.9239	0.7169	0.6143	**0.9013**	**0.7891**	**0.8660**	0.7064	0.5983	**0.8360**	**0.7517**

modeling decisions. The N3C Enclave [1] experiments underscore the importance of identifying comorbidity patterns and leveraging large-scale healthcare datasets to develop accurate, actionable predictive models. Overall, the ST model demonstrated its potential by uncovering meaningful connections between comorbid conditions, mainly through the comparison attribute scores across the four labels; the SADOR model was effect effectively distinguished with Diabetes (exclusive conditions), underscoring its utility in targeted healthcare interventions. Critical insight into gender-specific differences in diagnostic and treatment approaches:

Sleep Apnea (SA): Both genders consistently use diagnostic and treatment tools for Sleep Apnea, but distinct preferences emerge. Males and females adopt different strategies for managing their condition.

Diabetes (D): Gender-specific variations are evident, with males recorded to have a treatment that prioritizes respiratory and infusion devices while female care focuses on blood-related measurements and diagnostic tools. This divergence in management approaches underscores the need for personalized treatment strategies.

Intersection (SAD): Feature importance scores vary across different concepts, with certain items being particularly relevant. For instance, "Continuous positive airway pressure unit" stands out for males, while "Closed fracture of the skull" is notably significant for females. These scores emphasize the relative importance of each concept in the context of both conditions (SAD), offering valuable insights that could inform future research on their role in patient management.

Union (SAORD): In the female subset, the only drug-related concept is "bumetanide 2 MG Oral Tablet," indicating its role in managing fluid retention related to Sleep Apnea or Diabetes [26]. For both males and females, device-related concepts dominate. The top male concept, "Anchor/screw for opposing bone-to-bone or soft tissue-to-bone," has the highest importance (0.1721), followed by devices like "Guidewire" and "Ventilator." Similarly, in females, "Guidewire" leads with an importance score of (0.0720), along with "Treatment devices" and "Surgical supply."

In summary, we have shown that a data-driven approach could lead to more effective, targeted healthcare strategies, helping physicians prioritize treatments, reduce misdiagnoses, and ultimately improve patient outcomes. Our future work will investigate more advanced topic modeling techniques, autoencoders, or embedded next-generation sequencing (NGS) to capture those complex relationships. Additionally, utilizing SHAP or LIME could help capture complex relationships in the data, potentially improving model performance.

VII. ACKNOWLEDGMENT

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave (https://covid.cd2h.org) and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data, the organizations, and the scientists who have contributed [27].

REFERENCES

[1] N. I. of Health, "National covid cohort collaborative (n3c)," 2022, accessed: 2024-08-30. [Online]. Available: https://covid.cd2h.org/

[2] M. Elizondo, R. Musal, and et al., "Long covid challenge: Predictive modeling of noisy clinical tabular data," *The 11th IEEE International Conference on Healthcare Informatics*, 2023.

[3] H. T. Rauf, A. Freitas, and N. W. Paton, "Deep clustering for data cleaning and integration," 2023.

[4] W. H. et al., "Cleaning by clustering: a methodology for addressing data quality issues in biomedical metadata," *BMC Bioinformatics*, vol. 18, no. Suppl 14, p. 61, 2017.

[5] A. D. Boyd, R. Gonzalez-Guarda, K. Lawrence, C. L. Patil, M. O. Ezenwa, E. C. O'Brien, H. Paek, J. M. Braciszewski, O. Adeyemi, A. M. Cuthel, and et al., "Equity and bias in electronic health records data," *Contemporary Clinical Trials*, vol. 130, p. 107238, Jul 2023.

[6] B. Ehsani-Moghaddam, K. Martin, and et al., "Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data," *Health Information Management Journal*, vol. 50, no. 1-2, 2021.

[7] S. Ehrenberg, "Spectral clustering and variational autoencoders for compact patient representations from electronic health records," M. Eng. Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, May 2020, cataloged from the official PDF of thesis. Includes bibliographical references (pages 61-66). [Online]. Available: https://hdl.handle.net/1721.1/127522

[8] N. P. et al., "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in bioinformatics*, Jun 2022.

[9] E. Chang and J. Mostafa, "The use of snomed ct, 2013-2020: A literature review," *Journal of the American Medical Informatics Association*, vol. 28, no. 9, p. 2017–2026, Jun 2021.

[10] V. R. et al., "Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 163, no. 3, pp. 883–892.e11, Dec 2021. [Online]. Available: https://doi.org/10.1016/j.jtcvs.2021.10.060

[11] W. de Vargas et al., "Imbalanced data preprocessing techniques for machine learning: A systematic mapping study," *Knowledge and Information Systems*, vol. 65, no. 1, p. 31–57, Nov 2022.

[12] U. Buatoom, W. Kongprawechnon, and et al., "Document clustering using k-means with term weighting as similarity-based constraints," *Symmetry (Basel)*, vol. 12, no. 6, pp. 1–25, 2020.

[13] J. Li, W. Li, Q. Sun, and et al., "A document clustering approach based on hybrid lda and k-means," *Expert Systems with Applications*, vol. 149, p. 113263, 2020.

[14] N. Kumar, P. C. Panchariya, and et al., "An approach for documents clustering using k-means algorithm," in *Proceedings of the Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[15] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002360

[16] T. F. e. a. Guolin Ke, Qi Meng, "Lightgbm: A highly efficient gradient boosting decision tree," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3149–3157, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] L. Prokhorenkova, G. Gusev, and et al., "Catboost: Unbiased boosting with categorical features," Jan 2019. [Online]. Available: https://arxiv.org/abs/1706.09516

[18] X. Zhao and et al., "Stacked gradient boost machine for predictive modeling," in *Proceedings of the International Conference on Advanced Engineering*, 2022.

[19] S. Tayal, "Diabetic patients'readmissionn prediction," Aug 2020. [Online]. Available: https://www.kaggle.com/datasets/saurabhtayal/diabetic-patients-readmission-prediction

[20] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The effect of rebalancing techniques on the classification performance in cyberbullying datasets," *Neural Computing and Applications*, vol. 36, no. 3, p. 1049–1065, Nov 2023.

[21] O. C. W. Group, "Omop cdm v5.3," Aug 2021. [Online]. Available: https://ohdsi.github.io/CommonDataModel/cdm53.html

[22] N. C.-W. et al., "Association between undiagnosed obstructive sleep apnea and severe course of covid-19: a prospective observational study," *SleepBreath*, pp. 79–86, July 2023.

[23] A. D. Association, "Standards of medical care in diabetes—2023," 2023, accessed: 2024-08-19. [Online]. Available: https://professional.diabetes.org/standards-of-care

[24] A. Sardone, L. Franchin, and et al., "Management of vascular access in the setting of percutaneous mechanical circulatory support (pmcs): Sheaths, vascular access and closure systems," *Journal of Personalized Medicine*, vol. 13, no. 2, p. 293, Feb 2023.

[25] K. B. et al., "Gender differences in the utilization of health care services," *Journal of Family Practice*, vol. 49, no. 2, pp. 147–152, February 2000.

[26] Sep 2024. [Online]. Available: https://www.mayoclinic.org/drugs-supplements/bumetanide-oral-route/description/drg-20071274

[27] M. H. et al., "The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment," *Journal of the American Medical Informatics Association*, vol. 28, no. 3, pp. 427–443, 08 2020. [Online]. Available: https://doi.org/10.1093/jamia/ocaa196