Data-driven Clinical Attribute Importance Analysis

Mirna Elizondo*, Jelena Tešić* on behalf of N3C
* Department of Computer Science

***Abstract*—Objective: Machine learning (ML) presents unprecedented opportunities for predicting comorbidities and post-sequelae conditions in healthcare. This study explores advanced machine learning (ML) techniques to predict Diabetes (Type I and II) and Sleep Apnea using electronic health records (EHRs) from the N3C Enclave. To address high-dimensional data challenges, we employ feature reduction techniques such as Linear Discriminant Analysis (LDA), Hamming Clustering, and k-means Clustering. These techniques refine the feature sets, enhancing the predictive accuracy and robustness of the models. The analysis identifies significant predictors for each condition, improving our understanding of the comorbidities. A stacked ensemble model integrating Gradient Boosting, LightGBM, XGBoost, and CatBoost achieves superior performance, demonstrated by specific metrics (e.g., accuracy, AUC, F1-score). *Conclusion:* This study demonstrates the efficacy of combining feature reduction techniques with advanced ML models in predictive healthcare analytics. *Significance:* Overall, this study showcases the efficacy of integrating feature reduction techniques with advanced ML models in advancing predictive healthcare analytics. These findings hold promise for early intervention and management of co-morbid conditions, thereby significantly impacting clinical practice.**

***Index Terms*—gradient boosting, predictive modeling, noisy data, Long COVID, Diabetes, Sleep Apnea**

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. This paper explores the development of advanced predictive modeling techniques in healthcare analytics, utilizing large-scale datasets to address complex medical challenges. The COVID-19 pandemic has unveiled a phenomenon known as *Long COVID*,characterized by persistent symptoms following recovery from the acute phase of the disease. Addressing Long-term COVID and other post-acute sequelae (conditions resulting from a previous disease or injury) requires innovative approaches and interdisciplinary insights. Figure 1 presents estimates from the Household Pulse Survey, showing that the percentage of U.S. adults aged 18 and older who experienced Long COVID is less than 30% across various demographics including age, disability, education, gender identity, race, sex, sexual orientation, and state. Figure 1 indicates that Long COVID affects a diverse range of individuals across different population groups. *Mirna TODO: so what?* Our research initially focused solely on long-term COVID [1], but we encountered challenges in accurately predicting this condition due to the limited knowledge about it. The complexity and variability of these symptoms posed significant difficulties in creating precise predictive models. To develop robust and actionable predictive analytics, we shifted our focus to related comorbidities of long-term COVID, specifically Diabetes (Type I and II) and Sleep Apnea conditions.

In this paper, we address the inherent challenges of real-world medical electronic health record data, such as highly imbalanced attribute sets, missing values, and a long tail of rare attribute occurrences. First, we emphasize the need for robust data cleaning and integration methodologies. Next, we uncover patterns within these complex healthcare datasets to enhance predictive modeling capabilities and improve medical decision-making and patient outcomes. We examine the comorbidity patterns within the noisy, expansive National COVID Cohort Collaborative (N3C) Enclave collection of electronic health records (EHRs) from 22,649,720 patients, including 8,735,076 confirmed COVID-19 (+) cases and 216,806 possible COVID-19 (+) cases from 89 sites, spanning 33.1 billion rows [2]. For the initial proof of concept, we also utilize the Kaggle readmission dataset [3], focusing exclusively on diabetic patients.

Paper Contributions The study introduces feature space reduction techniques, including Linear Discriminant Analysis (LDA), Hamming Clustering, and k-means Clustering to create efficient feature sets. These methods effectively streamline the dataset, addressing the curse of dimensionality and enhancing model performance. LDA reduces dimensions by focusing on maximizing class separability, while Hamming Clustering and k-means Clustering group similar features, facilitating better pattern recognition and predictive accuracy. This multifaceted approach not only improves computational efficiency but also contributes to more robust and interpretable models for predicting Diabetes, Sleep Apnea, and Long COVID.

Paper Overview Initially, the research aimed to predict *Long COVID* but shifted to Diabetes and Sleep Apnea due to the complexity and similarity in Long COVID symptoms. Using extensive datasets from the National COVID Cohort Collaborative (N3C) [2] and the Kaggle readmission dataset [3], the study emphasizes robust data cleaning and feature space reduction techniques to address challenges in real-world medical data. Through feature selection, clustering, and advanced ensemble learning methods, the paper demonstrates improved predictive modeling capabilities, ultimately enhancing medical decision-making and patient outcomes. The findings underscore the importance of understanding comorbidity patterns and leveraging large-scale healthcare datasets to develop accurate, actionable predictive models.

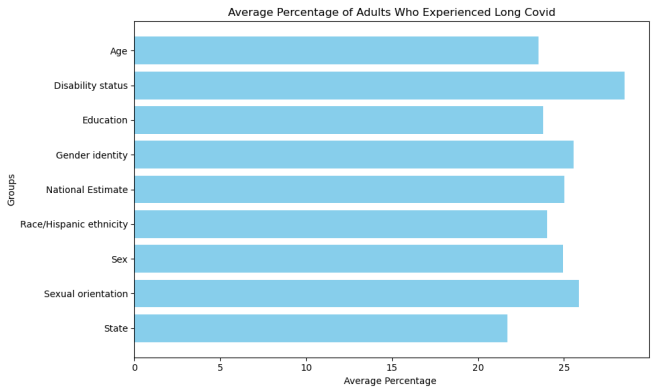

Fig. 1. Estimates from the Household Pulse Survey: Percentage of U.S. Adults 18 and Older Who Experienced Long COVID [4]

II. Related Work

The National Center for Health Statistics (NCHS) [4] has created the Household Pulse Survey in collaboration with the Census Bureau. This survey, initiated in April 2020, aims to provide timely insights into the impact of the pandemic on U.S. adults, including the prevalence of Long COVID symptoms and their effects on daily activities. Recent research has delved into understanding the lingering effects of COVID-19, commonly referred to as Long COVID or post-acute sequelae of COVID-19 (PASC). Studies have identified potential risk factors such as preexisting conditions such as asthma, chronic constipation, reflux, rheumatoid arthritis, seasonal allergies, and depression/anxiety, shedding light on the diverse impacts of the virus [5]. Efforts to define Long COVID subtypes underscore the importance of tailored monitoring and treatment programs [6]. Fuzzy entropy with similarity classification for the feature selection had shown promise in enhancing classification accuracy in medical datasets, particularly in contexts such as Diabetes diagnosis [7]. Relief-based feature selection algorithms and stability selection techniques have also demonstrated effectiveness in capturing complex associations in biomedical data mining [8].

Recent advancements in data cleaning and integration methodologies have contributed significantly to the analysis of electronic health records (EHRs). Deep clustering algorithms such as SDCN, EDESC, and SHGP have been evaluated for their ability to enhance clustering and integration of heterogeneous healthcare data sources [9]. Clustering methodologies have also been proposed to address data quality issues, offering insights into resolving discrepancies and standardizing biomedical metadata [10]. Moreover, comprehensive reviews of data cleaning methods in healthcare underscore the importance of maintaining high data quality for reliable analysis and decision-making [11]. In parallel, clustering approaches for grouping patient records in EHR systems have been evaluated, comparing traditional algorithms with newer techniques tailored for large and complex healthcare datasets [12]. Additionally, machine learning models applied to EHRs have shown promise in early prediction of clinical deterioration, facilitating timely interventions, and improving patient care [13]. This multidisciplinary research landscape, spanning from identifying risk factors for *Long COVID* to refining predictive modeling techniques in healthcare analytics, forms a cohesive tapestry aimed at advancing our understanding of complex medical phenomena and enhancing patient outcomes.

III. Methodology

Long tail data source aggregation refers to the process of collecting and integrating a large number of diverse data sources. In this context, ”long tail” refers to the statistical distribution where a vast number of data sources contribute relatively small amounts of information individually but collectively constitute a significant portion of the overall data. This approach involves aggregating data from a wide range of sources, including niche datasets, specialized databases, or sources with limited availability, to create a comprehensive and inclusive dataset. This can be a complex process in high-dimensional datasets where concepts can have large distinct counts.

A. Data Preprocessing

In our initial research, we developed a pipeline with 23 different feature categories which encompassed fifty-two thousand patients, with the following counts representing the number of OMOP concepts included in each category: Fever (2), Cough (3), Fatigue (2), COVID-19 (2), Renal (15), Obesity (6), Brain Injury (9), Deformity Foot (7), Respiratory Failure (13), Otalgia (6), Ventricular (7), Elevation (6), Bypass (1), Trial Fib (5), Disorders (3), Effusion of Joint (11), Hernia (2), Nutritional Deficiency (2), Pain in Limb (2), Pain in Hand (2), Cyst (1), Loss (1), and Seasonal (1) [1]. Upon further development, we were able to increase our datasets to include 12 million patients. The datasets and their respective counts are as follows: *condition_era* (19,424), *condition_occurrence* (18,004), *drug_era* (16,083), *procedure_occurrence* (14,591), *device_exposure* (2,848), *measurement* (9,318), and *observation* (4,632). These counts represent the number of unique concepts in each category, providing a broad and detailed foundation for our analysis.

1) Target Label Development: We formulated our predictive labels using three distinct OMOP concept sets within the N3C Enclave. For Long COVID, the label comprises two concept codes: ’Post-acute COVID-19’ and ’Post COVID-19 condition, unspecified’. The Sleep Apnea label encompasses nine conditions, including ’Obstructive sleep apnea of adult’, ’Sleep apnea’, ’Obstructive sleep apnea (adult) (pediatric)’, and ’Obstructive sleep apnoea syndrome’, among others. Additionally, for Diabetes Types 1 and 2, our label accounts for both complicated and uncomplicated cases, totaling 536 conditions. Further details on patient counts for each label are provided in Figure 2.

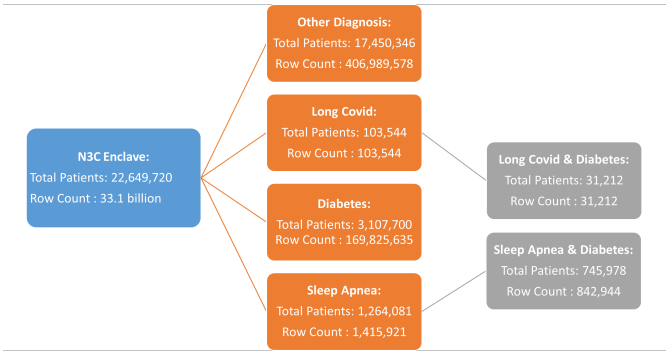

Fig. 2. Patient distribution across target conditions (Diabetes, Sleep Apnea, and Long COVID) with total dataset rows and intersection counts.

2) Addressing Concept Ambiguity: The presence of concept names such as 'Obstructive sleep apnoea syndrome' and 'Obstructive Sleep Apnea syndrome' underscores a significant challenge in medical data analysis. To mitigate confusion encountered by machine learning models, we employed clustering and grouping techniques to address concept ambiguity. By recognizing and consolidating similar conditions, our approach enhances the accuracy and representation of medical data, particularly within the context of our target labels for Sleep Apnea diagnosis. These concept names are integral to the OMOP concept set for Sleep Apnea, and the ambiguity extends to our target labels, which play a crucial role in determining a patient's diagnosed conditions.

B. Feature Space Reduction

However, we weren't able to correctly predict the Long COVID class well, so we decided to focus on reducing the feature space through various techniques: feature selection, topic modeling, clustering, and measuring concept term similarity using Hamming distance to remove redundancy. Hamming similarity serves as preprocessing steps, with the conditions being grouped afterward. LDA (Latent Dirichlet Allocation) and k-means are employed as dimensionality reduction techniques to group similar terms.

C. Baseline Modeling

Random Forest, a popular ensemble learning method, constructs multiple decision trees during training and aggregates their predictions to improve accuracy and robustness. It excels in handling high-dimensional data and is less prone to overfitting compared to individual decision trees. [14]. This section explores several gradient-boosting algorithms employed for various machine-learning tasks. We begin with Gradient Boosting (GB) [15], a foundational ensemble learning that builds a series of weak learners, progressively improving the model by focusing on minimizing the loss function gradient. Its simplicity and robustness to overfitting have made it a popular choice for diverse applications. Building upon this baseline, XGBoost [16] has gained recognition for its exceptional scalability and performance. Through optimizations for speed and efficiency using parallel and distributed computing, XGBoost excels at handling large datasets effectively. LightGBM further augments efficiency in training large-scale datasets [17]. Its innovative gradient-based approach for tree splitting and built-in support for categorical features includes an 'is_unbalanced' hyperparameter that directly contributes to its faster training times and superior performance compared to traditional methods. Finally, CatBoost [18] addresses the challenge of handling categorical features efficiently. CatBoost is particularly well-suited for real-world datasets that often contain categorical features. CatBoost leverages techniques like ordered boosting and oblivious trees, reducing the need for extensive preprocessing.

The final dataset contains seven dataframes, each representing a different category: conditions, observations, devices, drugs, visits, procedures, and measurements. Each patient can have between one and seven feature vectors, with corresponding counts indicating the number of patients possessing each specific number of feature vectors.

D. Proposed Modeling

Although machine learning techniques have shown promise in predictive modeling, challenges persist in optimizing model performance and generalization. To address this, Zhao et al. [19] proposed a novel Stacked Gradient Boost Machine (StackGBM) algorithm. StackGBM combines outputs from XGBoost, LightGBM, and Catboost models in a two-stage paradigm, followed by integration with a neural network system. Comprehensive experiments demonstrated that StackGBM outperformed existing methods, showcasing its potential to advance engineering practices.

The proposed modeling approach involves the use of several ensemble learning methods to predict the presence of Sleep Apnea based on a given dataset. The dataset is preprocessed by downsampling the majority class and combining it with the minority class to address class imbalance. Various classifiers including Gradient Boosting (due to its performance on our Baseline Experiment) , LightGBM, XGBoost, and CatBoost are trained on the data using the hyper-parameters found in the baseline modeling, and their predictions are combined to form a meta-dataset. This meta-dataset is then used to train a RandomForestClassifier, which acts as a meta-learner.

The performance of this stacked model is evaluated using metrics such as accuracy, precision, recall, and F1 score, with results visualized through a confusion matrix. The approach aims to leverage the strengths of different ensemble methods to improve predictive accuracy and robustness.

IV. Proof Of Concept

N3C data The National COVID Cohort Collaborative (N3C) leverages a standardized 'concept_id' system, known as the Observational Medical Outcomes Partnership (OMOP) standard data model [20], to harmonize electronic health record (EHR) data from diverse sources across the United States, facilitating large-scale investigations into COVID-19

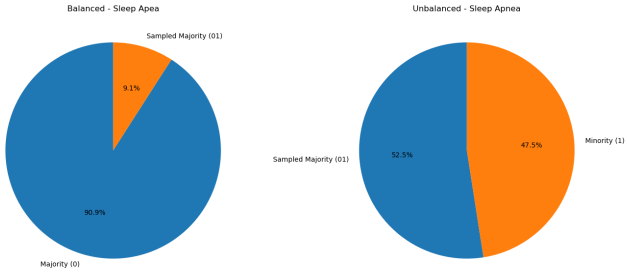

Fig. 3. (a) True Class Counts - Sleep Apnea and (b) Balanced Set Counts - Sleep Apnea

outcomes and treatments. Researchers utilize this system to aggregate and analyze EHR data, extract cohorts based on specific criteria, and transform data into a uniform format that is compatible with OMOP. Statistical and computational methods are then applied to explore patterns and associations, with advanced techniques like machine learning to aid in uncovering insights. Throughout the analysis process, stringent measures are upheld to ensure patient privacy and regulatory compliance. Overall, the N3C initiative provides a robust framework for collaborative research, enabling rapid knowledge generation to address the ongoing challenges posed by the pandemic.

TABLE I
ORIGINAL N3C ENCLAVE DATASETS - DATA CATALOG

Source	Concept ID Count	Concept Name Count	Columns	Rows
device_exposure	6504	5061	19	571,107,525
measurements	26932	26668	30	16,054,603,701
observations	14098	13935	25	3,228,355,577
procedures	57819	57259	19	1,248,124,769
condition_era	53075	53204	8	1,175,596,792
condition_occurrence	53020	53148	21	3,183,174,915
drug_era	37406	37111	9	1,138,041,266
visit_occurrence	57	11	23	1,786,570,960

The Figure 3 illustrates the sampling strategy adopted for each dataframe, stemming from our baseline Kaggle class sampling experiments. By examining the class distributions across various dataframes, we discern that only oversampling techniques were employed in the 'Drugs' dataset. This insight is crucial for understanding how different sampling methods impact the distribution of classes within each dataset. It highlights the effectiveness of oversampling in balancing class distributions, which was necessary to address class imbalance issues and accommodate memory constraints in the enclave. This understanding informs future decisions regarding data preprocessing and model training strategies.

A. Baseline Experiment: Hospital Readmission Prediction Using Kaggle Diabetic Patient Dataset

In this study, we utilize the Kaggle dataset on hospital readmission as a baseline for comparison against larger-scale datasets like the National COVID Cohort Collaborative (N3C), focusing on the aggregation of extended tail data sources. The Kaggle dataset serves as a reference point

TABLE II
PATIENT COUNTS BY DIAGNOSIS AND COVERAGE PERCENTAGE (COUNT/TOTAL COUNT)

Sleep Apnea	Long COVID	Diabetes	Other	Patient Count	Row Count	Overall Coverage
1	0	0	0	1,264,081	1,415,921	6.15%
0	1	0	0	103,544	103,544	0.50%
0	0	1	0	3,107,700	169,825,635	15.11%
0	0	0	1	17,450,346	406,989,578	84.87%
1	0	1	0	745,978	842,944	3.63%
0	1	1	0	31,212	31,212	0.15%

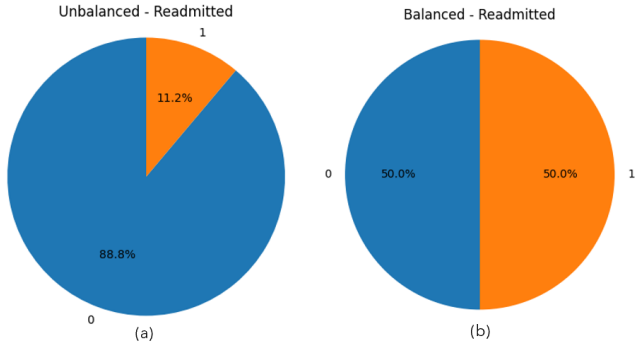

Fig. 4. (a) True Class Counts - Hospital Readmission and (b) Balanced Set Counts - Hospital Readmission [Majority: 86950; Minority: 10966]

for evaluating the performance and scalability of predictive models for transmission prediction tasks. In contrast, the N3C dataset offers a more extensive and diverse collection of Electronic Health Records (EHRs) from various healthcare institutions across the United States. By comparing these datasets, we aim to assess the efficacy and generalizability of predictive models in handling extended tail data sources characterized by uneven attribute distributions, rare occurrences, and complexities inherent in real-world healthcare data.

Our methodology employs the Random Forest algorithm with 50 trees, which is informed by our analysis of the Kaggle dataset. We observed minimal variation in feature selection among gradient boosting methods, with Random Forest tending to select a more significant number of features compared to the LassoCV method. To optimize feature selection, we establish a threshold at the 50th percentile of attribute importance (with feature importance scores ¿= 0.1 in N3C). Attributes surpassing this threshold are considered pertinent and incorporated into the final feature set, enhancing our predictive model's resilience. Leveraging this feature's vital information, we refine our feature selection process to include critical attributes that contribute significantly to our model's performance. This dual approach maximizes our model's predictive power while mitigating the risk of overfitting and enhancing its interpretability.

A similarity between the Kaggle Diabetic Hospital Readmission and N3C target conditions datasets is the noticeable imbalance in class distribution, as illustrated in Figure 4 and 3 respectively.

Our methodology utilizes the embedded Random Forest selection algorithm with 50 trees, which were informed by our analysis of the Kaggle dataset. Notably, our examination

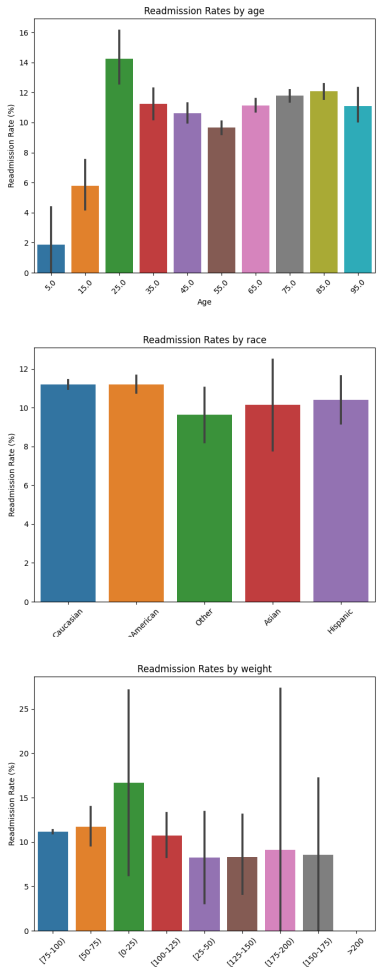

Fig. 5. Kaggle Demographic Characteristics

revealed minimal variation in the feature selected among gradient boosting methods. At the same time, Random Forest tended to choose a more significant number of features than the LassoCV method.

Initially, we harness the Random Forest algorithm, which inherently offers feature importance metrics such as the Gini importance or mean decrease impurity. To optimize the feature selection procedure, we suggest establishing a threshold at the 50^{th} percentile of attribute importance (with feature importance scores ≥ 0.1 in N3C). Attributes exceeding this threshold are deemed pertinent and are consequently incorporated into the final feature set. This method guarantees the integration of attributes identified as influential by the RF algorithm, thereby fortifying the resilience of our predictive model. By leveraging this feature importance information, we further refine our feature selection process, ensuring the inclusion of critical attributes that contribute significantly to our model's predictive performance. This dual approach allows us to harness the strengths of both algorithms, maximizing our model's predictive power while mitigating the risk of overfitting and enhancing its interpretability.

TABLE III
KAGGLE TOP 11 FEATURES BY FEATURE SELECTION METHODS: LASSOCV, RANDOMFORESTCLASSIFIER AND GRADIENT BOOSTING MODELS (LIGHTGBM, XGBOOST, CATBOOST)

LassoCV	**RandomForestClassifier**	**Gradient Boosting Models**
number_emergency	age	number_inpatient
number_inpatient	time_in_hospital	discharge_Discharged to Home
diag_1_250.41	num_lab_procedures	number_diagnoses
diag_1_250.42	num_procedures	time_in_hospital
diag_1_250.6	num_medications	age
diag_1_250.7	number_outpatient	num_lab_procedures
diag_1_434	number_emergency	number_emergency
diag_1_443	number_inpatient	num_medications
diag_1_787	number_diagnoses	diag_1_V58
diag_1_V58	race_AfricanAmerican	diag_1_434

The feature selection process of the Kaggle dataset led to the inclusion of the visits data frame in our N3C analysis. Table III presents the top 11 features selected using LassoCV, RandomForestClassifier, and Boosting Models (LightGBM, XGBoost, Catboost). These features were deemed significant for predictive modeling, highlighting attributes such as number of emergency visits, age, number of inpatient visits, time spent in hospital, discharge destination, number of diagnoses, number of procedures, number of medications, and specific diagnosis codes. The relevance of the visits data frame captures patient interaction with healthcare services and its importance in enhancing the predictive capabilities of our analysis.

TABLE IV
KAGGLE DATASET SAMPLING EXPERIMENTS BY MODEL: RANDOM FOREST, GRADIENT BOOSTING, LIGHTGBM

Sampling Technique	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Smote-RFC	0.8826	0.8216	0.8826	0.8395
Undersampled-RFC	0.6140	0.8384	0.6140	0.6856
Resampled-RFC	0.8818	0.8192	0.8818	0.8388
Smote-GBT	0.8498	0.8127	0.8498	0.8292
Undersampled-GBT	0.6383	0.8399	0.6383	0.7050
Resampled-GBT	0.8498	0.8127	0.8498	0.8292
Undersampled-LGBM	0.8498	0.8127	0.8498	0.8292
Smote-LGBM	0.8887	0.8473	0.8887	0.8412
Resampled-LGBM	0.8887	0.8473	0.8887	0.8412

Sampling experiments were conducted using various techniques, such as random forest, gradient boosting, and LightGBM algorithms. Table IV presents the results, with the SMOTE technique consistently outperforming undersampling and simple resampling methods across all algorithms. SMOTE yielded higher accuracy, precision, recall, and F1 scores, indicating its effectiveness in addressing class imbalance and improving predictive model performance. Conversely, undersampling techniques generally resulted in lower performance, suggesting potential information loss due to reduced samples from the majority class.

B. N3C Random Forest Feature Importance Scores

From our feature selection and sampling method results using the Kaggle Dataset, I selected random forest and gradient boosting models to show the performance of the three feature space reduction techniques. The feature importance scores for various medical concepts were analyzed across different targets, specifically Diabetes, Long COVID, and Sleep Apnea.

TABLE V
OVERALL COVERAGE OF FEATURE SPACE REDUCTION TECHNIQUES

Feature Technique	Missing Count	Included	Overall Coverage
Hamming Similarity	0	20,561,373	100%
k-means Clustering	9,098,998	11,462,375	56%
LDA	194,562	20,366,811	99%

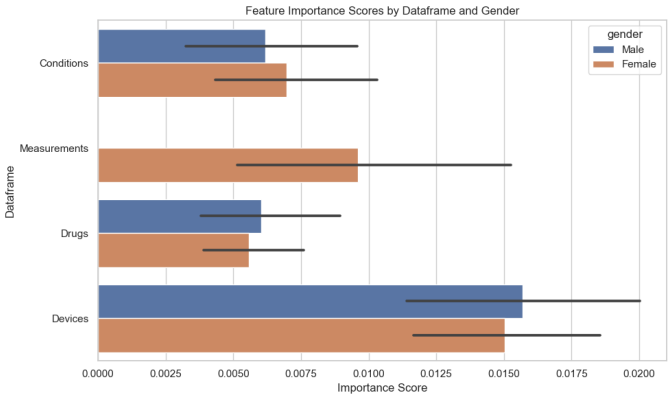

Fig. 6. Feature Importance Scores of Random Forest by data frame and Gender across Target Labels: Long COVID, Sleep Apnea, and Diabetes (Uncomplicated & Complicated)

Fig. 7. Feature Importance Scores of Random Forest by data frame and Target Labels: Long COVID, Sleep Apnea, and Diabetes (Uncomplicated & Complicated)

TABLE VI
TOP 10 SLEEP APNEA FEATURE IMPORTANCE FOR MALES

Dataframe	Concept Name	Importance Score
Devices	Technetium tc-99m macro aggregated albumin, diagnostic, per study dose, up to 10 millicuries	0.0458
Devices	Cervical, collar, semi-rigid, thermoplastic foam, two-piece with thoracic extension, prefabricated, off-the-shelf	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

C. Understanding the Connections Between the Targets

For Diabetes and Long COVID, the concept "Alginate or other fiber gelling dressing, wound cover, sterile, pad size 16 sq. in. or less, each dressing" showed an importance score of 0.0051 for Diabetes and 0.0230 for Long COVID in males. The "Closure device, vascular (implantable/insertable)" had importance scores of 0.0031 and 0.0166 for Diabetes and Long COVID in males, respectively, and 0.0112 and 0.0166 in

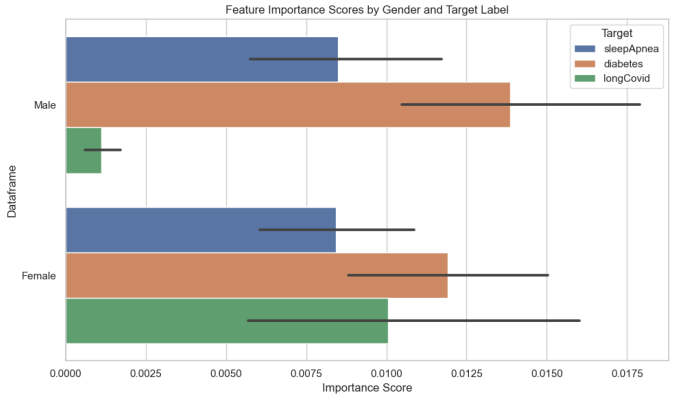

Fig. 8. Feature Importance Scores of Random Forest by Gender across Target Labels: Long COVID, Sleep Apnea, and Diabetes (Uncomplicated & Complicated)

TABLE VII
TOP 10 SLEEP APNEA FEATURE IMPORTANCE FOR FEMALES

Dataframe	Concept Name	Importance Score
Devices	Technetium tc-99m macro aggregated albumin, diagnostic, per study dose, up to 10 millicuries	0.0458
Devices	Cervical, collar, semi-rigid, thermoplastic foam, two-piece with thoracic extension, prefabricated, off-the-shelf	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

TABLE VIII
TOP 10 DIABETES FEATURE IMPORTANCE FOR MALES

Dataframe	Concept Name	Importance Score
Devices	Basic nasal oxygen cannula	0.0368
Devices	Non-covered item or service	0.0326
Devices	N3C:Room air	0.0117
Devices	Red blood cells, leukocytes reduced, each unit	0.0200
Devices	Ventilator	0.0103
Devices	Oxygen ventilator	0.0300
Devices	Guidewire	0.0034
Devices	Treatment devices, design, and construction; complex (irregular blocks, special shields, compensators, wedges, molds or casts)	0.0037
Devices	High flow oxygen nasal cannula	0.0954
Devices	Catheter, infusion, inserted peripherally, centrally, or midline (other than hemodialysis)	0.0610

TABLE IX
TOP 10 DIABETES FEATURE IMPORTANCE FOR FEMALES

Dataframe	Concept Name	Importance Score
Devices	Red blood cells, leukocytes reduced, each unit	0.0793
Devices	Catheter, transluminal angioplasty, non-laser (may include guidance, infusion/perfusion capability)	0.0671
Measurements	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, centrally, or midline (other than hemodialysis)	0.0413
Devices	Treatment devices, design, and construction; complex (irregular blocks, special shields, compensators, wedges, molds or casts)	0.0388
Devices	Introducer/sheath, guiding, intracardiac electrophysiological, fixed-curve, other than peel-away	0.0383
Devices	N3C:Room air	0.0378

TABLE X
TOP 10 LONG COVID FEATURE IMPORTANCE FOR FEMALES

Dataframe	Concept Name	Importance Score
Measurements	Culture, fungi; yeast	0.0013
Measurements	Urobilinogen in Urine by Automated test strip	0.0009
Measurements	Monocytes+Macrophages/100 leukocytes in Body	0.0012
Measurements	severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (coronavirus disease [COVID-19])	0.0039
Measurements	Special stain including interpretation and report	0.0025
Measurements	Varicella zoster virus DNA [Presence] in Specimen	0.0089
Measurements	Complement C3 [Mass/volume] in Serum or Plasma	0.0075
Measurements	Yersinia sp DNA [Identifier] in Specimen by NAA	0.0041
Measurements	Aldolase [Enzymatic activity/volume] in Serum or Plasma	0.0394
Measurements	Human metapneumovirus RNA [Presence] in the Lower respiratory	0.0145

TABLE XI
TOP 10 LONG COVID FEATURE IMPORTANCE FOR MALES

Dataframe	Concept Name	Importance Score
Devices	Tracheostomy speaking valve	0.0042
Drugs	Fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray [Flonase]	0.0032
Drugs	Salicylic acid 20 MG/ML / sulfur 20 MG/ML Medicated Shampoo	0.0027
Conditions	Arteritis	0.0022
Devices	FUL-GLO 1 MG OPTH STRIPS	0.0022
Conditions	Post-acute COVID-19	0.0012
Drugs	Acetazolamide 125 MG Oral Tablet	0.0010
Devices	Upper extremity fracture orthosis, humeral, prefabricated, includes fitting and adjustment	0.0009
Devices	Leukocyte reduced red blood cells, human	0.0006
Conditions	Bilateral localized swelling of lower limbs	0.0004

females. Tables VIII and IX for the top 10 Diabetes features; Tables XI and X for the top 10 Long COVID features.

In the comparison between Diabetes and Sleep Apnea, the concept "CVS TEST STRIP" had importance scores of 0.0011 for Sleep Apnea and 0.0616 for Diabetes in females, while in males, the scores were 0.0051 and 0.0616, respectively. The "Introducer/sheath, guiding, intracardiac electrophysiological, fixed-curve, other than peel-away" had scores of 0.0022 for Sleep Apnea and 0.0383 for Diabetes in females. The "N3C: Room air" concept showed scores of 0.0113 for Sleep Apnea, 0.0247 for Diabetes in females, and 0.0151 and 0.0247 in males. Please refer to the Tables VIII and IX for the top 10 Diabetes features; Tables VI and VII for the top 10 Sleep Apnea features.

Other 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," displayed varying importance scores for the targets, with notable differences across genders. These results highlight the differing significance of medical concepts in relation to specific conditions, emphasizing the need for gender-specific analysis in medical research.

The comparative analysis of specific health conditions, namely Diabetes, Sleep Apnea, and Long COVID, across genders, can be found in Figures 10 and 9. The X-axis represents the concept names found in both conditions, while the Y-axis denotes gender categories (Female and Male) separated by condition. The color intensity in each cell indicates the magnitude of the scores, with higher intensity representing higher scores. This visualization allows for the identification of disparities or similarities in condition scores between genders.

Shared Medical Concepts: Certain medical concepts are essential across multiple conditions, suggesting common underlying factors or overlapping pathophysiological mechanisms. For instance, the idea "Alginate or other fiber gelling dressing, wound cover" has higher importance for Long COVID compared to Diabetes in males, indicating its particular relevance in wound management for Long COVID patients. Tables VIII and IX for the top 10 Diabetes features; Tables XI and X for the top 10 Long COVID Features.

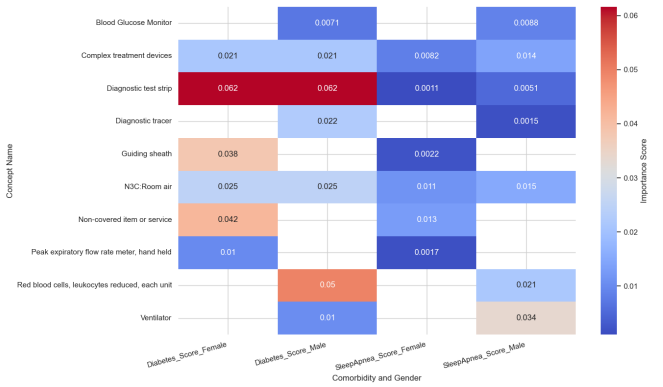

Fig. 9. Diabetes and Sleep Apnea - Feature Importance Scores by Gender

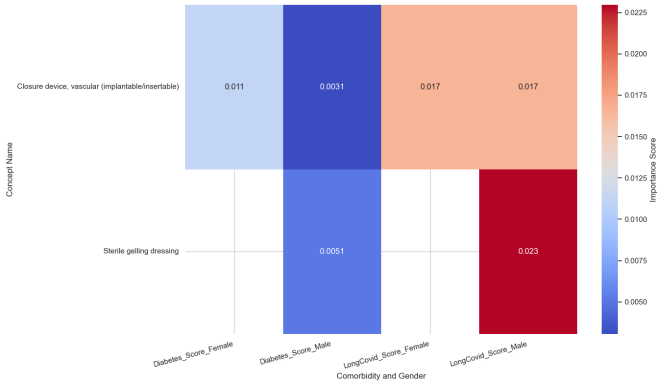

Fig. 10. Diabetes and Long Covid - Feature Importance Scores by Gender

Gender-Specific Differences: The importance scores vary 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 importance scores for Diabetes and Long COVID across genders, seen in Figure 10 which suggest that vascular complications might present differently or have other clinical implications in males versus females.

Condition-Specific Significance: Concepts like "CVS TEST STRIP" have drastically higher importance for Diabetes compared to Sleep Apnea, both in males and females. Since blood glucose monitoring is directly related to Diabetes management, this is expected. The contrast in importance scores shows that although these conditions have similarities, condition-specific significance is found in each individual feature.

Potential Overlaps and Comorbidities: Some concepts show moderate importance across conditions, indicating potential overlaps or comorbidities. For example, "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. Tables VIII and IX for the top 10 Diabetes features; Tables VI and VII for the top 10 Sleep Apnea Features.

V. Modeling Results

TABLE XII
KAGGLE DIABETIC PATIENT HOSPITAL READMISSION - BASELINE EXPERIMENT: MODELING SCORES

Sampling Technique	Accuracy	Precision	Recall	F1 Score
RandomForestClassifier	0.6059	0.6053	0.5866	0.5958
GradientBoostingClassifier	0.6181	0.6176	0.6004	0.6089
XGBClassifier	0.6044	0.6019	0.5939	0.5979
LGBMClassifier	0.6160	0.6167	0.5927	0.6045
CatBoostClassifier	0.6131	0.6127	0.5939	0.6032
Tensor(4-Hidden)(0.3)	0.5547	0.5309	0.8653	0.6580
Tensor(4-Hidden)(0.5)	0.5967	0.6266	0.4589	0.5298
Tensor(4-Hidden)(0.7)	0.5564	0.7432	0.1590	0.2619
Tensor(4-Hidden)(0.9)	0.5135	0.8519	0.0212	0.0413

In Figure 11, we aim to compare the performance metrics of different models across seven data frames (Conditions, Observations, Devices, Procedures, Visits, Measurements, Drugs) and three techniques (LDA, Hamming Similarity, and KMeans Clustering). We began by transforming the data frame, which contains columns such as 'Model', 'Technique', 'Dataframe', 'Accuracy', 'Precision_0', 'Recall_0', 'F1-Score_0', 'Precision_1', 'Recall_1', and 'F1-Score_1', into a long format using the melt function from pandas. This reshaping process involved melting the data frame with 'Model', 'Technique', and 'Dataframe' as identifier variables and the performance metrics ('Accuracy', 'Precision_0', 'Recall_0', 'F1-Score_0', 'Precision_1', 'Recall_1', 'F1-Score_1') as value variables.

TABLE XIII
HYPERPARAMETERS OF THE GRADIENT BOOSTING MODELS USED FOR STACKED MODEL

Model	Hyperparameter	Value
GradientBoostingClassifier	learning_rate	0.2
	max_depth	5
	n_estimators	100
LGBMClassifier	learning_rate	0.2
	n_estimators	50
	num_leaves	50
XGBClassifier	learning_rate	0.2
	max_depth	200
	min_child_weight	1
CatBoostClassifier	learning_rate	0.05
	max_depth	9
	iterations	200
	l2_leaf_reg	1

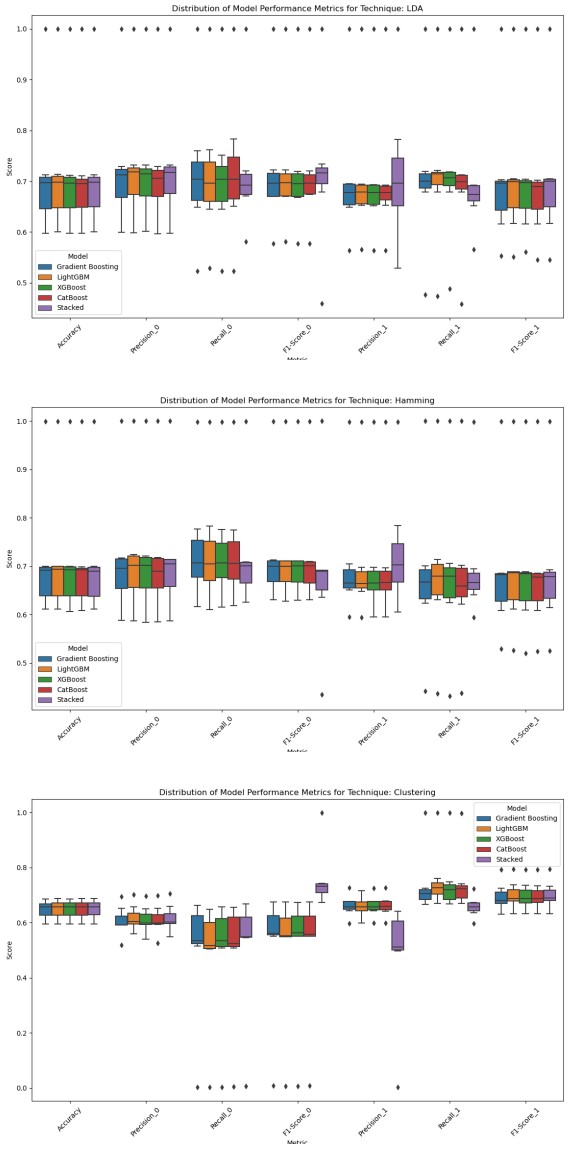

Fig. 11. Distribution of Model Performance Metrics by Technique ((top) LDA, (middle) Hamming Similarity (bottom) KMeans Clustering) for Sleep Apnea Target Label

VI. Conclusion and Future Work

A. Conclusion

The analysis reveals critical insights into predictive factors for Sleep Apnea, Diabetes, and Long COVID across genders. For Sleep Apnea, consistent reliance on specific diagnostic and treatment tools is evident for both genders. However, Diabetes showcases gender-specific variations, with males prioritizing respiratory and infusion devices while females emphasize blood-related measurements and diagnostic tools. *Long COVID* demonstrates even more pronounced gender disparities, with females emphasizing diagnostic tests for infections and metabolic markers, while males' data reflect a broader mix of devices, conditions, and drugs. These findings

underscore the necessity for gender-specific medical analyses to advocate for personalized treatment plans and optimize healthcare resources.

Furthermore, the stacked model, integrating multiple gradient-boosting models (Gradient Boosting, LightGBM, XGboost, CatBoost), demonstrates robust performance, particularly in accurately predicting co-morbid conditions across genders. This highlights the effectiveness of ensemble learning techniques in enhancing predictive analytics in healthcare. Additionally, the analysis of gradient-boosting model performance across various datasets underscores the need for dataset-specific optimization strategies to maximize model accuracy and generalizability.

B. Future Work

Future research will focus on several key areas to build upon these findings:

- **Enhanced Feature Selection Methods**: Implementing advanced feature selection techniques to refine the predictive models further and identify more granular gender-specific differences.
- **Expanded Datasets**: Incorporating more extensive and more diverse open-source datasets to validate and generalize the findings across different populations and healthcare settings.
- **Temporal Analysis**: Analyzing temporal patterns in the feature is essential to understanding how the predictive factors evolve and their impact on disease progression and treatment efficacy.
- **Integrative Approaches**: Combining feature space reduction techniques with other machine learning methodologies, such as deep learning, to improve the accuracy and robustness of the predictive models.

By addressing these areas, future work can enhance the precision and applicability of predictive models in healthcare, ultimately contributing to more effective and personalized patient care.

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 and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the ongoing development of this community resource [https://doi.org/10.1093/jamia/ocaa196]."

References

[1] Mirna Elizondo, Rasim Musal, June Yu, and Jelena Tesic. Long covid challenge: Predictive modeling of noisy clinical tabular data. *The 11th IEEE International Conference on Healthcare Informatics*, 2023.

[2] National Institute of Health. National covid cohort collaborative (n3c), 2022. Accessed: 2023-01-12.

[3] Saurabh Tayal. Diabetic patients' re-admission prediction, Aug 2020.

[4] National Center for Health Statistics. Common Core. https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm, June 2022. Public access level: Data asset is publicly available to all without restrictions (public).

[5] Elizabeth T. Jacobs, Collin J. Catalfamo, Paulina M. Colombo, Sana M. Khan, Erika Austhof, Felina Cordova-Marks, Kacey C. Ernst, Leslie V. Farland, and Kristen Pogreba-Brown. Pre-existing conditions associated with post-acute sequelae of covid-19. *Journal of Autoimmunity*, 135:102991, 2023.

[6] Skyler Resendez, Steven H. Brown, H. Sebastian Ruiz, Prahalad Rangan, Jonathan R. Nebeker, Diane Montella, and Peter L. Elkin. Defining the subtypes of long covid and risk factors for prolonged disease. *medRxiv*, 2023.

[7] Pasi Luukka. Feature selection using fuzzy entropy measures with similarity classifier. *Expert Systems with Applications*, 38(4):4600–4607, 2011.

[8] Ryan J. Urbanowicz, Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85:189–203, 2018.

[9] Hafiz Tayyab Rauf, Andre Freitas, and Norman W. Paton. Deep clustering for data cleaning and integration, 2023.

[10] Zhipeng Cai, Daniel Zeng, Naixue Zhang, Thanh Nguyen, and Xin Gao. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *BMC Bioinformatics*, 18(Suppl 14):61, 2017.

[11] Behrouz Ehsani-Moghaddam, Ken Martin, and John A Queenan. Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. *Health Information Management Journal*, 50(1-2), 2021. https://orcid.org/0000-0002-5038-1118 (Behrouz Ehsani-Moghaddam).

[12] Shira 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).

[13] Victor M. Ruiz, Michael P. Goldsmith, Lingyun Shi, Allan F. Simpao, Jorge A. Gálvez, Maryam Y. Naim, Vinay Nadkarni, J. William Gaynor, and Fuchiang Rich Tsui. Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records. *The Journal of Thoracic and Cardiovascular Surgery*, 163(3):883–892.e11, Dec 2021.

[14] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[15] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 2001.

[16] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[17] Thomas Finley et al. 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*, pages 3149–3157, 2017.

[18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: Unbiased boosting with categorical features, Jan 2019.

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

[20] OHDSI CDM Working Group. Omop cdm v5.3, Aug 2021.