Diagnosis and Conditions: can relations be uncovered from the Big Data?

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

***Abstract*—This study investigates the role of linked comorbidities in female patients to deepen our understanding of Long COVID-19. Specifically focusing on diabetic females as a high-risk subgroup, we explore how comorbid conditions influence the development and persistence of Long COVID-19 symptoms. Long COVID-19 presents a spectrum of symptoms across various bodily systems, including fatigue, fever, respiratory and heart symptoms like difficulty breathing and chest discomfort, neurological symptoms such as headaches and disrupted sleep, as well as digestive issues like diarrhea. [1] Leveraging the Kaggle Hospital Re-admission dataset [2], we analyze the interconnections between comorbidities and the likelihood of long-term COVID-19 complications in female diabetic individuals. Additionally, utilizing the extensive National COVID Cohort Collaborative (N3C) dataset [3], we discern patterns and associations between different conditions and Long COVID-19 outcomes specifically in female patients. Employing advanced analytical methods, including machine learning algorithms, we aim to develop predictive models to accurately identify female patients at risk of Long COVID-19 based on their comorbidity profiles. The insights from this research refine risk assessment, guide clinical decision-making, and tailor targeted interventions for Long COVID-19 among female patients with linked comorbidities. Future research will explore additional comorbidity relationships unique to female patients and integrate diverse data sources to advance personalized healthcare strategies in the context of Long COVID-19.**

***Index Terms*—gradient boosting, predictive modeling, noisy data, long covid, diabetes**

I. Introduction

The COVID-19 pandemic has swiftly emerged as an unparalleled global health crisis, profoundly impacting every facet of human existence, ranging from individual well-being to economic stability, social dynamics, and healthcare systems worldwide. Since its initial outbreak in late 2019, the novel coronavirus, SARS-CoV-2, has spread rapidly across the globe, affecting millions of individuals and leading to substantial mortality rates. The pandemic has gone beyond immediate health consequences and has introduced a myriad of challenges and disruptions to societies on a global scale. In addition to the direct impact on mortality and morbidity, the aftermath of the pandemic has revealed the emergence of a condition known as Long COVID-19, further exacerbating the challenges faced by individuals and healthcare systems. Long COVID, Post-COVID Conditions (PCC), or post-acute sequelae of SARS-CoV-2 infection (PASC), denote a variety of lingering symptoms and complications that persist in individuals following their recovery from the initial phase of COVID-19. [1]

Various risk factors associated with COVID-19 have been identified, including the presence of severe acute symptoms, advancing age, female sex, and preexisting comorbidities. Preexisting conditions that have been identified encompass diabetes, lung disease, frailty, chronic obstructive pulmonary disease, the use of medications for autoimmune disorders, asthma, multiple sclerosis, and depression/anxiety [4].

The primary aim of this research is to discern relationships within extensive electronic health records, providing evidence-based guidance that incorporates long-term ramifications. Through enhancing predictive model performance, the study endeavors to equip healthcare providers and policymakers with the requisite tools for informed decision-making, encompassing optimal resource allocation, effective intervention strategies, and tailored support for high-risk individuals. Secondary objectives entail pinpointing specific risk factors linked to long-term COVID-19 in female diabetic patients and formulating targeted intervention approaches. Furthermore, this study seeks to address challenges in diabetic patient care, particularly those posed by Long COVID, by bridging the gap between data analysis and actionable insights. Electronic Health Records (EHR) present a wealth of information, yet their complexity, including high dimensionality and sparsity, poses significant challenges in extracting meaningful insights, even with a large number of subjects available.

This study utilizes the National COVID Cohort Collaborative (N3C) system, which consists of a vast collection of electronic health records (EHRs) from multiple healthcare institutions. The N3C system grants us access to patient information contributed by 76 healthcare centers spanning 49 out of 50 states in the United States. Additionally, we utilize data from the Kaggle Hospital Re-admission dataset, specifically focusing on diabetic patients. This dataset provides valuable insights into the hospital readmission patterns of diabetic individuals, complementing our analysis within the N3C dataset. The combined dataset represents 19 million patients, among which there are 8.4 million cases of confirmed COVID-positive individuals. This comprehensive dataset consists of over 25 billion rows of valuable data, providing a rich and extensive resource for analysis [5].

Our initial data cleaning, data integration, and data analysis reveal characteristics typical of tabular data from multiple heterogeneous sources. Tabular data in the wild consists of an uneven distribution of attributes, missing, overlapping, noisy values, and a mix of categorical and numerical data

attributes. To address the challenges posed by the unique entry values and complexities of the N3C, this study aggregates and organizes the variables relevant to diabetic re-hospitalization.

II. Related Work

The National Center for Health Statistics (NCHS) 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.

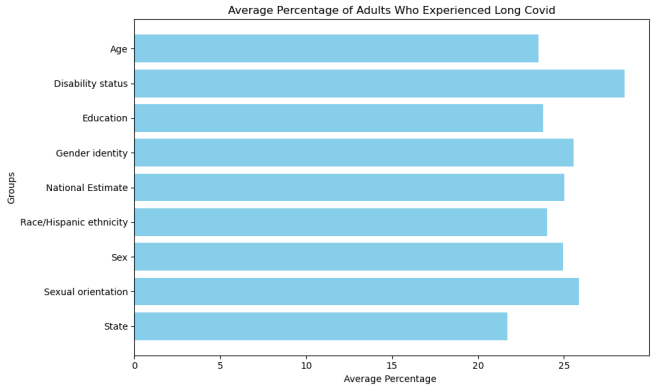

Fig. 1. Estimates from the Household Pulse Survey: Percentage of U.S. Adults 18 and Older Who Experience Long Covid

A. COVID-19

Recent studies have investigated the association between preexisting conditions and the risk of experiencing post-acute sequelae of COVID-19 (PASC), highlighting conditions such as asthma, chronic constipation, reflux, rheumatoid arthritis, seasonal allergies, and depression/anxiety as potential risk factors [4]. Additionally, efforts have been made to define long COVID and identify medical-specialty subtypes, emphasizing the importance of monitoring and treatment programs for at-risk individuals [6].

B. Maternal and Perinatal Outcomes

A systematic review and meta-analysis explored the impact of the COVID-19 pandemic on maternal, fetal, and neonatal outcomes, revealing increases in stillbirth and maternal death rates, particularly in high-income countries. However, overall, preterm birth rates did not show significant changes. Notably, high-income countries experienced reductions in preterm births, along with increased rates of surgically managed ectopic pregnancies [7].

C. Diabetes

Feature selection using fuzzy entropy with a similarity classifier has shown promise in improving classification accuracy in medical datasets, including those related to diabetes. The method enhances model transparency and simplifies diagnoses, making it valuable in medical contexts. Relief-based feature selection algorithms and stability selection techniques have also been highlighted for their effectiveness in capturing complex associations in biomedical data mining [8], [9].

D. Modeling

Comparative studies have evaluated the effectiveness of gradient boosting algorithms, including GBM, XGBoost, LightGBM, and CatBoost, for classification tasks, with a focus on hyperparameter optimization techniques to enhance performance. These models have shown promise in predicting diabetes and mortality in heart failure patients, as well as classifying individuals at risk of COVID-19 infection [10]–[12]. For example, in a study comparing the ability of two machine learning models to predict myocardial infarction (MI), an XGBoost model trained on EHR data outperformed a deep neural network (DNN) model trained on ECG traces [13]. Similarly, a stacking ensemble learning model incorporating multiple machine learning algorithms has been proposed to predict mortality in heart failure patients during ICU admission [14]. Moreover, ensemble learning models, such as optimized XGBoost, have been effective in classifying individuals at risk of COVID-19 infection, highlighting the potential of machine learning in public health initiatives [12].

III. N3C Data Characteristics

The National COVID Cohort Collaborative (N3C) leverages a standardized 'concept_id' system, such as the Observational Medical Outcomes Partnership (OMOP) common data model, to harmonize electronic health record (EHR) data from diverse sources across the United States, facilitating large-scale investigations into COVID-19 outcomes and treatments. Researchers utilize this system to aggregate and analyze EHR data, extracting cohorts based on specific criteria and transforming data into a uniform format compatible with OMOP. Statistical and computational methods are then applied to explore patterns and associations, with advanced techniques like machine learning aiding 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 datasets

Datasets	Rows x Columns	# Patients	# Concepts	DataFrame
condition_era	1,204,337,597 x 8	19,876,988	65,581	conditions
condition_occurrence	3,306,921,179 x 21	20,140,630	59,985	
visit_occurrence	1,689,971,115 x 23	21,781,439	60	visits
drug_era	1,095,058,492 x 9	18,638,834	32,528	drugs
death	687,970 x 11	689,305	13	deceased
person	22,062,107 x 27	22,062,107	48	demographics
measurement	14,757,204,903 x 30	21,562,580	27,501	measurement
observation	2,881,068,441 x 25	21,262,196	12,394	observations
procedure_occurrence	1,162,065,398 x 19	17,467,610	59,513	procedures
device_exposure	544,709,418 x 19	6,472,493	5,798	devices

Analyzing the Tables I and II reveals several key insights. Firstly, the disparity in patient population size between the

TABLE II
FEMALE COHORT N3C DATASETS

Datasets	Rows x Columns	# Patients	# Concepts	data frame
condition-era	1,057,039,277 x 8	18,282,351	63,630	conditions
condition_occurrence	2,757,267,130 x 21	18,539,963		
visit_occurrence	1,490,120,555 x 23	19,761,967	60	visits
drug_era	968,271,120 x 9	17,180,741	32,180	drugs
death	402,954 x 11	382,420	12	deceased
person	12,290,501 x 27	12,290,501	48	demographics
measurement	15,305,292,900 x 30	21,899,515	27,487	measurement
observation	2,616,365,368 x 25	19,381,019	12,499	observations
procedure_occurrence	947,996,351 x 19	15,673,100	57,970	procedures
device_exposure	493,232,614 x 19	5,666,679	5,693	devices

complete N3C datasets and the female cohort datasets is noteworthy. The former comprises 22,062,107 patients, while the latter contains 12,290,501 patients. This substantial difference underscores the importance of deliberate cohort selection to ensure the representation of pertinent demographic groups, thereby mitigating potential biases and enhancing the generalizability of our findings.

Secondly, both datasets encompass similar healthcare data concepts, including condition occurrences, visit occurrences, drug eras, and measurements. However, subtle variations in the number of concepts captured within each dataset are observed. Specifically, the complete N3C datasets exhibit a slightly higher average number of concepts compared to the female cohort datasets. This discrepancy underscores the nuanced differences in healthcare utilization and medical conditions among diverse patient populations. Consequently, the selection of the female cohort allows for a more targeted analysis of gender-specific healthcare needs and outcomes, thereby enriching the depth and breadth of our research insights.

Moreover, a closer examination of the death dataset within the original N3C datasets reveals a more significant number of patients and concepts compared to its counterpart in the female cohort datasets. His divergence may reflect inherent disparities in mortality rates and causes of death between genders, underscoring the imperative to consider gender-specific factors in healthcare research and policy formulation. By leveraging the insights gleaned from these datasets, our academic research endeavors can effectively address gender-specific healthcare needs and contribute to advancing equitable and inclusive healthcare practices.

IV. LONG TAIL DATA SOURCE AGGREGATION

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.

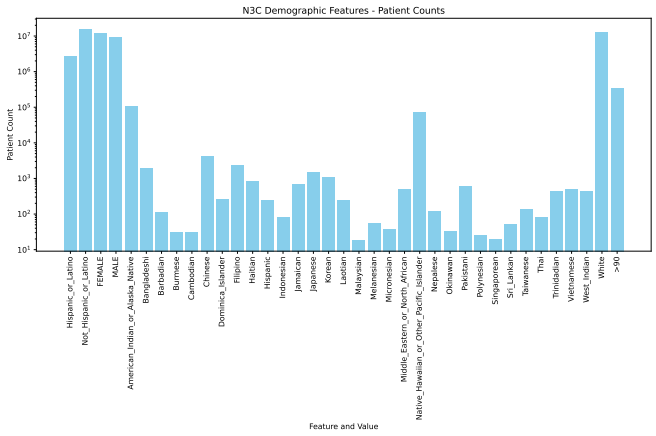

Fig. 2. Distribution over aggregated gender, race, ethnicity, and age group categories in the Demographics data frame

A. Kaggle Hospital Readmission

In this study, we employ the Kaggle dataset on hospital readmission as a baseline for comparison against a larger-scale dataset, such as the National COVID Cohort Collaborative (N3C), with a focus on long tail data source aggregation. y utilizing the Kaggle dataset, which serves as a reference point for evaluating the performance and scalability of predictive models applied to hospital readmission prediction tasks. In contrast, the N3C dataset offers a substantially more extensive and diverse collection of Electronic Health Records (EHRs) from multiple healthcare institutions across the United States. The comparison between these datasets allows us to assess the efficacy and generalizability of predictive models in handling extended tail data sources, characterized by uneven distribution of attributes, rare occurrences, and complexities inherent in real-world healthcare data.

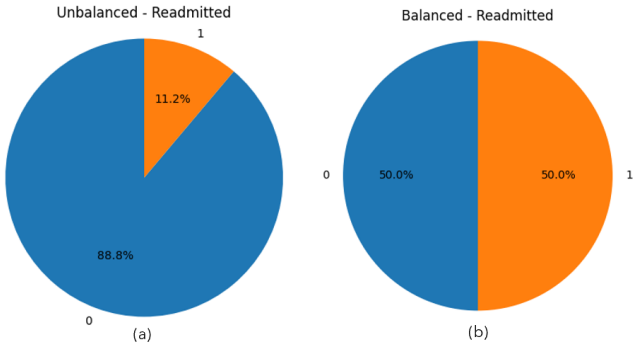

Fig. 3. (a) True Class Counts - Hospital Readmission and (b) Balanced Set Counts - Hospital Readmission [Majority: 86950; Minority: 10966]

The N3C system grants us access to patient records, which represent 19 million patients, among whom there are 8.4 million cases of confirmed COVID-positive individuals. In this study, we will be focusing on the identified female cohort; see Table II.

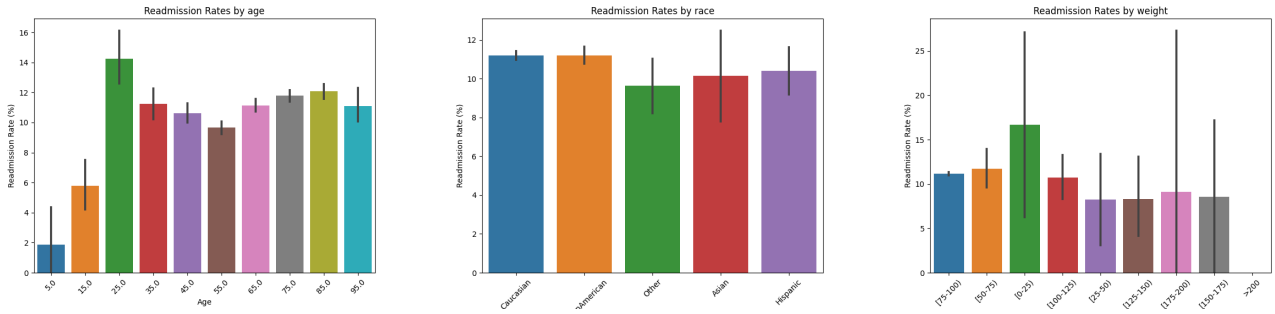

Fig. 4. Kaggle Demographic Characteristics

B. Data Pre-Processing

1) Enclave Pipeline: The **Demographics** data frame consolidates information from the **person** and **death** datasets, as outlined in Table II. It encompasses 48 attributes, including *person_id*, age, 41 race and ethnicity indicators, one deceased label and three target label attributes: one for long COVID, one for diabetes type 1, and one for diabetes type 2.

Patient **Conditions** and their occurrence and duration records were sourced from *condition_occurrence* and *condition_era* datasets (refer to Table II), resulting in the aggregation of 18,282,351 patient records. Each patient had at least one condition out of 63,630 unique conditions.

Observations records were aggregated from the *observation* dataset (see Table II), comprising 19,381,019 patients and 12,449 unique observation types. Each observation can range from one to a 'long-term stay' in the hospital.

Drug records were obtained from the *drug_era* dataset, detailed in Table II. Integration included records for 17,180,741 patients and 32,180 unique drug values.

Device records were aggregated from the *device_exposure* dataset, outlined in Table II, comprising records for 5,666,679 patients and 5,693 unique device concept values.

Measurement records originated from the *measurement* dataset, detailed in Table II. Integration included records for 21,899,515 patients and 27,487 unique device concept values.

Procedure records were aggregated from the *procedure_occurrence* dataset, outlined in Table II. Integration encompassed records for 15,673,100 patients and 57,970 unique device concept values.

Visit records were sourced from the *visit_occurrence* dataset, detailed in Table II. These records, integrated solely within the diabetic datasets, comprised 19,761,967 patients and 60 unique visit concepts.

For the **Target Labels** in this study, we will utilize the definition provided by [6] to identify relevant ICD-10-CM diagnosis codes for **long Covid**. From those, we will create our long covid label. The diabetes labels, **diabetes_1** and **diabetes_2**, refer to the 407 OMOP concepts codes linked to complicated diabetes and 127 OMOP concepts codes linked to uncomplicated diabetes. See Table III for target class distribution.

TABLE III
TARGET LABELS FOR DATAFRAMES

DataFrame	Long Covid		Diabetes Type 1		Diabetes Type 2	
Target Label	Positive - 1	Negative - 0	Positive - 1	Negative - 0	Positive - 1	Negative - 0
Drugs	1,231	4,084,873	1,042	2,616,764	1,416,166	2,616,764
Devices	272	1,275,318	1,320,700	4,491,089	775,877	5,035,912
Conditions	1,472	4,602,799	1,046	2,692,158	1,441,288	2,692,158
Observations	1,484	4,515,715	1,046	2,664,544	1,429,581	2,664,544
Measurements	1,311	4,407,944	1,046	21,898,469	1,437,147	20,462,368
Procedures	996	3,671,669	1,042	15,672,058	1,352,568	14,320,532
Visits	–	–	1,046	19,760,921	1,407,536	18,354,431

2) Feature Selection: Our methodology utilizes the embedded RF and LightGBM feature selection algorithms, which were informed by our analysis of the Kaggle dataset. Notably, our examination 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. Here, we leverage Random Forest (RF) alongside LightGBM algorithms for feature selection, capitalizing on their respective strengths to enhance the robustness and efficacy of our approach.

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. The efficiency and accuracy of the LightGBM algorithm offer built-in feature importance metrics. Similar to Random Forest, LightGBM identifies the most influential features within the dataset. 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.

The feature selection process of the Kaggle dataset led to

TABLE IV
KAGGLE TOP 11 FEATURES - LASSOCV, RANDOMFORESTCLASSIFIER AND BOOSTING MODELS - LIGHTGBM, XGBOOST, CATBOOST

LassoCV	RandomForestClassifier	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 inclusion of the visits dataframe in our N3C analysis. Table IV 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.

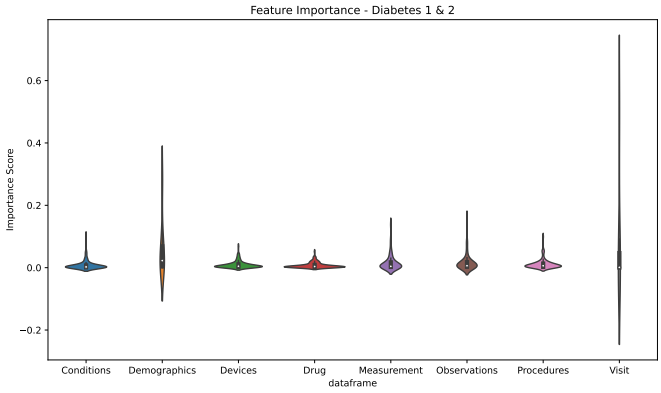

Fig. 5. N3C - Diabetes Feature Importance using Random Forest embedded feature scoring method

A discernible pattern emerges upon the top 30 features. Recurrent occurrences of acute upper respiratory infection, headache, nausea, and vomiting signify a potential association between these clinical manifestations and the likelihood of experiencing long-term consequences of COVID-19 infection. Table V includes various medical devices and procedures, indicating a potential link between specific medical interventions and the development of long COVID-19 symptoms. For example, the presence of devices such as oxygen nasal cannulas and aerosol oxygen masks may suggest a correlation between respiratory support measures and the persistence of COVID-19-related symptoms.

Furthermore, the inclusion of observations related to patient encounter status and tobacco use status underscores the

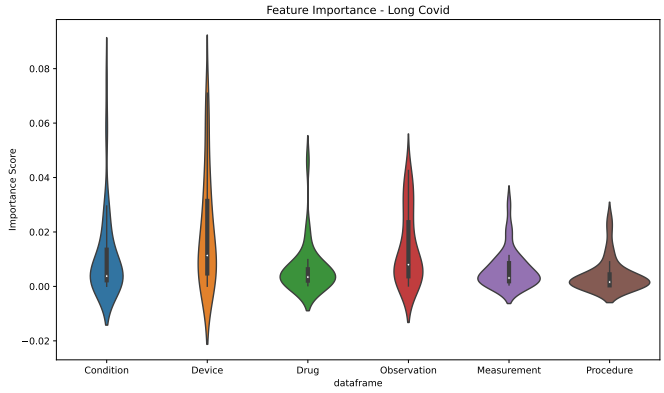

Fig. 6. N3C - Long Covid Feature Importance using Random Forest embedded feature scoring method

importance of considering individual health behaviors and healthcare utilization patterns in understanding the risk factors for long COVID-19.

3) Experiments: The experiments were designed to address two pivotal questions aimed at optimizing the predictive modeling process and enhancing the insights derived from the dataset:

- Sampling Strategy Evaluation: Determine the most effective sampling strategy (e.g., SMOTE, undersampling, resampling) to handle class imbalance
- Clustering for Insights: Apply clustering algorithm k-nearest neighbors (KNN) algorithm to cluster predictions based on different data frames (devices, conditions, drugs, observations, procedures, measurements).

The results of the sampling experiments provide valuable insights into the performance of different sampling techniques with various machine learning algorithms. Across all three algorithms (Random Forest, Gradient Boosting, LightGBM), the SMOTE technique consistently yielded the highest performance metrics in terms of accuracy, precision, recall, and F1 score. Specifically, when using the Random Forest classifier, SMOTE achieved an accuracy of 0.8826, outperforming both undersampling and simple resampling techniques. Similarly, for Gradient Boosting and LightGBM, SMOTE demonstrated superior performance compared to the other sampling methods. The e results suggest that addressing class imbalance through oversampling with SMOTE can lead to better predictive models, as evidenced by the higher accuracy and balanced precision-recall trade-offs. Conversely, undersampling techniques generally resulted in lower performance, indicating potential information loss due to the reduction in the majority class samples.

C. Modeling

Baseline Models, such as Random Forest and Gradient Boosting, will be utilized for comparison in this study.

1. Random Forest: Random Forest, a popular ensemble learning method, constructs multiple decision trees during

TABLE V
TOP 30 OVERALL CONCEPT IDS - RANDOM FOREST EMBEDDED FEATURE SELECTION - TARGET LABEL: LONGCOVID

Concept Id	Concept Name	Importance
condition_concept_id_257011	Acute upper respiratory infection	0.0771
device_concept_id_45110833	OPTIRAY 350 (350 MG/ML) SYRN	0.0710
device_concept_id_40664904	Injection, gadobutrol, 0.1 ml	0.0609
condition_concept_id_378253	Headache	0.0609
device_concept_id_4224038	Oxygen nasal cannula	0.0594
condition_concept_id_257011	Acute upper respiratory infection	0.0547
drug_concept_id_732893	bupivacaine	0.0462
device_concept_id_2614897	Surgical trays	0.0442
observation_concept_id_40217302	Clinical decision support mechanism national decision support company, as defined by the medicare appropriate use criteria program	0.0426
device_concept_id_2720868	Low osmolar contrast material, 100-199 mg/ml iodine concentration, per ml	0.0392
observation_concept_id_36307579	Current some day user	0.0389
observation_concept_id_443364	Patient encounter status	0.0376
device_concept_id_2615740	Anchor/screw for opposing bone-to-bone or soft tissue-to-bone (implantable)	0.0361
observation_concept_id_36305168	Smokeless tobacco status	0.0353
device_concept_id_2720522	Red blood cells, leukocytes reduced, each unit	0.0337
condition_concept_id_4273307	Platelet count - finding	0.0327
observation_concept_id_45881517	Current every day smoker	0.0322
observation_concept_id_40481872	Multigravida of advanced maternal age	0.0319
measurement_concept_id_2213115	Infectious agent detection by nucleic acid (DNA or RNA); Chlamydia trachomatis, amplified probe technique	0.0301
condition_concept_id_439658	Disorder of pregnancy	0.0295
observation_concept_id_4224504	Pulse	0.0278
condition_concept_id_4193704	Type 2 diabetes mellitus without complication	0.0266
observation_concept_id_4188893	History of clinical finding in subject	0.0252
procedure_concept_id_42628505	Drug test(s)	0.0249
condition_concept_id_195867	Noninflammatory disorder of the vagina	0.0249
condition_concept_id_37311061	COVID-19	0.0248
device_concept_id_2615762	Catheter, infusion, inserted peripherally, centrally or midline (other than hemodialysis)	0.0245
device_concept_id_4145694	Aerosol oxygen mask	0.0244
condition_concept_id_27674	Nausea and vomiting	0.0240

TABLE VI
SAMPLING EXPERIMENTS - RANDOM FOREST, GRADIENT BOOSTING, LIGHTGBM - KAGGLE DATASET

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

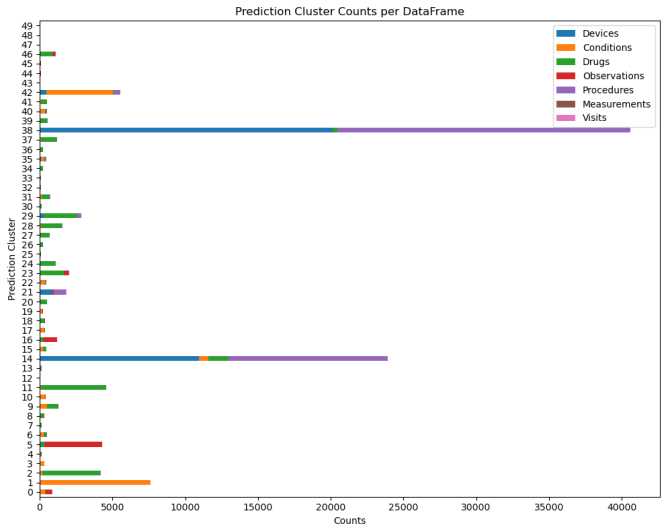

Fig. 7. N3C - Counts of prediction clusters per data frame using the KNN algorithm

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.

2. Gradient boosting: Gradient boosting, a powerful ensemble learning technique, sequentially trains weak learners to refine predictive models by focusing on residual errors. This approach captures complex data relationships, achieving high predictive accuracy in regression and classification tasks.

Its versatility has led to the development of advanced frameworks like LightGBM, enhancing performance and efficiency further.

State-of-the-art gradient boosting models, such as LightGBM and XGBoost, have made significant advancements in the field of machine learning as they further enhance performance and efficiency.

1. LightGBM: LightGBM is an advanced gradient-boosting framework that focuses on improving training speed and memory efficiency. It introduces the concepts of "Gradient-based One-Side Sampling" (GOSS) and "Exclusive Feature Bundling" (EFB) to accelerate training. GOSS selectively samples instances based on their gradients, prioritizing the ones that contribute more to the overall loss. EFB bundles mutually exclusive features together to reduce memory consumption. LightGBM's innovative techniques enable faster training while maintaining competitive performance.

2. XGBoost: XGBoost, an optimized gradient boosting library, implements parallel and distributed computing to improve scalability and performance. It incorporates regular-

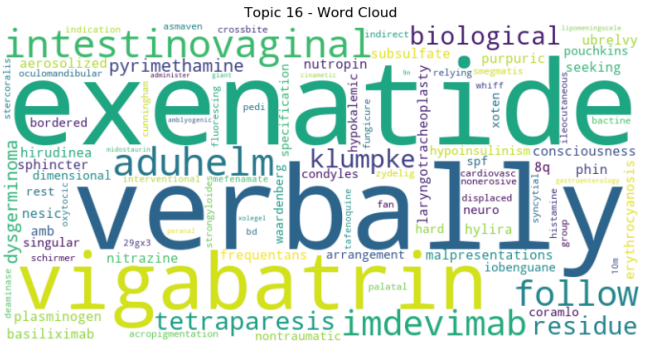

Fig. 8. Wordcloud depicting the distribution of terms in Topic 16 generated using k-means clustering with k=50.

TABLE VII
KAGGLE MODELING SCORES

Model	Accuracy	Precision	Recall	F1
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

ization techniques to prevent overfitting and is renowned for its speed and efficiency in handling large datasets. XGBoost's versatility and robustness make it a preferred choice for various machine-learning tasks, ranging from classification to ranking problems.

TABLE VIII
LONG COVID - GRADIENT BOOSTING MODELING SCORES FOR ALL ATTRIBUTES BY DATA FRAME

data frame	train	test	class_0	class_1	accuracy	precision_0	recall_0	f1-score_0	precision_1	recall_1	f1-score_1
Observations	250	63	164	165	0.4762	0.5	0.4242	0.459	0.4571	0.5333	0.4923
Conditions	2372	593	1493	1472	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Procedures	1614	404	1022	996	0.4233	0.4157	0.3366	0.372	0.4286	0.5126	0.4668
Measures	2104	526	1319	1311	0.4278	0.407	0.3068	0.3499	0.4404	0.5496	0.489
Devices	448	113	289	272	0.4956	0.5455	0.1034	0.1739	0.4902	0.9091	0.6369
Drugs	1979	495	1243	1231	0.4505	0.4428	0.3574	0.3956	0.4558	0.5447	0.4963

V. CONCLUSION AND FUTURE WORK

In this study, we investigated the predictive modeling of long-term consequences of COVID-19 infection, mainly focusing on the risk factors associated with long-term COVID-19 and diabetic patients. Leveraging the National COVID Cohort Collaborative (N3C) dataset, encompassing electronic

TABLE IX
LONG COVID - GRADIENT BOOSTING MODELING SCORES FOR SELECTED ATTRIBUTES BY DATA FRAME

data frame	train	test	class_0	class_1	accuracy	precision_0	recall_0	f1-score_0	precision_1	recall_1	f1-score_1
Observations	212	54	124	136	0.4074	0.4000	0.5600	0.4667	0.4211	0.2759	0.3333
Conditions	2292	573	1388	1428	0.4066	0.0923	0.0209	0.0341	0.4469	0.7937	0.5718
Procedures	1204	302	771	735	0.4272	0.3875	0.2000	0.2638	0.4414	0.6667	0.5312
Measures	2033	509	1284	1258	0.4637	0.4574	0.3346	0.3865	0.4673	0.5952	0.5236
Devices	299	75	192	185	0.4800	0.4872	0.5000	0.4935	0.4722	0.4595	0.4658
Drugs	1331	333	853	811	0.4054	0.3247	0.1462	0.2016	0.4297	0.6790	0.5263

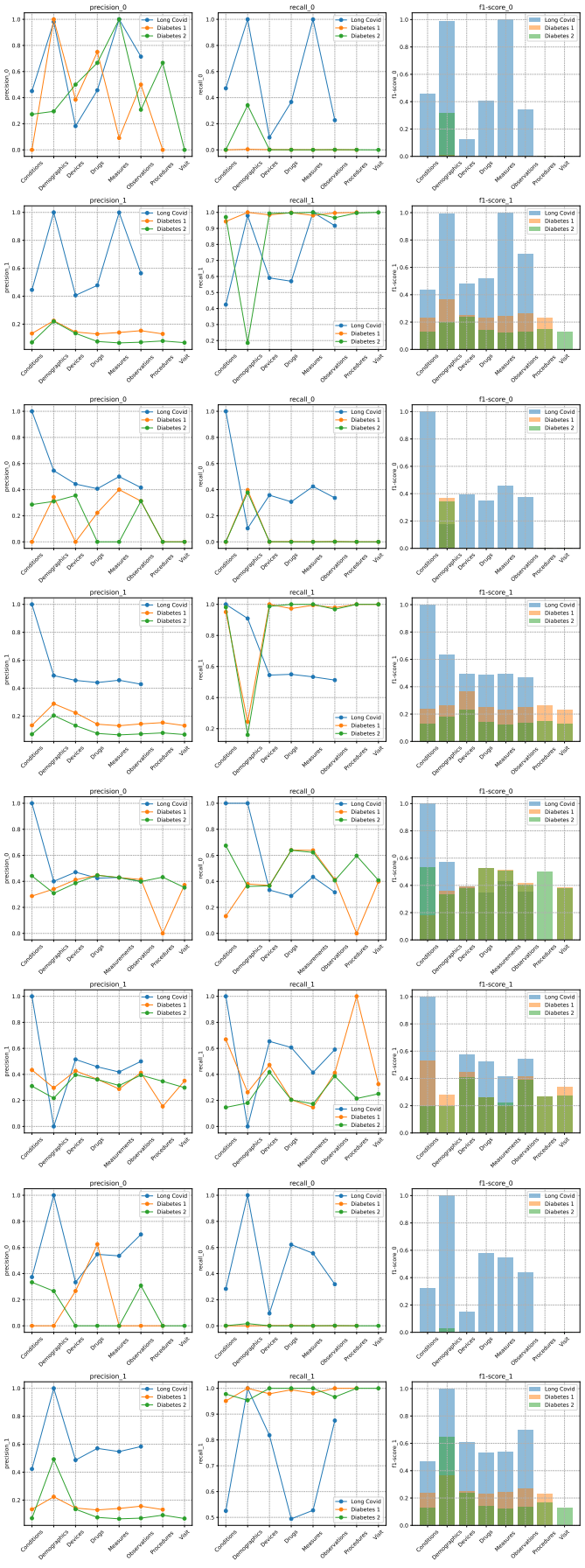

Fig. 9. Precision, Recall, and F1-score for both Class 0 and Class 1 across all attributes, categorized by data frame, for models Random Forest, Gradient Boosting, LightGBM, and XGBoost

Fig. 10. Precision, Recall, and F1-score for both Class 0 and Class 1 across selected attributes, categorized by data frame, for models Random Forest, Gradient Boosting, LightGBM, and XGBoost

TABLE X
LONG COVID - GRADIENT BOOSTING MODELING SCORES FOR KMEANS CLUSTERS WITH K=50 BY DATA FRAME

DataFrame	Train	Test	Class_0	Class_1	Accuracy	Precision_0	Recall_0	F1-score_0	Precision_1	Recall_1	F1-score_1
Observations	2378	595	449899	1508	0.4403	0.4383	0.4846	0.4603	0.4428	0.3974	0.4188
Conditions	2346	587	457384	1492	0.5077	0.3333	0.0035	0.0069	0.5086	0.9933	0.6727
Procedures	1989	498	365846	1268	0.4137	0.3763	0.2992	0.3333	0.4375	0.5236	0.4767
Measures	2142	536	406995	1360	0.4254	0.4035	0.3485	0.3740	0.4416	0.5000	0.4690
Devices	2372	593	452857	1506	0.4165	0.3969	0.3562	0.3755	0.4320	0.4751	0.4525
Drugs	824	206	127569	526	0.4272	0.3333	0.1683	0.2237	0.4581	0.6762	0.5462

health records from diverse healthcare institutions, we conducted comprehensive data analysis and modeling experiments to elucidate critical insights and develop accurate predictive models. Our analysis revealed significant associations between various clinical manifestations, medical interventions, and the likelihood of experiencing long-term COVID-19 symptoms. Through feature selection techniques, we identified critical attributes contributing to the prediction of long-term COVID-19 and diabetic outcomes ...*TODO: features identified in both –Mirna*

Utilizing state-of-the-art gradient boosting algorithms, including LightGBM and XGBoost, we developed robust predictive models capable of accurately identifying individuals at risk of long-term COVID-19 and diabetic complications.

Overall, this research contributes to advancing our understanding of the long-term health consequences of COVID-19 and diabetic conditions, providing valuable insights for healthcare providers, policymakers, and researchers. By leveraging large-scale electronic health records and sophisticated machine learning techniques, we pave the way for targeted intervention strategies, improved resource allocation, and personalized healthcare delivery. In future work, it is crucial to delve deeper into population-specific analysis to understand how demographic and socioeconomic factors influence long COVID-19 and diabetic outcomes. The investigation should consider disparities in healthcare access, social determinants of health, and cultural differences, aiming to address health inequities and tailor interventions accordingly. Additionally, integrating diverse data modalities, such as genomic, proteomic, and environmental factors, holds promise for capturing comprehensive patient profiles and enhancing predictive modeling performance. Furthermore, incorporating longitudinal data into analyses can provide valuable insights into the progression of long COVID-19 symptoms and diabetic complications over time. This approach enables the identification of dynamic risk factors and temporal patterns, facilitating more accurate predictions and personalized healthcare interventions.

VI. ACKNOWLEDGMENT

”The analyses described in this [publication/report/presentation] 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 and [insert additional funding agencies or sources and reference numbers]. 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] Centers for disease control and prevention$_2$023, $Jul2023$.

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

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

[4] 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.

[5] Melissa A Haendel, Christopher G Chute, Tellen D Bennett, David A Eichmann, Justin Guinney, Warren A Kibbe, Philip R O Payne, Emily R Pfaff, Peter N Robinson, Joel H Saltz, Heidi Spratt, Christine Suver, John Wilbanks, Adam B Wilcox, Andrew E Williams, Chunlei Wu, Clair Blacketer, Robert L Bradford, James J Cimino, Marshall Clark, Evan W Colmenares, Patricia A Francis, Davera Gabriel, Alexis Graves, Raju Hemadri, Stephanie S Hong, George Hripscak, Dazhi Jiao, Jeffrey G Klann, Kristin Kostka, Adam M Lee, Harold P Lehmann, Lora Lingrey, Robert T Miller, Michele Morris, Shawn N Murphy, Karthik Natarajan, Matvey B Palchuk, Usman Sheikh, Harold Solbrig, Shyam Visweswaran, Anita Walden, Kellie M Walters, Griffin M Weber, Xiaohan Tanner Zhang, Richard L Zhu, Benjamin Amor, Andrew T Girvin, Amin Manna, Nabeel Qureshi, Michael G Kurilla, Sam G Michael, Lili M Portilla, Joni L Rutter, Christopher P Austin, Ken R Gersing, and the N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*, 28(3):427–443, 08 2020.

[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] Barbara Chmielewska, Imogen Barratt, Rosemary Townsend, Erkan Kalafat, Jan van der Meulen, Ipek Gurol-Urganci, Pat O'Brien, Edward Morris, Tim Draycott, Shakila Thangaratinam, and et al. Effects of the covid-19 pandemic on maternal and perinatal outcomes: A systematic review and meta-analysis. *The Lancet Global Health*, 9(6), 2021.

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

[9] 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.

[10] Piotr Florek and Adam Zagdański. Benchmarking state-of-the-art gradient boosting algorithms for classification, May 2023.

[11] Hiroe Seto, Asuka Oyama, Shuji Kitora, Hiroshi Toki, Ryohei Yamamoto, Jun'ichi Kotoku, Akihiro Haga, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, and et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific Reports*, 12(1), 2022.

[12] Saumendra Kumar Mohapatra, Abhishek Das, and Mihir Narayan Mohanty. An optimized ensemble model for covid detection. *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*, 2022.

[13] Linyuan Jing, John M. Pfeifer, Dustin Hartzel Martin Kang, Sushravya Raghunath, Brandon K. Fornwalt, and Christopher M. Haggerty. Poster: An ehr-based machine learning model predicts myocardial infarction better than an ecg-based machine learning model and the pooled cohort equations. In *AHA Scientific Session 2022: Measuring Outcomes in ACS: Our Lives Depend on It*, October 2022.

[14] Chih-Chou Chiu, Chung-Min Wu, Te-Nien Chien, Ling-Jing Kao, Chengcheng Li, and Han-Ling Jiang. Applying an improved stacking ensemble model to predict the mortality of icu patients with heart failure. *Journal of Clinical Medicine*, 11(21):6460, 2022.