Long COVID Challenge: Predictive Modeling of Noisy Clinical Tabular Data

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

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

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, also known as post-acute sequelae of SARS-CoV-2 infection (PASC), refers to a range of persistent symptoms and complications experienced by individuals even after recovering from the acute phase of COVID-19.

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. [1]

The primary objective of this study is to enhance the accuracy of re-hospitalization prediction, aiming to provide evidence-based recommendations that also consider the long-term consequences. By improving the performance of predictive models, this research seeks to assist healthcare providers and policymakers with the necessary tools to make informed decisions regarding optimal resource allocation, effective intervention strategies, and targeted support for high-risk individuals. By bridging the gap between data analysis and actionable insights, this study aims to contribute to the improvement of diabetic patient care, including the challenges posed by Long COVID. Electronic Health Medical Records have a long tail of definite concepts, as illustrated in Table **??**. The high dimensionality and sparsity of the datasets make it highly challenging to elicit meaningful insights, even when the number of subjects is high.

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. The dataset for this study 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. [2]

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

TODO: write clear outline and documents (working on rewrite) –Mirna

1) Covid: A recent study investigated the link between preexisting conditions and the risk of experiencing post-acute sequelae of COVID-19 (PASC). The study revealed that individuals with conditions such as asthma, chronic constipation, reflux, rheumatoid arthritis, seasonal allergies, and depression/anxiety may face an increased risk of developing PASC. It strongly emphasized the importance of individuals with these conditions taking extra precautions to protect themselves against COVID-19 [1].

In the aftermath of the COVID-19 pandemic, a definition for long COVID was established by considering the signs, symptoms, and diagnoses that exhibited a higher frequency after a positive COVID test compared to before testing. This definition encompasses 323 ICD-10-CM diagnosis codes that are grouped into 143 functional groups, identifying 17 medical-specialty long COVID subtypes. It was observed that patients with more severe cases of COVID-19 and multiple comorbidities had a greater likelihood of developing long COVID. This underscores the urgent need for monitoring and treatment programs specifically designed for individuals identified as being at risk [3].

2) Females: A systematic review and meta-analysis were conducted to investigate the impact of the COVID-19 pandemic on maternal, fetal, and neonatal outcomes. The analysis revealed significant increases in stillbirth and maternal death, while overall preterm birth rates did not show significant changes. However, high-income countries experienced reductions in preterm births, along with increased rates of surgically managed ectopic pregnancies [4].

3) Diabetes: Feature selection using fuzzy entropy measures with a similarity classifier is vital in classification due to its ability to simplify models, reduce computational costs, and decrease the need for extensive measurements in practical applications. Moreover, it enhances transparency, making diagnoses more understandable, which is especially valuable in medical contexts. In a study using a feature selection method based on fuzzy entropy measures with four medical datasets, improved classification accuracy was achieved with fewer features compared to the original datasets. For example, the Parkinson's dataset achieved a mean accuracy of 85.03% using only two features out of the original 22. In comparison, the dermatology dataset achieved a mean accuracy of 98.28% with 29 features instead of the original 34. [5] Logistic Regression favors one of the highly correlated attributes, and it is not a good baseline for our problem [**?**].

4) Feature Selection: Another paper explores the significance of feature selection in biomedical data mining, emphasizing the importance of computationally efficient methods that capture complex associations, such as feature interactions. Relief-based algorithms, including the original Relief algorithm and its variants, are discussed in detail, showcasing their effectiveness in balancing computational efficiency and sensitivity to complex patterns of association. [6] Additionally, stability selection, which combines feature selection with resampling, can be used as a technique that measures the frequency of feature selection across multiple random subsets of data, identifying stable and informative features.

5) Modeling: The effectiveness of different gradient-boosting models has been applied to a multitude of research problems. In this study, we will be using the original GBM algorithm as a baseline and three newer variants: XGBoost, LightGBM, and CatBoost. In comparative research, four state-of-the-art gradient boosting methods were implemented (GBM, XGBoost, LightGBM, and CatBoost) for classification tasks. They focused on hyperparameter optimization techniques, such as randomized search and Bayesian optimization, to enhance their performance, aiming to identify a gradient boosting variant that balances effectiveness, reliability, and ease of use . [7]

The results indicate that the GBDT model, implemented with LightGBM, exhibits higher reliability and accuracy in diabetes prediction compared to LR, highlighting the potential of machine learning for developing reliable prediction models in diabetes prevention [8]

Electronic health records (EHRs) encompass a wide array of medical data, including medical history, diagnoses, medications, laboratory results, and other relevant information. The EHR-aware gradient boosting method leverages this valuable data source to enhance the algorithm's predictions and make them more applicable in healthcare settings. When comparing the performance of various models, the EHR-based XGBoost model exhibited superior performance, achieving an impressive area under the receiver operating characteristic curve (AUROC) of 81%. The EHR-based model identifies a subset of patients at high risk for MI who may benefit from targeted interventions. [9]

Leveraging the MIMIC-III dataset, this study focuses on predicting mortality in heart failure patients during their ICU admission. A stacking ensemble learning model is proposed, incorporating six machine learning algorithms as first-level classifiers and LightGBM as the second-level classifier. The study aims to improve mortality prediction by harnessing the power of ensemble learning techniques [10]

This paper proposes an ensemble learning model to classify individuals at risk of COVID-19 infection. The optimized XGBoost model achieves the highest accuracy of 100%, surpassing the other models, demonstrating its potential in COVID-19 risk classification [11]

III. N3C Data Characteristics

TABLE I
ORIGINAL N3C DATASETS

Dataset	# Patients	# Concepts	rows	columns
condition_occurrence	20,140,630	59,985	2,989,641,029	21
condition_era	19,876,988	65,140	1,135,412,171	8
death	664,492	13	687,970	11
person	21,357,315	N/A	21,357,315	27
device_exposure	6,203,914	5,784	539,290,184	19
observation	21,162,196	12,394	2,881,068,441	25
drug_era	18,638,834	31,252	1,072,070,035	9

TABLE II
FEMALE COHORT N3C DATASETS

Dataset	# Patients	# Concepts	rows	columns
condition_occurrence	19,020,523	58,990	2,761,751,044	21
condition_era	18,762,634	64,128	1,057,039,277	8
death	379,517	12	402,954	11
person	11,911,975	N/A	11,911,975	27
device_exposure	5,811,994	5,688	525,437,270	19
observation	19,882,126	12,182	2,695,283,903	25
drug_era	17,624,458	31,042	991,189,063	9

IV. Diabetic Data Characteristics

V. Kaggle Dataset Characteristics

VI. Long Tail Data Source Aggregation

A. Kaggle Hospital Readmission

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. For this study, we will be focusing on the identified female cohort; see Table **??**.

TABLE III
DISTRIBUTION OVER AGGREGATED GENDER, RACE, ETHNICITY, AND AGE GROUP CATEGORIES IN THE DEMOGRAPHICS DATA FRAME

Gender	Female	11623774
	Male	9121849
Ethnicity	Hispanic or Latino	2661276
	Not Hispanic or Latino	15891727
Race	African_American	0
	American_Indian_or_Alaska_Native	108550
	Asian_Indian	0
	Bangladeshi	1913
	Barbadian	112
	Bhutanese	6
	Burmese	31
	Cambodian	31
	Chinese	4191
	Dominica_Islander	258
	Filipino	2387
	Haitian	833
	Hispanic	243
	Hmong	1
	Indonesian	81
	Jamaican	676
	Japanese	1461
	Korean	1070
	Laotian	248
	Madagascar	7
	Malaysian	18
	Melanesian	55
	Micronesian	38
	Middle_Eastern_or_North_African	496
	Multiples	0
	Native_Hawaiian_or_Other_Pacific_Islander	73951
	Nepalese	118
	Okinawan	33
	Other	0
	Pakistani	590
	Polynesian	26
	Singaporean	20
	Sri_Lankan	51
	Taiwanese	134
	Thai	82
	Trinidadian	424
	Vietnamese	483
	West_Indian	425
	White	13122200
Age	>90	344567
	70-90	2944989
	50-70	545169
	30-50	5466857
	10-30	423623
	0-10	2673069

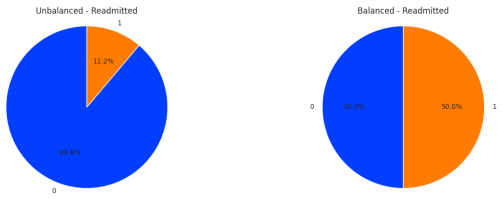

Fig. 1. True Class Counts - Hospital Readmission (left) and Balanced Set Counts - Hospital Readmission (right) *TODO: cut the white around circles, put them on the same slide, add (a) and (b) and include as one figure. – Mirna*

TABLE IV
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

TABLE V
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

B. Data Pre-Processing

1) Enclave Pipeline: The data frame **Demographics** aggregates the information provided in the **person** and **death** data sets, described in Table **??**, it has 49 attributes: *person_id*, age, two binary genders, 41 binary race and ethnicity, and one longCovid, one diabetes_1, one diabetes_2 label attribute.

Conditions per patient and their occurrence and duration records were obtained from *condition_occurrence*, *condition_era* (see Table **??**) resulting in aggregated 19, patient records. Each patient had at least one condition out of 59,595 unique conditions.

Observations records are aggregated from *observation* data set see (Table **??**), and the records were recovered for 20,380,71 patients for 1,225 unique observations. Each observation can last from 1 to a 'long-term stay' in the hospital, and multiple observations can be observed in a single patient.

Drugs *drug_era* and *drug_exposure* data sources, described in Table **??**. We integrated records for 18,357,01 patients and 4,186 unique drug values. We integrated all those records that can be observed in a single patient.

Devices records are aggregated from *device_exposure* data source, described in Table **??**. We integrated records for 33,816,215 patients and 4186* unique drug values. We integrated all those records that can be observed in a single patient.

Target Labels

In this study, we will utilize the definition provided by [3] to identify relevant ICD-10-CM diagnosis codes for **long-Covid**, see Table **??** for class distribution. From those, we will create our Long Covid label. The Diabetes labels, **diabetes_1**

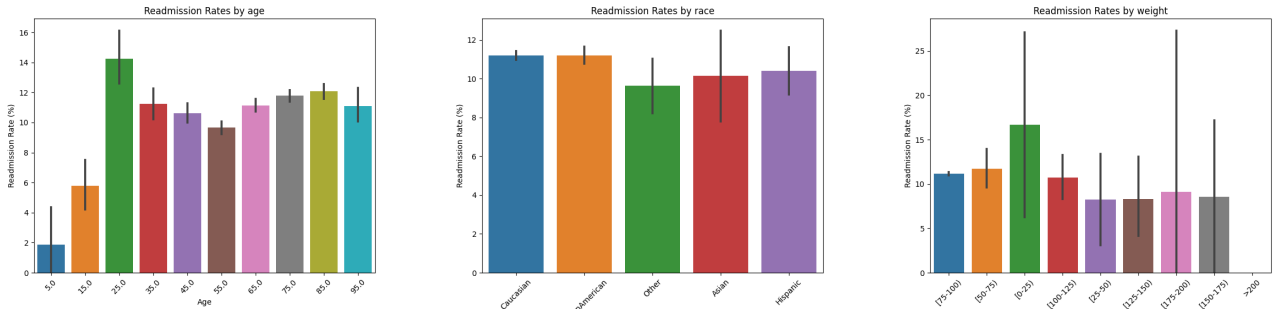

Fig. 2. Kaggle Demographic Characteristics

and **diabetes_2**, refer to the 407 OMOP concepts codes linked to complicated diabetes and 127 OMOP concepts codes linked to uncomplicated diabetes.

C. Feature Selection

(1) Random Forest embedding has a built-in feature importance measured by the Gini importance or mean decrease impurity. We propose to set the 50th percentile threshold for the importance of the attribute to include a relevant attribute in the final set. (2) LightGBM has a built-in feature that is important because of its efficiency and accuracy. Similar to Random Forest, LightGBM also provides feature importance metrics. This helps identify the most influential features in the dataset.

Random Forest

LightGBM

D. Experiments

E. Modeling

Baseline: Gradient boosting is a popular machine-learning technique used for both regression and classification tasks. It is an ensemble method that combines multiple weak or base learners to create a robust predictive model. The key idea behind gradient boosting is to sequentially train new models that focus on correcting the mistakes made by the previous models in the ensemble.

State-of-the-art gradient boosting models, such as XGBoost, LightGBM, and CatBoost, have made significant advancements in the field of machine learning.

1. XGBoost (Extreme Gradient Boosting): XGBoost is an optimized implementation of the gradient boosting algorithm. It introduces several improvements to the traditional GBM, making it more efficient and powerful. XGBoost incorporates techniques such as parallelization, regularization, and tree pruning to enhance model performance. It also provides a flexible interface and supports various objective functions and evaluation metrics. XGBoost has gained popularity due to its high scalability, speed, and ability to handle large-scale datasets.

2. LightGBM: LightGBM is another advanced gradient-boosting framework that focuses on improving training speed and memory efficiency. It introduces the concept 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.

3. CatBoost: CatBoost is a gradient boosting framework that specializes in handling categorical features effectively. It incorporates novel strategies to handle categorical variables, such as applying a combination of target statistics and ordered boosting to encode and utilize categorical information. CatBoost also implements advanced techniques like the "Ordered Boosting" algorithm, which reduces the complexity of creating and selecting split points during tree construction. This results in improved accuracy and faster training on datasets with categorical features.

VII. Conclusion and Future Work

TODO: –Mirna

VIII. 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] 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.

TABLE VI
TOP 30 CONCEPT ID - 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 VII
TOP 11 RANDOM FOREST DEVICE CONCEPTS

Device Concept Name	**Importance**
OPTIRAY (350 MG/ML) Syringe	0.0710
Injection, gadobutrol, 0.1 ml	0.0609
Oxygen nasal cannula	0.0594
Surgical trays	0.0442
Low osmolar contrast material, 100-199 mg/ml iodine concentration, per ml	0.0392
Anchor/screw for opposing bone-to-bone or soft tissue-to-bone (implantable)	0.0361
Red blood cells, leukocytes reduced, each unit	0.0337
Catheter, infusion (other than hemodialysis)	0.0245
Aerosol oxygen mask	0.0244
Tissue marker, implantable, any type, each	0.0221

TABLE VIII
TOP 11 RANDOM FOREST CONDITION CONCEPTS

Condition Concept Name	**Importance**
Acute upper respiratory infection	0.0771
Headache	0.0609
Acute upper respiratory infection	0.0547
Platelet count - finding	0.0327
Disorder of pregnancy	0.0295
Type 2 diabetes mellitus without complication	0.0266
Noninflammatory disorder of the vagina	0.0249
COVID-19	0.0248
Nausea and vomiting	0.0240
condition duration	0.0222
Hypothyroidism	0.0189

TABLE IX
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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

TABLE X
KNN - PREDICTION CLUSTERS COUNTS PER DATAFRAME - N3C

Prediction Cluster	Devices	Conditions	Drugs	Observations	Procedures	Measurements
0	29	354	36	434	29	4
1	1	7600	24	2	1	5
2	35	129	3990	3	35	1
3	10	294	19	1	10	1
4	2	9	99	13	2	1
5	18	11	312	3937	18	4
6	42	254	168	1	42	1
7	3	5	124	10	3	1
8	1	44	253	1	1	1
9	4	496	761	3	4	3
10	3	385	2	13	3	6
11	2	88	4471	6	2	1
12	1	12	15	9	1	1
13	34	26	9	28	34	1
14	10938	636	1410	2	10938	4
15	3	215	222	2	3	1
16	4	6	204	981	4	1
17	2	316	35	14	2	1
18	1	1	343	36	1	1
19	15	174	16	15	15	1
20	2	8	502	10	2	1
21	854	6	31	77	854	1
22	7	343	44	40	7	1
23	10	46	1626	320	10	4
24	29	22	998	8	29	1
25	1	2	80	19	1	1
26	8	42	180	3	8	1
27	2	5	699	2	2	1
28	10	83	1440	9	10	1
29	280	9	2281	28	280	1
30	2	49	75	1	2	1
31	3	125	572	8	3	2
32	21	5	19	15	21	2
33	14	22	13	16	14	0
34	1	2	236	7	1	0
35	77	248	35	17	77	0
36	14	1	206	7	14	0
37	51	43	1045	10	51	0
38	20121	38	315	6	20121	0
39	5	34	471	17	5	0
40	56	295	106	1	56	0
41	3	97	408	1	3	0
42	462	4569	36	6	462	0
43	1	47	13	2	1	0
44	13	31	9	25	13	0
45	4	15	28	62	4	0
46	4	4	872	202	4	0
47	1	17	25	1	1	0
48	9	3	2	1	9	0
49	9	3	36	2	9	0

TABLE XI
LIGHTGBM MODELING SCORES FOR ALL ATTRIBUTES BY DATAFRAME

DataFrame	train	test	class_0	class_1	accuracy	precision_0	recall_0	f1_score_0	precision_1	recall_1	f1_score_1
Measures	35332	8833	440613	141	0.3717	0.9997	0.3716	0.5419	0.0002	0.5000	0.0004
Drugs	32757	8190	40936	118	0.9984	0.9998	0.9987	0.9992	0.0000	0.0000	0.0000
Devices	10158	2540	12695	26	0.6413	1.0000	0.6412	0.7814	0.0011	1.0000	0.0022
Procedures	29398	7350	36737	11	0.6190	0.9996	0.6192	0.7647	0.0000	0.0000	0.0000
Conditions	36980	9245	46209	146	0.7233	0.9997	0.7234	0.8394	0.0004	0.3333	0.0008
Observations	36174	9044	45202	163	0.5179	0.9996	0.5180	0.6824	0.0002	0.3333	0.0005

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

TABLE XII
LIGHTGBM MODELING SCORES FOR SELECTED ATTRIBUTES BY DATAFRAME

DataFrame	train	test	class_0	class_1	accuracy	precision_0	recall_0	f1_score_0	precision_1	recall_1	f1_score_1
Measures	35846	8962	447233	135	0.3121	0.9996	0.3121	0.4756	0.0002	0.5000	0.0003
Drugs	24506	6127	30625	94	0.6729	1.0000	0.6728	0.8044	0.0010	1.0000	0.0020
Devices	6988	1748	8736	0	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
Procedures	24009	6003	30002	10	0.9032	0.9996	0.9035	0.9491	0.0000	0.0000	0.0000
Conditions	33969	8493	42309	1405	0.6859	0.9979	0.6861	0.8132	0.0071	0.6129	0.0140
Observations	29586	7397	36970	130	0.9467	0.9997	0.9470	0.9726	0.0025	0.3333	0.0051

TABLE XIII
LIGHTGBM MODELING SCORES FOR PREDICTION CLUSTER ATTRIBUTES BY DATAFRAME

DataFrame	train	test	class_0	class_1	accuracy	precision_0	recall_0	f1_score_0	precision_1	recall_1	f1_score_1
Measures	32696	8174	40732	138	0.0948	0.9934	0.0923	0.1689	0.0031	0.8214	0.0062
Drugs	36428	9107	45388	147	0.1337	0.9942	0.1317	0.2327	0.0028	0.7586	0.0055
Devices	10205	2552	12702	55	0.0486	0.9913	0.0449	0.0858	0.0041	0.9091	0.0082
Procedures	36107	9027	44991	143	0.1549	0.9942	0.1530	0.2652	0.0027	0.7241	0.0055
Conditions	36789	9198	45832	155	0.0467	0.9926	0.0439	0.0840	0.0032	0.9032	0.0063
Observations	29389	7348	36603	134	0.8483	0.9965	0.8507	0.9178	0.0046	0.1852	0.0089