Diabetes Mellitus Prediction in ICU Patients

HENA GHONIA & SHANAYA MEHTA

Motivation

- Diabetes is a serious chronic health condition affecting millions of people across the planet

- In the ICU, **sick patients** needing urgent care might not be able to provide all information beforehand to care providers. Giving the wrong medicine might affect patient outcomes negatively

- Focus is to develop a predictive model to predict diabetes in sick ICU patients using data their data generated from the **first 24 hours of ICU admission**.

- According to doctors, the data from first 24 hours is critical to a patient's survival. In the absence of diabetes-specific tests, a predictive model helps in making **quick and informed decisions** and thus help save patient lives

Some statistics about diabetes

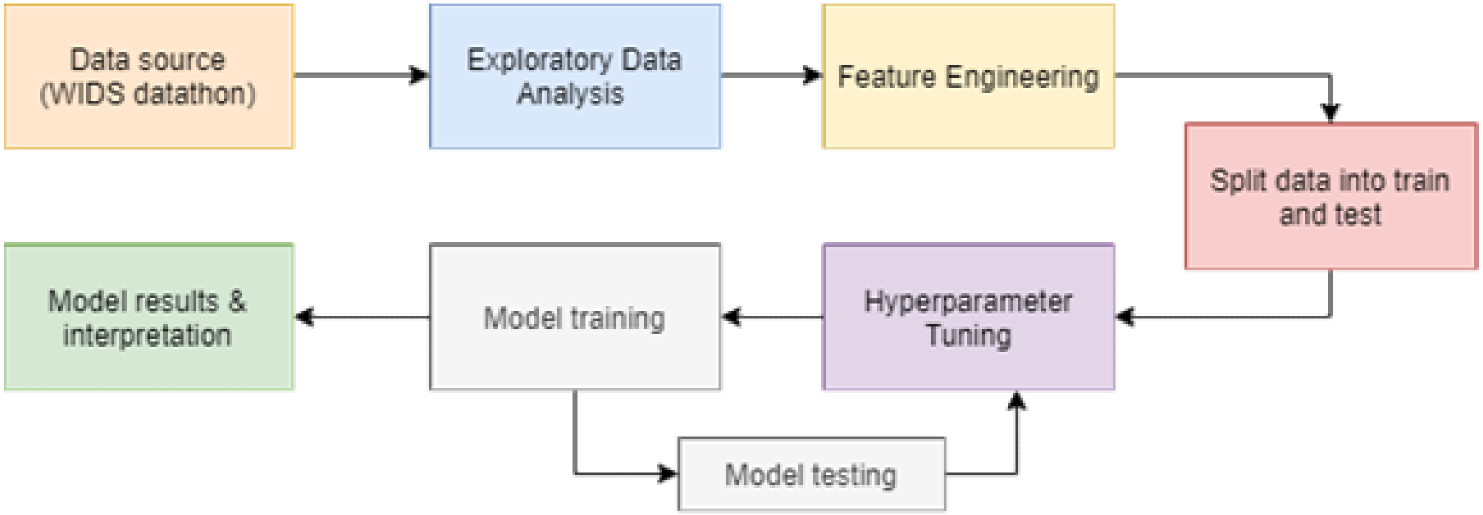
- 463 million adults aged 20-79 years have diabetes in the world today
- In 2045, 700 million adults might have diabetes globally
- 4.2 million deaths have occurred due to diabetes in 2019 alone
- It is the 7th top cause of death due to a disease
- 232 million people are predicted to be undiagnosed with this disease
- Diabetes affects patient outcomes and severity of disease for those having the Coronavirus
- In China in 2020, those having coronavirus with diabetes had a mortality rate of 20% while those who didn’t have diabetes had a mortality rate of 10%

Hypothesis

- Diabetes prediction is sensitive to blood glucose levels
- All features are stationary
- It is possible to predict diabetes in sick patients using a predictive model

Methodology

Methodology



The Data

Dataset Description

- The data is provided by the Global Open Source Severity of Illness Score (GOSSIS) consortium of the Massachusetts Institute of Technology (MIT)

- The consortium collects electronic health records (EHRs) of ICU patients from the USA, Australia and New Zealand.

- The collected data is primarily used to create an open source system for assessing the severity of illness in patients in critical care.

- Sample size of training data: 130,157
- Sample size of test data: 10,234
- Number of features: 180

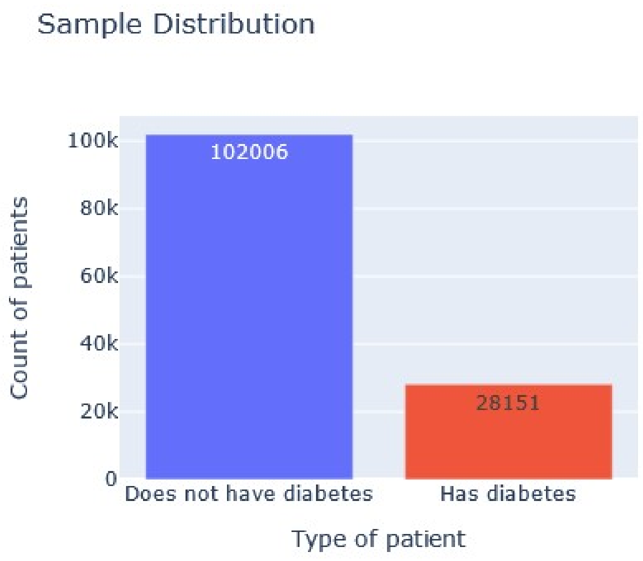
Features

FEATURE CATEGORY	COUNT
Lab Results	60
Vitals	52
APACHE* Covariate	28
Labs Blood Gas	16
Demographic	15
APACHE* Comorbidity	7
Identifier	2
Target	1

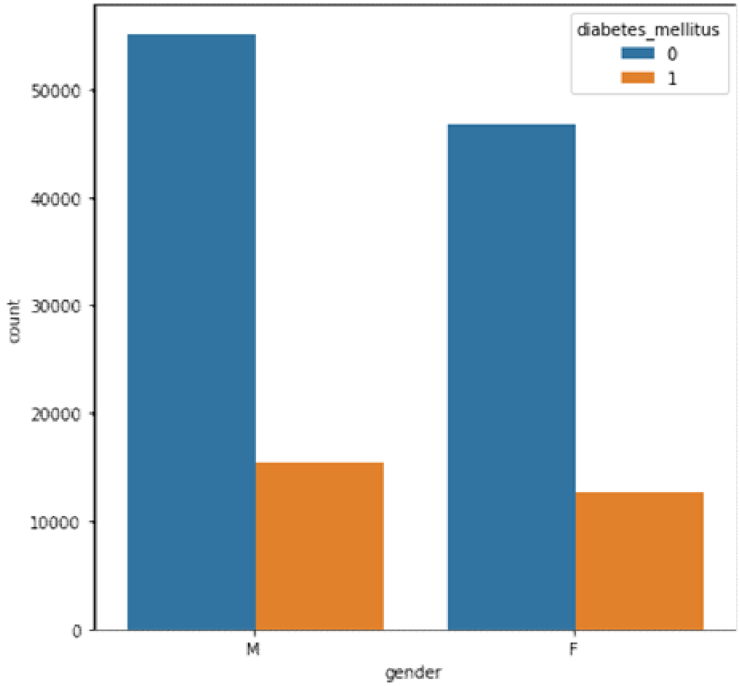
* APACHE: Acute Physiology and Chronic Health Evaluation

Exploratory Data Analysis

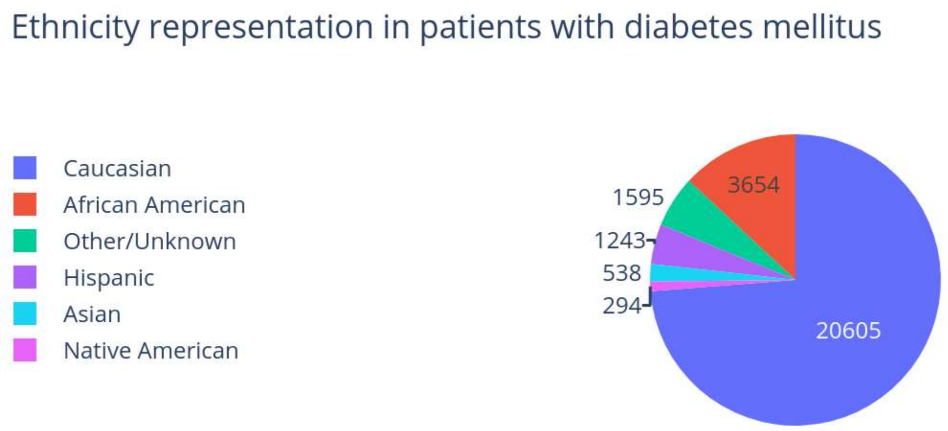
Target feature distribution



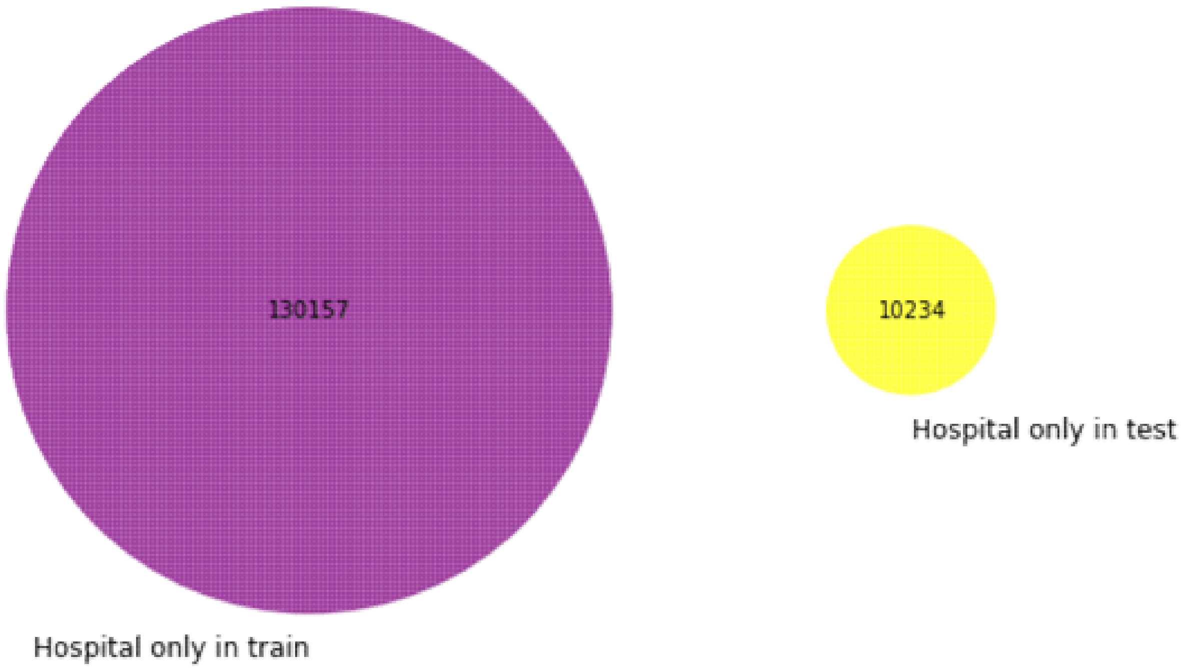
Gender distribution



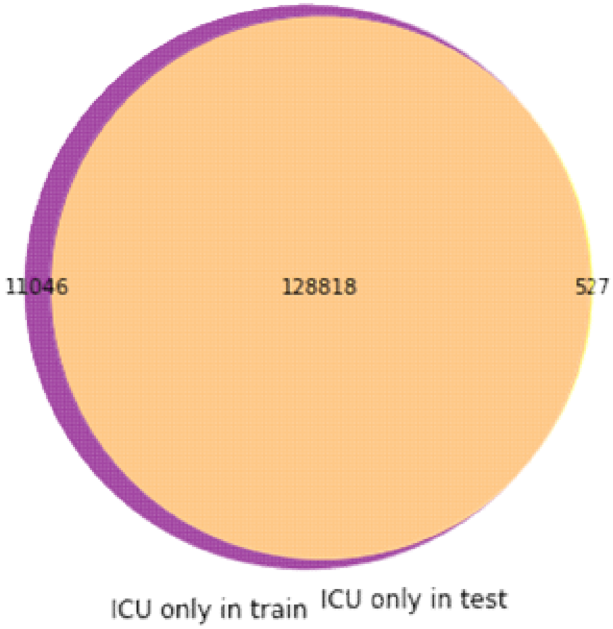
Ethnicity distribution



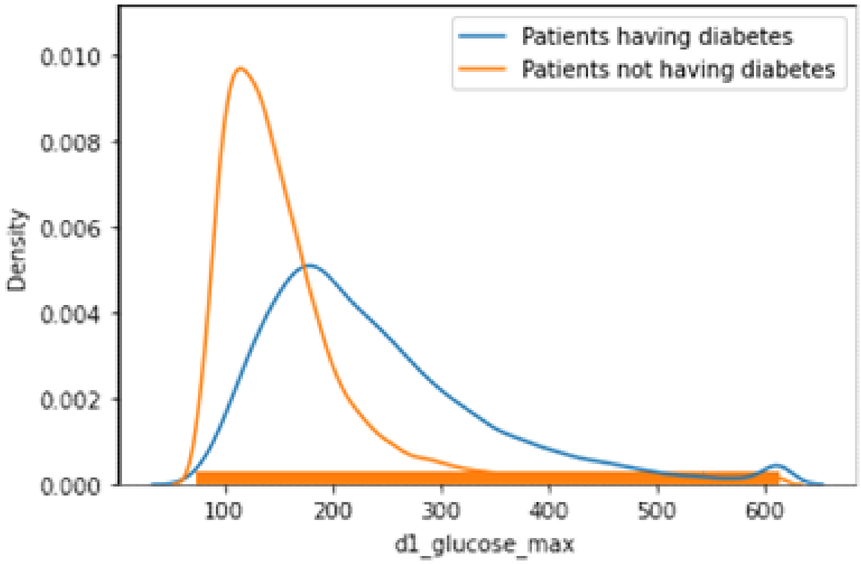
Hospital IDs in train and test data



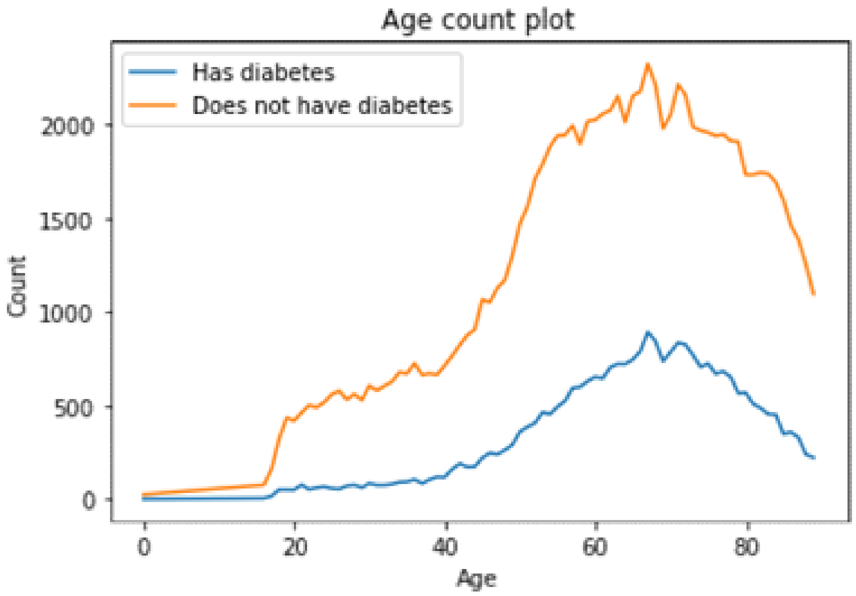
ICU IDs in train and test data



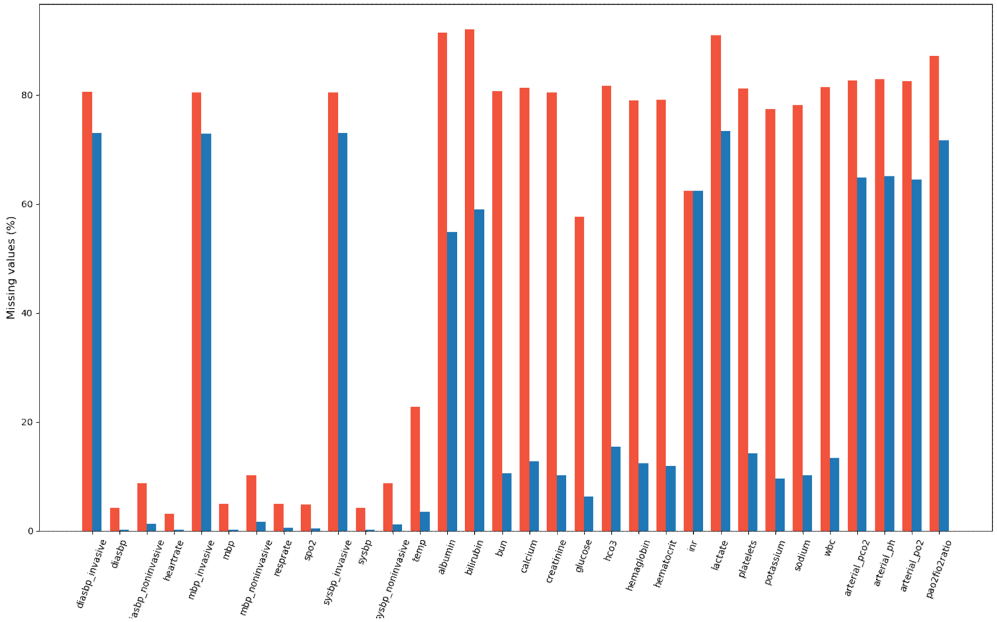
Max glucose readings distribution



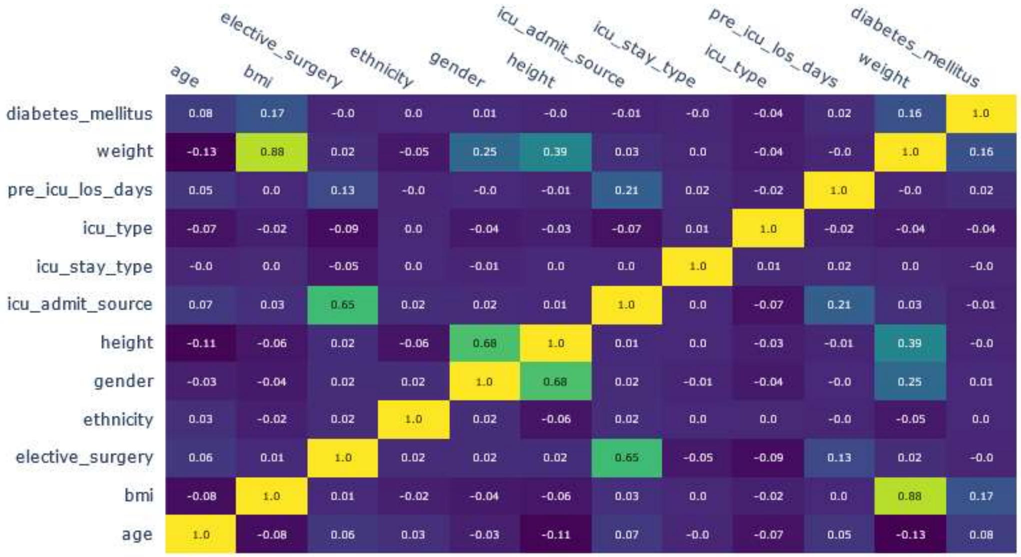
Age distribution



Missing values in hourly and full day readings

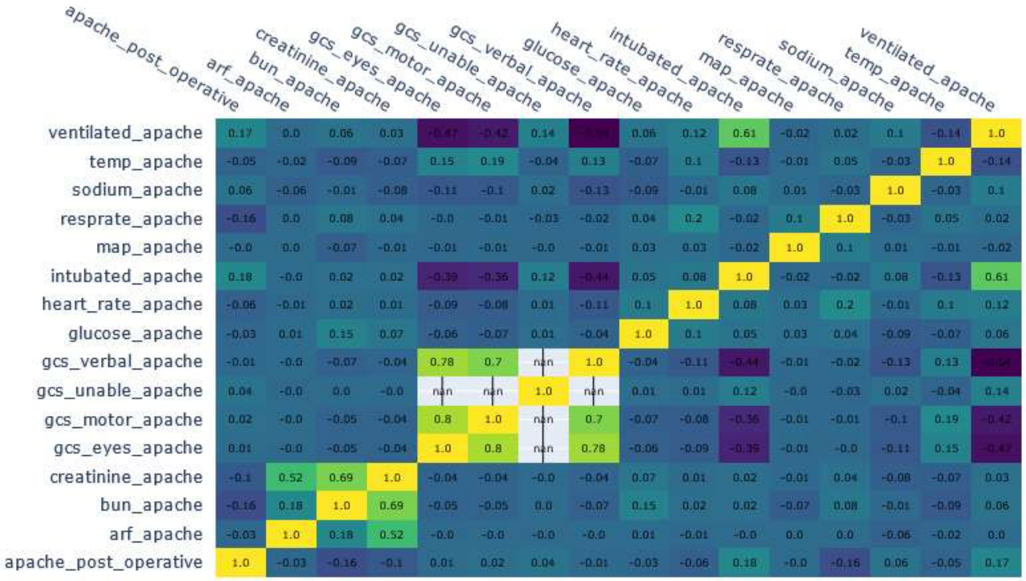


Correlation Analysis – Patient Demographics



	age	bmi	elective_surgery	ethnicity	gender	height	icu_admit_source	icu_stay_type	icu_type	pre_icu_los_days	weight	diabetes_mellitus
diabetes_mellitus	0.08	0.17	-0.0	0.0	0.01	-0.0	-0.01	-0.0	-0.04	0.02	0.16	1.0
weight	-0.13	0.88	0.02	-0.05	0.25	0.39	0.03	0.0	-0.04	-0.0	1.0	0.16
pre_icu_los_days	0.05	0.0	0.13	-0.0	-0.0	-0.01	0.21	0.02	-0.02	1.0	-0.0	0.02
icu_type	-0.07	-0.02	-0.09	0.0	-0.04	-0.03	-0.07	0.01	1.0	-0.02	-0.04	-0.04
icu_stay_type	-0.0	0.0	-0.05	0.0	-0.01	0.0	0.0	1.0	0.01	0.02	0.0	-0.0
icu_admit_source	0.07	0.03	0.65	0.02	0.02	0.01	1.0	0.0	-0.07	0.21	0.03	-0.01
height	-0.11	-0.06	0.02	-0.06	0.68	1.0	0.01	0.0	-0.03	-0.01	0.39	-0.0
gender	-0.03	-0.04	0.02	0.02	1.0	0.68	0.02	-0.01	-0.04	-0.0	0.25	0.01
ethnicity	0.03	-0.02	0.02	1.0	0.02	-0.06	0.02	0.0	0.0	-0.0	-0.05	0.0
elective_surgery	0.06	0.01	1.0	0.02	0.02	0.02	0.65	-0.05	-0.09	0.13	0.02	-0.0
bmi	-0.08	1.0	0.01	-0.02	-0.04	-0.06	0.03	0.0	-0.02	0.0	0.88	0.17
age	1.0	-0.08	0.06	0.03	-0.03	-0.11	0.07	-0.0	-0.07	0.05	-0.13	0.08

Correlation Analysis – APACHE Features



	apache_post_operative	arf_apache	bun_apache	creatinine_apache	gcs_eyes_apache	gcs_motor_apache	gcs_unable_apache	gcs_verbal_apache	glucose_apache	heart_rate_apache	intubated_apache	map_apache	resprate_apache	sodium_apache	temp_apache	ventilated_apache
ventilated_apache	0.17	0.0	0.06	0.03	-0.47	-0.42	0.14	-0.54	0.06	0.12	0.61	-0.02	0.02	0.1	-0.14	1.0
temp_apache	-0.05	-0.02	-0.09	-0.07	0.15	0.19	-0.04	0.13	-0.07	0.1	-0.13	-0.01	0.05	-0.03	1.0	-0.14
sodium_apache	0.06	-0.06	-0.01	-0.08	-0.11	-0.1	0.02	-0.13	-0.09	-0.01	0.08	0.01	-0.03	1.0	-0.03	0.1
resprate_apache	-0.16	0.0	0.08	0.04	-0.0	-0.01	-0.03	-0.02	0.04	0.2	-0.02	0.1	1.0	-0.03	0.05	0.02
map_apache	-0.0	0.0	-0.07	-0.01	-0.01	-0.01	-0.0	-0.01	0.03	0.03	-0.02	1.0	0.1	0.01	-0.01	-0.02
intubated_apache	0.18	-0.0	0.02	0.02	-0.39	-0.36	0.12	-0.44	0.05	0.08	1.0	-0.02	-0.02	0.08	-0.13	0.61
heart_rate_apache	-0.06	-0.01	0.02	0.01	-0.09	-0.08	0.01	-0.11	0.1	1.0	0.08	0.03	0.2	-0.01	0.1	0.12
glucose_apache	-0.03	0.01	0.15	0.07	-0.06	-0.07	0.01	-0.04	1.0	0.1	0.05	0.03	0.04	-0.09	-0.07	0.06
gcs_verbal_apache	-0.01	-0.0	-0.07	-0.04	0.78	0.7	nan	1.0	-0.04	-0.11	-0.44	-0.01	-0.02	-0.13	0.13	-0.54
gcs_unable_apache	0.04	-0.0	0.0	-0.0	nan	nan	1.0	nan	0.01	0.01	0.12	-0.0	-0.03	0.02	-0.04	0.14
gcs_motor_apache	0.02	-0.0	-0.05	-0.04	0.8	1.0	nan	0.7	-0.07	-0.08	-0.36	-0.01	-0.01	-0.1	0.19	-0.42
gcs_eyes_apache	0.01	-0.0	-0.05	-0.04	1.0	0.8	nan	0.78	-0.06	-0.09	-0.39	-0.01	-0.0	-0.11	0.15	-0.47
creatinine_apache	-0.1	0.52	0.69	1.0	-0.04	-0.04	-0.0	-0.04	0.07	0.01	0.02	-0.01	0.04	-0.08	-0.07	0.03
bun_apache	-0.16	0.18	1.0	0.69	-0.05	-0.05	0.0	-0.07	0.15	0.02	0.02	-0.07	0.08	-0.01	-0.09	0.06
arf_apache	-0.03	1.0	0.18	0.52	-0.0	-0.0	-0.0	-0.0	0.01	-0.01	-0.0	0.0	0.0	-0.06	-0.02	0.0
apache_post_operative	1.0	-0.03	-0.16	-0.1	0.01	0.02	0.04	-0.01	-0.03	-0.06	0.18	-0.0	-0.16	0.06	-0.05	0.17

Feature Engineering

Feature Engineering

- Target Encoding Categorical Features
  - Replacing categorical values with the mean of the target variable
- Recalculating Numerical Demographic Features
- Dropping Highly Correlated Variables
- Calculating New Features

Feature Engineering – New Features

- Differences between related features:
  - First hour readings and first day readings
  - Invasive and non-invasive readings
  - Blood pressure readings
  - Min and max readings of all variables (min and max readings are provided in the data)
  - Recalculated and recorded BMI
- Flag indicator variables that show if two features were equivalent
- Counts of missing values in APACHE readings, first hour and first day readings
- Average value of first day readings
- Recalculated demographic features

Feature Engineering – New Features

- New features from demographics such as age/weight, age/bmi
- Comorbidity score calculation
- Patient profile from existing demographics
- New categorical columns by adding two categorical columns
- Grouping data by a categorical feature
- Mean and standard deviation of numerical readings
- Normalization of numerical readings

Total number of columns after feature engineering: 1182

Baseline

Classifier	Accuracy
Dummy Classifier	78.37%
Random Forest	81.70%
Logistic Regression	80.00%

Algorithms in consideration

Traditional algorithms:

- LightGBM
- XGBoost
- CatBoost
- Stacked Ensembles
- Weighted Ensembles

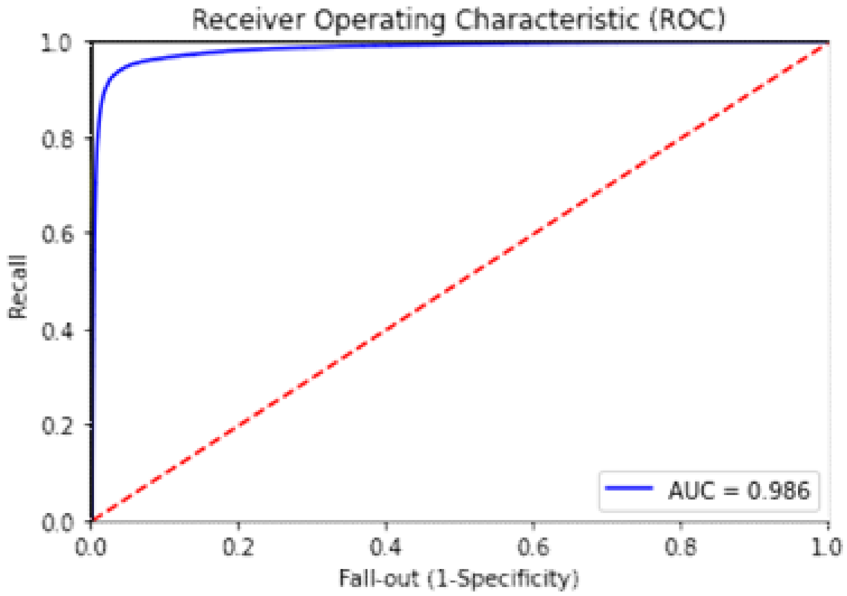
Model details

Model	Model Details
XGBoost	Single XGBoost
LightGBM	Ensemble of 2 LightGBMs
PyStackNet 1	2 layers, 2 classifiers: LightGBM & XGBoost
PyStackNet 2	2 Layers, 3 classifiers: CatBoost, XGBoost & LightGBM

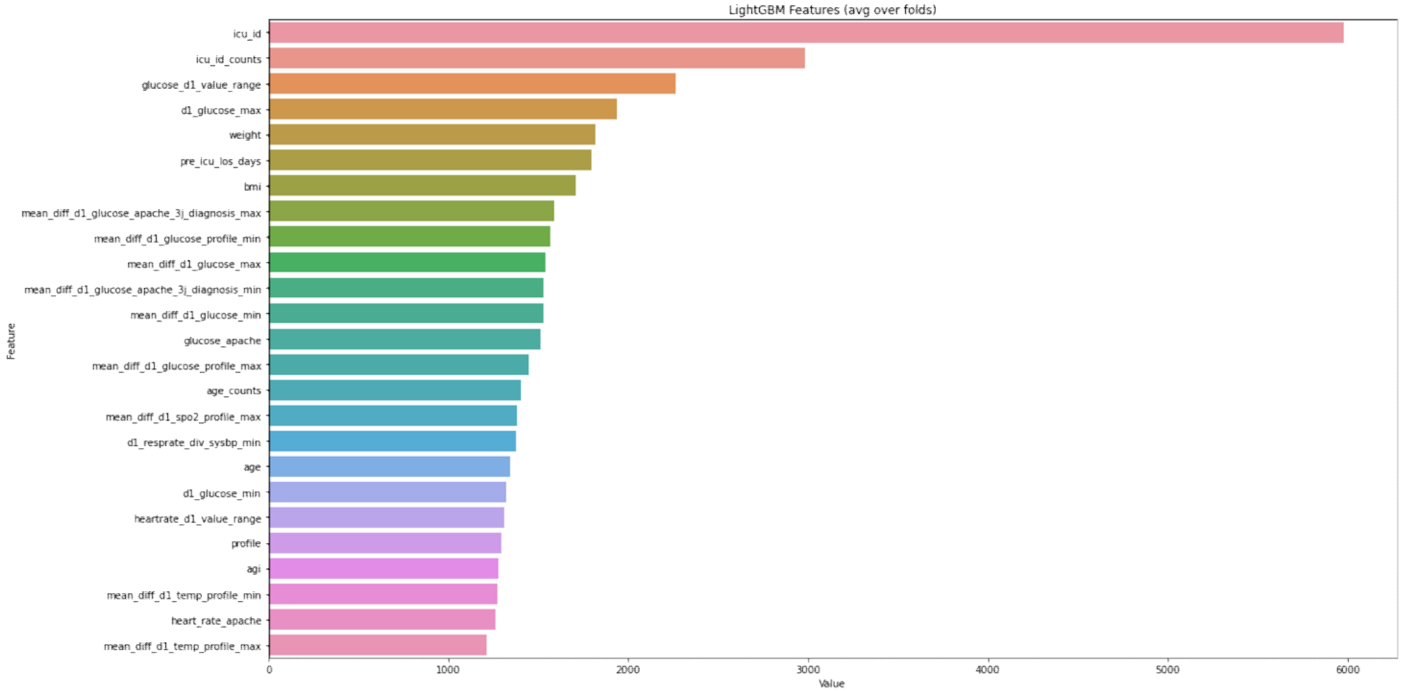
Result – AUC Scores

Model	Valid AUC	Test AUC
Ensemble of LightGBMs	**0.9860**	**0.8767**
XGBoost with Ensemble of SVM & KNN	0.8450	0.8572
StackNet with LightGBM& XGBoost	0.8766	0.8680
StackNet with CatBoost, LightGBM & XGBoost	0.8790	0.8683

ROC Curve – LightGBM Ensemble

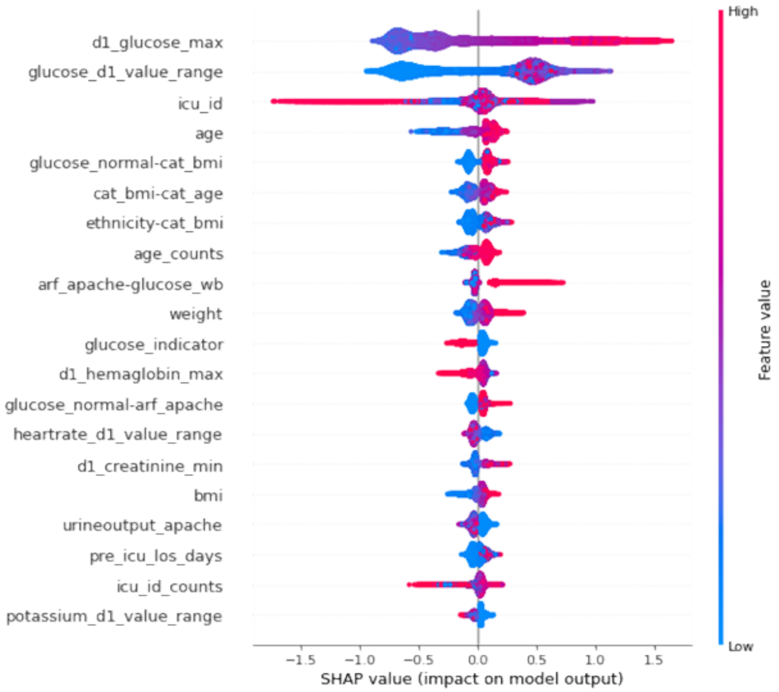


Feature Importance – LightGBM Ensemble



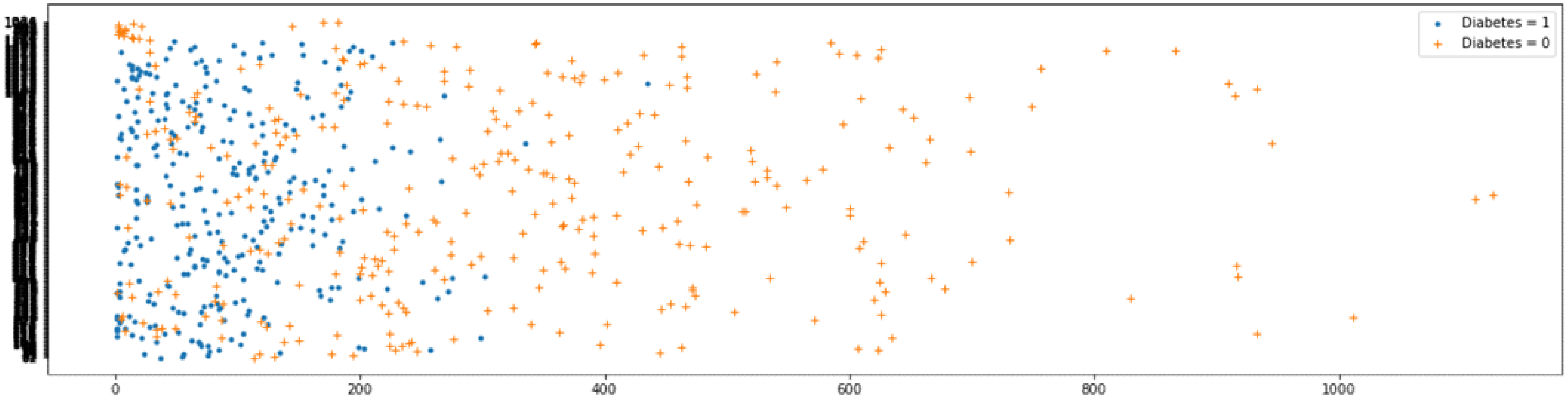
Model Explainability

SHAP Plot Analysis



ICU ID

- One reason ICU ID can be important is that different hospitals have different kinds of ICUs and equipment in the ICUs can vary as well. The ID can be an indicator of all such kind of information and thus without having more data related to the ICUs and hospitals in particular, it is hard to explain why it contributes so strongly to the prediction.
- The graph below shows the ICU IDs spread over the dataset.



Conclusion & Future Work

- We were able to predict diabetes mellitus for ICU patients with a decent AUC score.

- The results of the model were made explainable to understand how the model makes the decision.

- Having more information on hospitals and ICUs can make the model more robust.

- Deep learning models can be explored.

Questions?

References

[1] Wu, H., S. Yang, Z. Huang, J. He, and X. Wang (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked 10*, 100–107.

[2] Jashwanth Reddy, D., B. Mounika, S. Sindhu, T. Pranayteja Reddy, N. Sagar Reddy, G. Jyothsna Sri, K. Swaraja, K. Meenakshi, and P. Kora (2020). Predictive machine learning model for early detection and analysis of diabetes. *Materials Today: Proceedings.*

[3] Priyadarshini, R., N. Dash, and R. Mishra (2014). A novel approach to predict diabetes mellitus using modified extreme learning machine. In *2014 International Conference on Electronics and Communication Systems (ICECS)*, pp. 1–5.

[4] Vijayan, V. V. and C. Anjali (2015). Prediction and diagnosis of diabetes mellitus — a machine learning approach. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 122–127.

References

[5] Han, L., S. Luo, J. Yu, L. Pan, and S. Chen (2015). Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes. I*EEE Journal of Biomedical and Health Informatics* 19(2), 728–734.

[6] Jahangir, M., H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. In *2017 Intelligent Systems Conference (IntelliSys)*, pp. 722–728.

[7] Anand, A. and D. Shakti (2015). Prediction of diabetes based on personal lifestyle indicators. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, pp. 673–676

[8] Dinh, A., S. Miertschin, A. Young, and S. D. Mohanty (2019, Nov). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making 19(1)*, 211