RDS Project Report

Maham Arif, Wayne Zhu

Disclaimer: While this report with 47 pages might seem daunting, it's in fact only 5800 words including an appendix and citations and problem prompts. The real content word count is much closer to 5000, which is expected for a 10-page single-spaced paper. We tried our best to keep it as succinct as possible but the dataset has 186 columns so it's difficult.

1 Background: General Information about ADS

(a) What is the purpose of this ADS? What are its stated goals?

The purpose of this ADS is to use the first 24 hours of intensive care data to predict the patient's death probability in the ICU. The goal is that by generating the most accurate prediction of survivability, medical resources can be allocated in the most efficient way. Before this competition, the industry used the logistic regression predictor Apache IV which has an AUC score of 0.868. The machine learning ensemble solution by Seffi Cohen, with an AUC of 0.915, ranked 1st place out of all solutions[Coh+21]. It uses a StackNet-based ensemble method consisting of 42 different models including K-nearest neighbors, gradient boosting, random forest, neural network, and logistic regression. Due to some unresolved bugs in their GitHub repository, we weren't able to train the StackNet to completion. Thus, we used their earlier CatBoost classifier which is another ensemble network model with an accuracy of 90.7. The rest remains the same. A significant contribution noted by the authors in the paper, the researchers also tried to address discrimination based on gender, ethnicity, and age with test time augmentation techniques to both improve fairness and accuracy.

(b) If the ADS has multiple goals, explain any trade-offs that these goals may introduce.

From the impossibility result, we know that there will be a tradeoff between accuracy and fairness—especially when different protected groups' prevalence differs.

2 Input and Output

(a) Describe the data used by this ADS. How was this data collected or selected?

Released with a privacy certificate from Harvard Privacy Lab, this dataset is the MIT Global Open Source Severity of Illness Score (GOSSIS) dataset consisting of 131,051 hospital Intensive Care Unit (ICU) visits and outcomes, spanning a one-year timeframe in various countries. The data contains patient demographics data (such as age, gender, ethnicity, etc.), hospitalization

conditions, physiologic data, chronic health conditions (such as hepatic failure, immunosuppression, lymphoma, leukemia, etc.), pre-computed APACHE scores, and various medical measures, including the minimum and maximum lab test metrics for the first hour and first 24 hours of the ICU admission. Overall, the provided dataset included 185 features. The target variable to predict was the death likelihood in the ICU. The training set was unbalanced, with 7,915 deaths out of the total 91,713 admissions. 0 represents alive, while 1 represents death.

(b) For each input feature, describe its datatype, give information on missing values and on the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting and appropriate.

The data contains 186 features in total, that mainly fall into the following categories:

1. Identifiers (such as patient id, hospital id, etc.)
2. Demographics (various attributes about the patient such as gender, ethnicity, age, height, and weight)
3. Apache Covariate (a series of measures for patient's survivability)
4. Vital Measures
5. Lab Test Results (mostly numerical values)
6. Lab Blood Gas
7. Apache Comorbidity Diseases (AIDS, Diabetes, mostly of boolean datatype)

2.1 Datatypes and Missing Values

According to the paper, 50% of features have more than 50% of missing values. Features with a high percentage of missing values should be removed altogether.

2.1.1 Identifiers

The following table provides a summary of the data type, unit of measurement, the count of missing values, the count of unique values, the number of null values, and uniqueness for each feature that belongs to the identifier category in the data.

feature	datatype	unit	null_count	null_ratio	num_uniques	uniqueness
encounter_id	integer	None	0	0.0	91713	1.0
patient_id	integer	None	0	0.0	91713	1.0
hospital_id	integer	None	0	0.0	147	0.0

As shown in the table above, the features encounter_id and patient_id (both integers) have zero null values and uniqueness 1. This indicates that these two features are possible candidate

keys for the given data.

feature	min	max	mean	std
encounter_id	1.0	131051.0	65606.08	37795.09
patient_id	1.0	131051.0	65537.13	37811.25
hospital_id	2.0	204.0	105.67	62.85

2.1.2 Demographics

feature	datatype	null_count	null_ratio	num_uniques	uniqueness
hospital_death	binary	0	0.0	2	0.0
age	numeric	4228	0.05	74	0.0
bmi	string	3429	0.04	34888	0.38
elective_surgery	binary	0	0.0	2	0.0
ethnicity	string	1395	0.02	6	0.0
gender	string	25	0.0	2	0.0
height	numeric	1334	0.01	401	0.0
hospital_admit_source	string	21409	0.23	15	0.0
icu_admit_source	string	112	0.0	5	0.0
icu_id	integer	0	0.0	241	0.0
icu_stay_type	string	0	0.0	3	0.0
icu_type	string	0	0.0	8	0.0
pre_icu_los_days	numeric	0	0.0	9757	0.11
readmission_status	binary	0	0.0	1	0.0
weight	numeric	2720	0.03	3409	0.04

The table above provides a summary of the data type, the count of missing values, the count of unique values, the percentage of null values, and the uniqueness of each feature that belongs to the demographics category in the data. Following are some of the interesting parts of the statistics:

1. hospital_death: This is the target binary variable in the dataset, with 0 null values.
2. age: This column has a lot of null values, around 4228. It shows that various patients do not report their age in hospital records.
3. ethnicity: The data contains 6 races, namely Caucasian, Hispanic, African American, Asian, Native American, and Other. This column has a relatively less number of null values than age.
4. gender: The dataset contains two gender values males and females, and very few null values (only 25).

Elective surgery, ICU id, and ICU stay type are some of the features that have 0 null counts. Other demographics columns have some null values but the overall null ratio for each feature

is very low, below 0.3. The table below shows the minimum, maximum, mean, and standard deviation for these features.

feature	unit	min	max	mean	std
hospital_death	None	0.0	1.0	0.09	0.28
age	Years	16.0	89.0	62.31	16.78
bmi	kilograms/metres^2	14.84	67.81	29.19	8.28
elective_surgery	None	0.0	1.0	0.18	0.39
ethnicity	None				
gender	None				
height	centimetres	137.2	195.59	169.64	10.8
hospital_admit_source	None				
icu_admit_source	None				
icu_id	None	82.0	927.0	508.36	228.99
icu_stay_type	None				
icu_type	None				
pre_icu_los_days	Days	-24.95	159.09	0.84	2.49
readmission_status	None	0.0	0.0	0.0	0.0
weight	kilograms	38.6	186.0	84.03	25.01

The dataset contains people of almost all ages, but it doesn't include any data about children.

2.1.3 APACHE Covariates

The table below provides a summary of each feature that belongs to the APACHE covariate category. These features are the results of several measurements that provide an indication of the severity of the disease for adult patients admitted to intensive care units.

A lot of these features have a large number of null values. This is completely understandable as every measurement may not be relatable for every patient, depending on their specific disease. Only one feature, apache_post_operative in this category has a 0 null count. This is a binary attribute that indicates the APACHE operative status; 1 for post-operative, 0 for non-operative.

Following is the description of some other interesting features in the table above that have a low null ratio:

1. heart_rate_apache: It is a numeric feature that represents the heart rate measured during the first 24 hours.
2. temp_apache: It is also a numeric attribute. It represents the temperature measured during the first 24 hours.
3. ventilated_apache: This is a binary feature that represents whether the patient was invasively ventilated at the time of the highest scoring arterial blood gas using the oxygenation

scoring algorithm, including any mode of positive pressure ventilation delivered through a circuit attached to an endotracheal tube or tracheostomy.

4. resprate_apache: It is a numeric feature that represents the respiratory rate measured during the first 24 hours.

feature	datatype	null_count	null_ratio	num_uniques
albumin_apache	numeric	54379	0.59	35
apache_2_diagnosis	string	1662	0.02	44
apache_3j_diagnosis	string	1101	0.01	399
apache_post_operative	binary	0	0.0	2
arf_apache	binary	715	0.01	2
bilirubin_apache	numeric	58134	0.63	362
bun_apache	numeric	19262	0.21	476
creatinine_apache	numeric	18853	0.21	1127
fio2_apache	numeric	70868	0.77	82
gcs_eyes_apache	integer	1901	0.02	4
gcs_motor_apache	integer	1901	0.02	6
gcs_unable_apache	binary	1037	0.01	2
gcs_verbal_apache	integer	1901	0.02	5
glucose_apache	numeric	11036	0.12	565
heart_rate_apache	numeric	878	0.01	149
hematocrit_apache	numeric	19878	0.22	353
intubated_apache	binary	715	0.01	2
map_apache	numeric	994	0.01	161
paco2_apache	numeric	70868	0.77	704
paco2_for_ph_apache	numeric	70868	0.77	704
pao2_apache	numeric	70868	0.77	2003
ph_apache	numeric	70868	0.77	555
resprate_apache	numeric	1234	0.01	74
sodium_apache	numeric	18600	0.2	119
temp_apache	numeric	4108	0.04	191
urineoutput_apache	numeric	48998	0.53	24772
ventilated_apache	binary	715	0.01	2
wbc_apache	numeric	22012	0.24	3075

The table below shows the minimum, maximum, mean, and standard deviation for these features. Following are some of the interesting statistics from this table:

1. The minimum reported temperature for patients in Degrees Celsius is 32.1, whereas the maximum temperature is 39.7.

2. The normal respiration rates for an adult person at rest range from 12 to 16 breaths per

minute. However, the minimum value for resprate_apache is 4.0, whereas the maximum value is 60 breaths per minute.

3. A normal resting heart rate should be between 60 to 100 beats per minute, but it can vary from minute to minute. In the given dataset, the minimum reported heart rate is 30 beats per minute. On the other hand, the maximum reported heart rate is 178 beats per minute.

feature	unit	min	max	mean	std
albumin_apache	g/L	1.2	4.6	2.9	0.68
apache_2_diagnosis	None	101.0	308.0	185.4	86.05
apache_3j_diagnosis	None	0.01	2201.05	558.22	463.27
apache_post_operative	None	0.0	1.0	0.2	0.4
arf_apache	None	0.0	1.0	0.03	0.16
bilirubin_apache	micromol/L	0.1	51.0	1.15	2.17
bun_apache	mmol/L	4.0	127.0	25.83	20.67
creatinine_apache	micromol/L	0.3	11.18	1.48	1.53
fio2_apache	Fraction	0.21	1.0	0.6	0.26
gcs_eyes_apache	None	1.0	4.0	3.47	0.95
gcs_motor_apache	None	1.0	6.0	5.47	1.29
gcs_unable_apache	None	0.0	1.0	0.01	0.1
gcs_verbal_apache	None	1.0	5.0	3.99	1.56
glucose_apache	mmol/L	39.0	598.7	160.33	90.79
heart_rate_apache	Beats per minute	30.0	178.0	99.71	30.87
hematocrit_apache	Fraction	16.2	51.4	32.99	6.87
intubated_apache	None	0.0	1.0	0.15	0.36
map_apache	Millimetres of mercury	40.0	200.0	88.02	42.03
paco2_apache	Millimetres of mercury	18.0	95.0	42.18	12.38
paco2_for_ph_apache	Millimetres of mercury	18.0	95.0	42.18	12.38
pao2_apache	Millimetres of mercury	31.0	498.0	131.15	83.61
ph_apache	None	6.96	7.59	7.35	0.1
resprate_apache	Breaths per minute	4.0	60.0	25.81	15.11
sodium_apache	mmol/L	117.0	158.0	137.97	5.28
temp_apache	Degrees Celsius	32.1	39.7	36.41	0.83
urineoutput_apache	Millilitres	0.0	8716.67	1738.28	1448.16
ventilated_apache	None	0.0	1.0	0.33	0.47
wbc_apache	$10^9/L$	0.9	45.8	12.13	6.92

2.1.4 Vitals

Vitals-related features measure the body's most basic functions. Some of these features have a very low null count, indicating that these measurements are taken for every patient regardless

of their disease.

feature	datatype	null_count	null_ratio	num_uniques
d1_diasbp_invasive_max	numeric	67984	0.74	145
d1_diasbp_invasive_min	numeric	67984	0.74	85
d1_diasbp_max	numeric	165	0.0	120
d1_diasbp_min	numeric	165	0.0	78
d1_diasbp_noninvasive_max	numeric	1040	0.01	120
d1_diasbp_noninvasive_min	numeric	1040	0.01	78
d1_heartrate_max	numeric	145	0.0	120
d1_heartrate_min	numeric	145	0.0	154
d1_mbp_invasive_max	numeric	67777	0.74	285
d1_mbp_invasive_min	numeric	67777	0.74	118
d1_mbp_max	numeric	220	0.0	125
d1_mbp_min	numeric	220	0.0	91
d1_mbp_noninvasive_max	numeric	1479	0.02	122
d1_mbp_noninvasive_min	numeric	1479	0.02	91
d1_resprate_max	numeric	385	0.0	79
d1_resprate_min	numeric	385	0.0	55
d1_spo2_max	numeric	333	0.0	43
d1_spo2_min	numeric	333	0.0	101
d1_sysbp_invasive_max	numeric	67959	0.74	225
d1_sysbp_invasive_min	numeric	67959	0.74	163
d1_sysbp_max	numeric	159	0.0	143
d1_sysbp_min	numeric	159	0.0	120
d1_sysbp_noninvasive_max	numeric	1027	0.01	143
d1_sysbp_noninvasive_min	numeric	1027	0.01	120
d1_temp_max	numeric	2324	0.03	186
d1_temp_min	numeric	2324	0.03	209
h1_diasbp_invasive_max	numeric	74928	0.82	103
h1_diasbp_invasive_min	numeric	74928	0.82	86
h1_diasbp_max	numeric	3619	0.04	107
h1_diasbp_min	numeric	3619	0.04	92
h1_diasbp_noninvasive_max	numeric	7350	0.08	108
h1_diasbp_noninvasive_min	numeric	7350	0.08	93
h1_heartrate_max	numeric	2790	0.03	119
h1_heartrate_min	numeric	2790	0.03	109
h1_mbp_invasive_max	numeric	74844	0.82	247

The table below shows the minimum, maximum, mean, and standard deviation for these features.

feature	unit	min	max	mean	std
d1_diasbp_invasive_max	Millimetres of mercury	37.0	181.0	78.76	21.73
d1_diasbp_invasive_min	Millimetres of mercury	5.0	89.0	46.74	12.86
d1_diasbp_max	Millimetres of mercury	46.0	165.0	88.49	19.8
d1_diasbp_min	Millimetres of mercury	13.0	90.0	50.16	13.32
d1_diasbp_noninvasive_max	Millimetres of mercury	46.0	165.0	88.61	19.79
d1_diasbp_noninvasive_min	Millimetres of mercury	13.0	90.0	50.24	13.34
d1_heartrate_max	Beats per minute	58.0	177.0	103.0	22.02
d1_heartrate_min	Beats per minute	0.0	175.0	70.32	17.12
d1_mbp_invasive_max	Millimetres of mercury	38.0	322.0	114.89	49.45
d1_mbp_invasive_min	Millimetres of mercury	2.0	119.0	62.32	18.06
d1_mbp_max	Millimetres of mercury	60.0	184.0	104.65	20.81
d1_mbp_min	Millimetres of mercury	22.0	112.0	64.87	15.68
d1_mbp_noninvasive_max	Millimetres of mercury	60.0	181.0	104.59	20.7
d1_mbp_noninvasive_min	Millimetres of mercury	22.0	112.0	64.94	15.7
d1_resprate_max	Breaths per minute	14.0	92.0	28.88	10.7
d1_resprate_min	Breaths per minute	0.0	100.0	12.85	5.06
d1_spo2_max	Percentage	0.0	100.0	99.24	1.79
d1_spo2_min	Percentage	0.0	100.0	90.45	10.03
d1_sysbp_invasive_max	Millimetres of mercury	71.0	295.0	154.27	32.29
d1_sysbp_invasive_min	Millimetres of mercury	10.0	172.0	93.81	24.98
d1_sysbp_max	Millimetres of mercury	90.0	232.0	148.34	25.73
d1_sysbp_min	Millimetres of mercury	41.0	160.0	96.92	20.68
d1_sysbp_noninvasive_max	Millimetres of mercury	90.0	232.0	148.24	25.79
d1_sysbp_noninvasive_min	Millimetres of mercury	41.03	160.0	96.99	20.71
d1_temp_max	Degrees Celsius	35.1	39.9	37.28	0.69
d1_temp_min	Degrees Celsius	31.89	37.8	36.27	0.75
h1_diasbp_invasive_max	Millimetres of mercury	33.0	135.0	67.97	16.26
h1_diasbp_invasive_min	Millimetres of mercury	19.0	104.0	56.14	14.14
h1_diasbp_max	Millimetres of mercury	37.0	143.0	75.35	18.41
h1_diasbp_min	Millimetres of mercury	22.0	113.0	62.84	16.36
h1_diasbp_noninvasive_max	Millimetres of mercury	37.0	144.0	75.81	18.48
h1_diasbp_noninvasive_min	Millimetres of mercury	22.0	114.0	63.27	16.42
h1_heartrate_max	Beats per minute	46.0	164.0	92.23	21.82
h1_heartrate_min	Beats per minute	36.0	144.0	83.66	20.28
h1_mbp_invasive_max	Millimetres of mercury	35.62	293.38	94.88	30.81

Following is the description of some of the interesting features of vitals:

- d1_heartrate_max: This attribute represents the patient's highest heart rate during the first 24 hours of their unit stay.
- d1_mbp_max: This feature represents the patient's highest mean blood pressure during

the first 24 hours of their unit stay, either non-invasively or invasively measured.

- d1_resprate_max: It is a numerical feature that represents patients' highest respiratory rate during the first 24 hours of their unit stay.

2.1.5 Labs Statistics

These features are also measurements of body functions, similar to vitals. All of the features in this category are of a numeric data type. The null ratio for some of these features is very high, indicating that these are for specific diseases. The data profiling tables summarizing the are shown in the appendix for reference.

2.1.6 Lab Blood Gas Results

All of the features in this category are also of a numeric data type, similar to vitals. The null ratio for some of these features is very high, indicating that these are for specific diseases. The data profiling tables summarizing the are shown in the appendix for reference.

2.1.7 APACHE Comorbidity Diseases

The following tables provide a summary of basic profiling for each of the features that belong to this category. Each of these features is a binary attribute. It is interesting to note that all of them have the same null count.

feature	datatype	null_count	null_ratio	num_uniques
aids	binary	715	0.01	2
cirrhosis	binary	715	0.01	2
diabetes_mellitus	binary	715	0.01	2
hepatic_failure	binary	715	0.01	2
immunosuppression	binary	715	0.01	2
leukemia	binary	715	0.01	2
lymphoma	binary	715	0.01	2
solid_tumor_with_metastasis	binary	715	0.01	2

feature	unit	min	max	mean	std
aids	None	0.0	1.0	0.0	0.03
cirrhosis	None	0.0	1.0	0.02	0.12
diabetes_mellitus	None	0.0	1.0	0.23	0.42
hepatic_failure	None	0.0	1.0	0.01	0.11
immunosuppression	None	0.0	1.0	0.03	0.16
leukemia	None	0.0	1.0	0.01	0.08
lymphoma	None	0.0	1.0	0.0	0.06
solid_tumor_with_metastasis	None	0.0	1.0	0.02	0.14

2.2 Value Distributions

The dataset contains 186 features. Only interesting features are plotted below.

Moreover, the Kaggle competition provided separate training and test data sets for this ADS. However, their testing data is unlabelled and we cannot use it for our analysis. Therefore, we split the training data of the total 91,713 admissions, into 80/20 subsets for training and testing respectively. The value distribution results shown below, however, are for the whole dataset because we assumed that our random split didn't change the data distributions.

2.2.1 Hospital Death

This feature is a binary attribute that indicates whether the patient died during this hospitalization. Overwhelming majority of patients do not die.

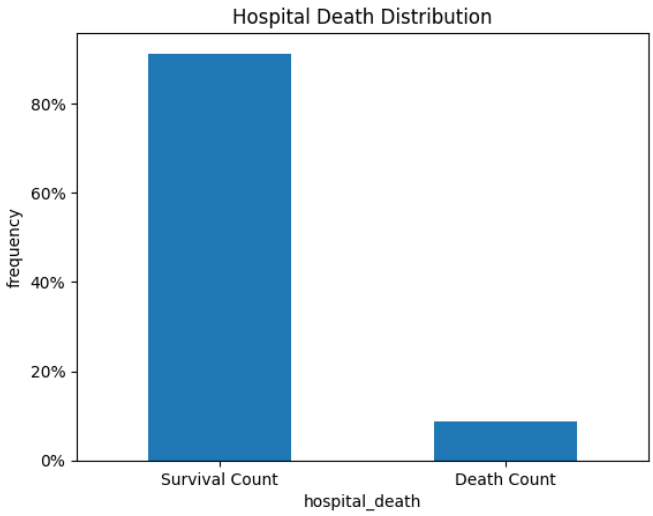

Figure 1: Distribution of Hospital Death

2.2.2 Gender

Figure 2 shows the gender distribution in the dataset, as well as the distribution of the target variable with respect to gender. The majority of patients are male. Moreover, the survival rate in males is slightly higher than in females. However, the difference is small enough to ignore.

2.2.3 Ethnicity

Figure 3 shows the distribution of ethnicity in the dataset, as well as the distribution of target variable hospital death with respect to race. As shown in the figure above, the overwhelming majority of the patients are Caucasian. The distribution of other races is very low in the dataset.

However, as evident in the graph, the death rate only slightly differs among races. Noticeably, the survival rate is the lowest, and the death rate is the highest for Hispanic people.

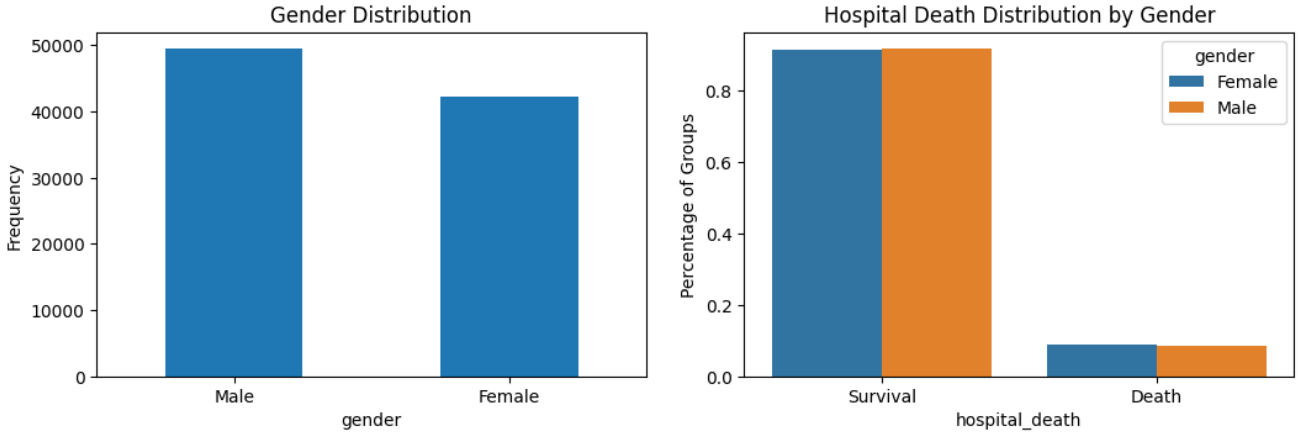

Figure 2: Gender Distribution

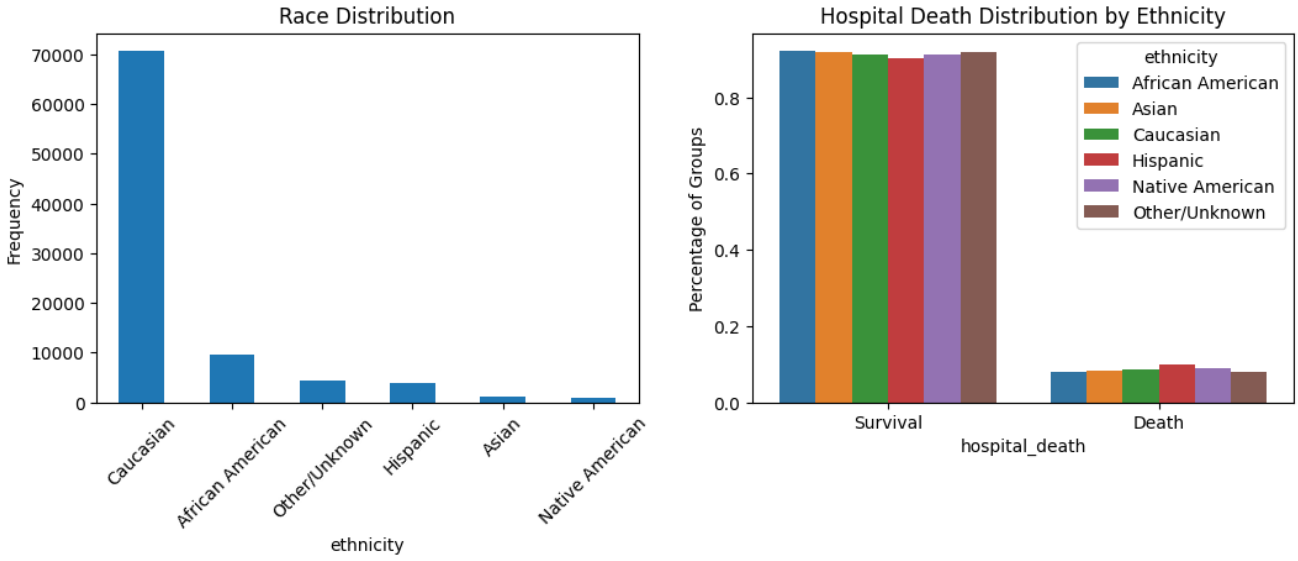

Figure 3: Race Distribution

2.2.4 Age

Figure 4 shows the value distribution of age, as well as the distribution of the target variable with respect to age. It is evident from the graph that age forms a left-skewed normal distribution with no kids. The distribution of the target variable clearly shows that the survival rate decreases with increasing age, which makes sense since lots of conditions are cumulative.

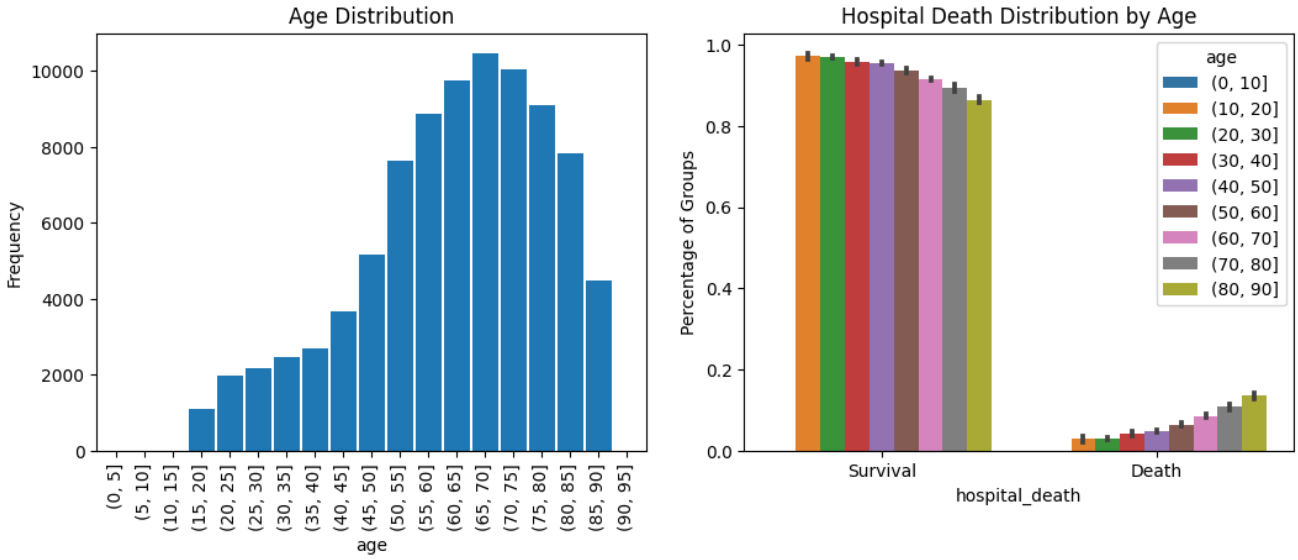

Figure 4: Age Distribution

2.2.5 BMI

This feature indicates the body mass index of the person on unit admission, measured in $kilograms/metres^2$. Figure 5 shows the distribution of BMI, as well as the distribution of hospital death with respect to BMI.

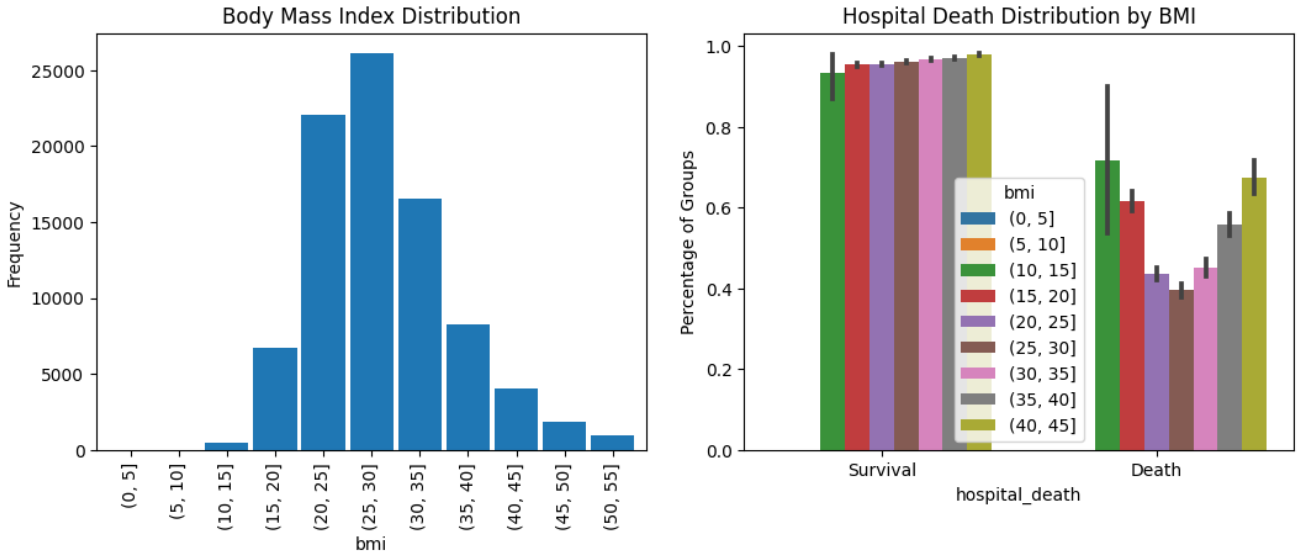

Figure 5: BMI Distribution

The BMI forms a slightly right-skewed normal distribution as no people have BMI below 10. The normal BMI for a person is 18.5 to 24.9, which falls within the Healthy Weight range. If the BMI is 25.0 to 29.9, it falls within the overweight range. Moreover, if the BMI is 30.0 or higher, it falls within the obese range. Most of the patients were overweight with a few underweight.

The survival rate increases slightly with increasing BMI. However, the death rate is higher for patients who fall into either the under-weight or over-weight category. The death rate is not correlated with height alone.

2.2.6 Heart Rate (beats per minute)

A normal resting heart rate should be between 60 to 100 beats per minute, but it can vary from minute to minute. As shown in the figure above, some patients do have heart rates below the normal range, which can be due to their specific diseases. The majority of patients have a heartbeat within the normal range, while a heart rate too low or too high is indicative of death.

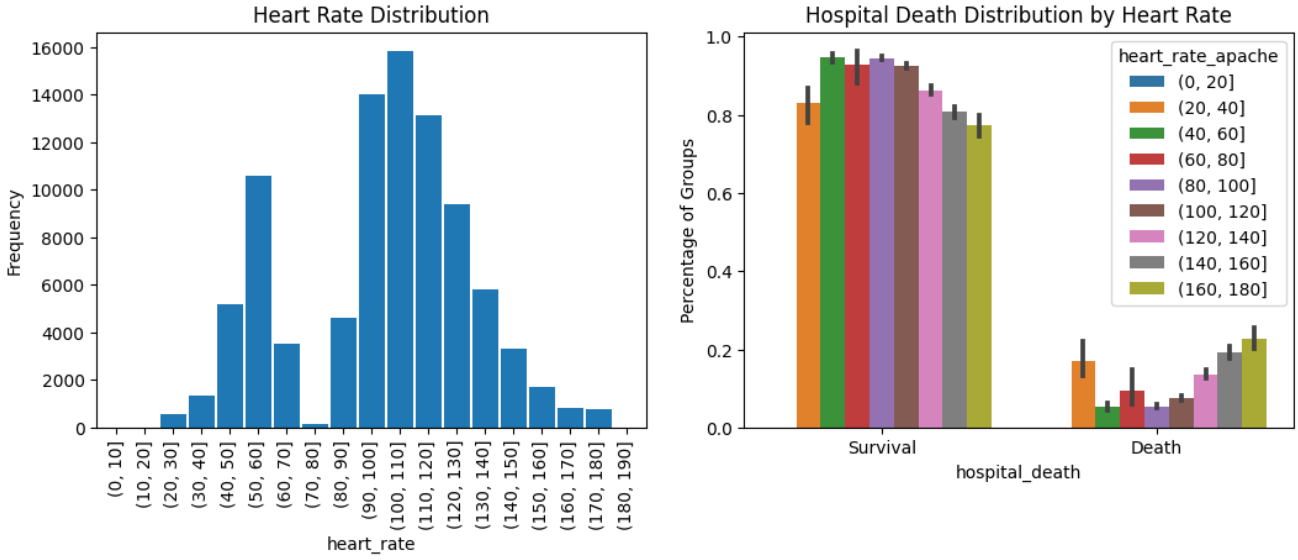

Figure 6: Heart Rate Distribution

2.2.7 Temperature (Celsius)

The death rate with respect to the patient's body temperature is shown in the appendix for reference. The survival rate is higher for normal temperature ranges, and the death rate is lowest for these ranges. This is understandable as the temperature of a person always rises with any disease.

2.2.8 Respiratory Rate (Breaths per minute)

The distribution of the respiratory rate of the patients, as well as the distribution of the target variable with respect to respiratory rate is shown in the appendix for reference.

2.3 Pairwise Correlations

For each of the 7 categories aforementioned, we plot the mutual-information heatmap. We took a sample of 10000 records for each heatmap. The explanations for some of the medical features can be found here:
`https://www.kaggle.com/code/jayjay75/wids2020-lgb-starter-adversarial-validation`

2.3.1 Identifiers

Not interesting. The pairwise correlation between attributes in the identifier category is shown in the appendix for reference.

2.3.2 Demographics

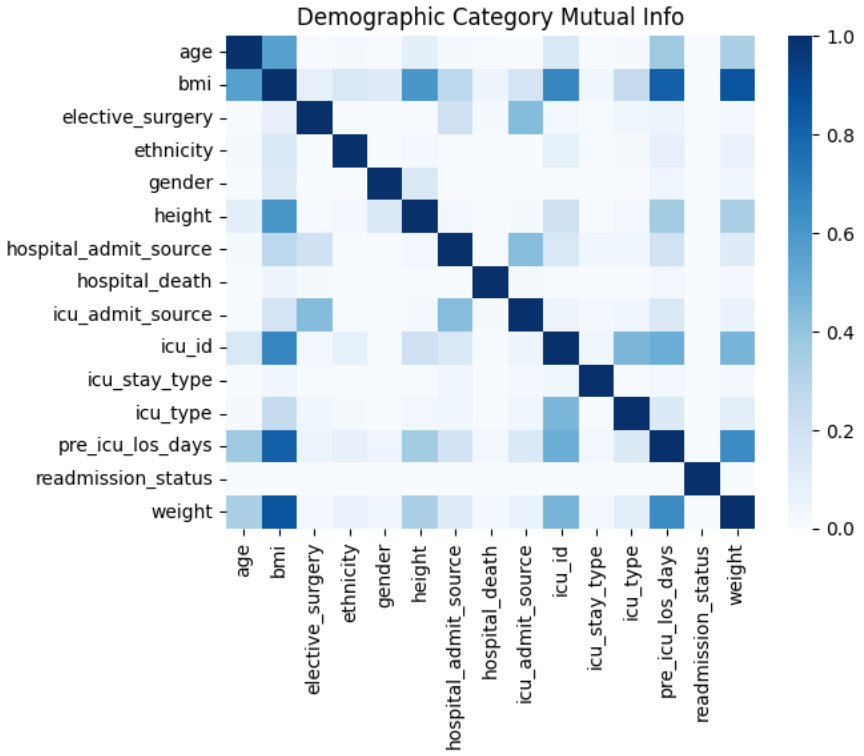

Figure 7: Pairwise Correlation between Demographics Attributes

As shown in the figure above, only weight, height, and BMI are strongly correlated to age as expected. Weight is also strongly correlated with days spent in ICU.

2.3.3 APACHE Covariate

Not all features are listed on the side due to space limitations. The GCS motor scores are correlated. Urine output and white blood cell counts are correlated. The paco scores measuring carbon dioxide and PH scores are correlated as well.

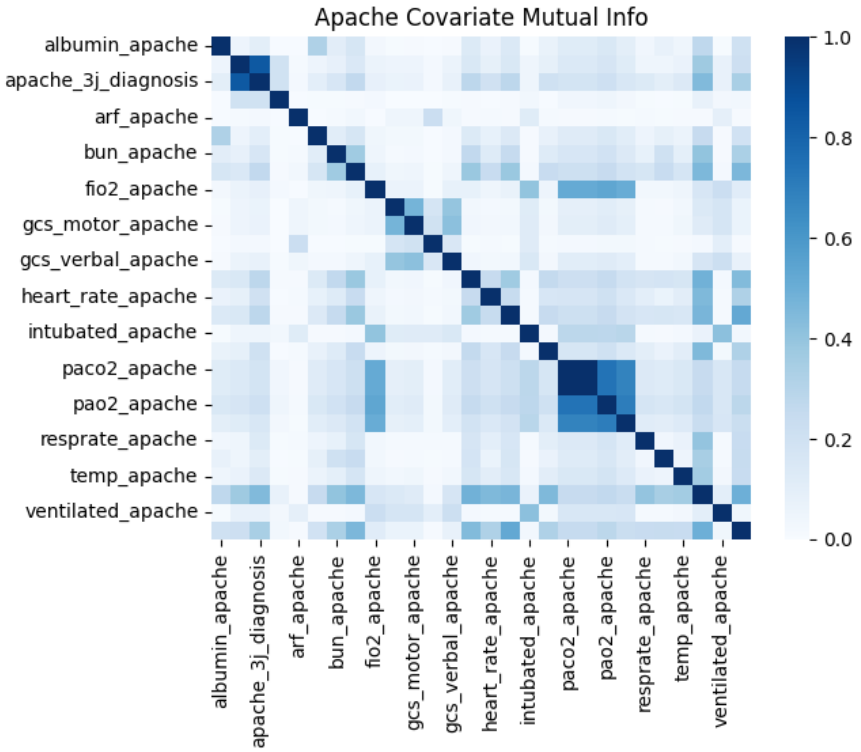

Figure 8: Pairwise Correlation between APACHE Covariate Attributes

2.3.4 Vitals

Figure 9 shows the pairwise correlation between attributes in the vitals category, but they are difficult to understand for people without domain knowledge. Most of the vital outcomes measuring different signs of the body are not correlated.

2.3.5 Lab Blood Gas

Figure 9 shows the pairwise correlation between attributes in the lab blood gas category. However, in contrast to the vitals, these features show the highest correlation between them.

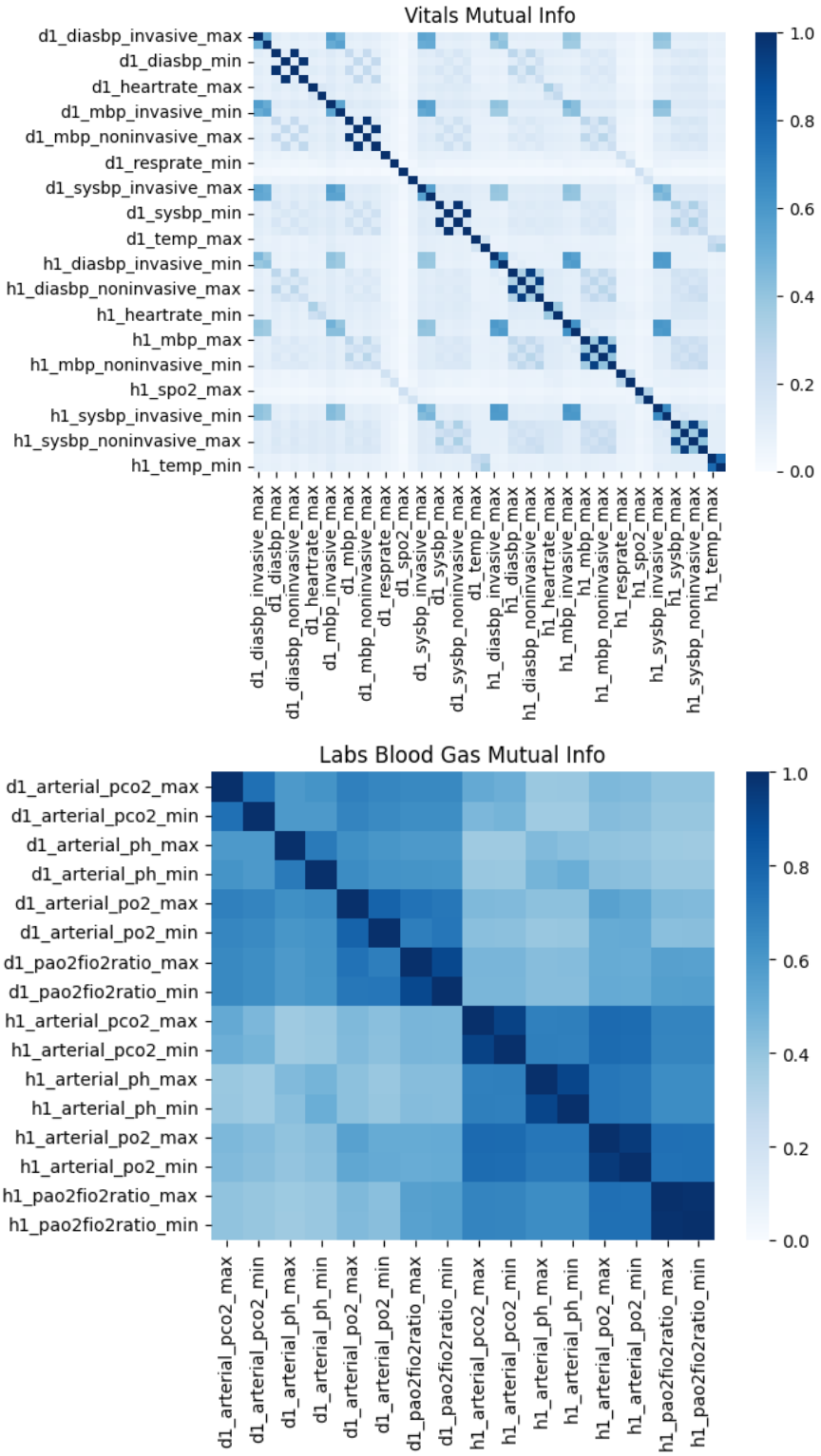

Figure 9: Pairwise Correlation between Vitals and Lab Blood Gas Statistics

2.3.6 Labs Statistics

The pairwise correlation between attributes in the labs' category is shown in the appendix for reference. These features are also measures of various body functions, similar to vitals. However, these measures do have various correlations between them. But again these are medical-related measures so these correlations are not very interesting.

2.3.7 APACHE Comorbidity

Each of these features is a binary attribute that indicates whether the patient has a history of that particular disease. All of these features seem to be correlated with each other except for diabetes mellitus.

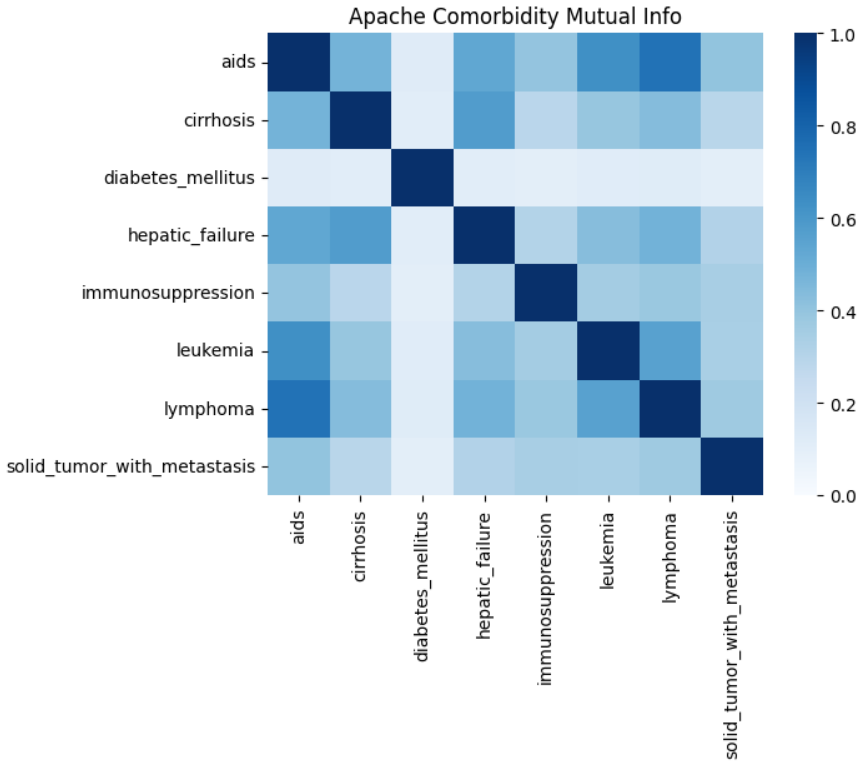

Figure 10: Pairwise Correlation between APACHE Comorbidity

c. What is the output of the system (e.g., is it a class label, a score, a probability, or some other type of output), and how do we interpret it?

The output is a likelihood of the 'hospital_death' attribute ranging from 0 to 1. If it's smaller than 0.5, then we predict 0 (survive). If it's larger than 0.5, we predict 1 (death).

3 Implementation and Validation

(a) Describe data cleaning and any other pre-processing.

More than 50% of the features had more than 50% missing values. The following methods are used by the ADS developers for pre-processing:

1. **Predictive Value Imputation:** This method uses linear regression to predict the feature using other features, e.g., height can be implied from weight, ethnicity, age, and gender. It also helped the developers to deduce other features using linear regression based on domain knowledge.

2. **Distribution-based Imputation:** This method uses mean or median according to the feature distribution to impute missing values. However, if a feature distribution is skewed, such as AIDS condition, the mode was used.

3. **Categorization:** Since missingness is often very predictive in medicine, researchers added a missing category bin when converting numerical variables to meaningful bins. Collectively, these preprocessing steps improve the AUC by 0.003.

They also removed features that were not useful via feature selection based on the following criteria: more than 80 percent of missing values, collinearity of 0.99, 0 standard deviation, 0 SHAP values, and adversarial validation.

(b) Give high-level information about the implementation of the system

On Kaggle, the team provides a simple model that uses a CatBoost classifier, which is a gradient-boosting decision tree algorithm that is particularly good for categorical data (which we had a lot in the dataset). They then performed test-time data augmentation (TAA) for producing the final mortality prediction for a record. That is, for each test record, they generated 3 additional records, each by changing either age, gender, or ethnicity. Then, a prediction was produced using the average of the 4 predictions (one for each additional record).

(c) How was the ADS validated? How do we know that it meets its stated goal(s)?

The experimenters compared their outcomes with multiple industrial benchmarks such as Apache IV, Deep Learning, and H2O AutoML (which involves the automatic training and tuning of various models). The paper's method won first place in the competition and was therefore under great scrutiny. It was also peer-reviewed and published in IEEE.

4 Outcomes

Before describing the outcomes of this ADS, let's formulate the problem and understand the stakeholder's perspectives:

1. **Patients:** Based on their likelihood of death prediction they will be assigned medical resources. Usually, the patient with the most critical condition aka higher likelihood of death will be assigned more treatment. But when resources are scarce such as in the early covid-19 pandemic, those patients would receive less treatment. In most cases, the patients would prefer a high likelihood of death so they can receive more care.
2. **Hospital/Medical Staff:** The hospital medical staffs are more interested in balancing resources and quality of care. When there are ample nurses and medicine, the hospital would willingly care for patients with a higher likelihood of death. This disadvantage does not necessarily harm patients with a lower likelihood of death. When the resources are low, the hospital would prioritize patients with a lower likelihood of death. Thus, this harms patients closer to death.

In our analysis, we have compared the ADS outcomes before and after this test time augmentation to evaluate whether that actually reduces bias.

Some protected attributes like age have a biological influence on a patient's survival that is not the result of discrimination. Further research suggests that lack of proper training and pressure to restrict healthcare costs and limited resources are primary barriers to the survival rate of elders who consume most of the healthcare ([MA97; WSB18]). These conditions do not apply to the pre-covid relatively developed countries where nurses receive proper training and where most of the dataset is collected. Even in a developing country like Egypt where there is less training on caring for the elderly, researchers survey nurses' ageist attitude toward elderly patient and found that the majority of this health-aware group do not demonstrate any negative bias against the elderly([EES20]). Therefore, we focus our analysis on gender and ethnicity.

We analyzed the accuracy and fairness of the ADS across the following subpopulations:

1. Gender (Males/Females)
2. Ethnicity (White/Black/Asian/Other)
3. Ethnicity and Gender Intersection

(a) Analyze the accuracy of the ADS by comparing its performance across different subpopulations, with respect to different accuracy metrics. Carefully justify your choice of accuracy metrics.

4.1 ADS Accuracy (Overall)

To analyze the accuracy of the ADS, we used the following three metrics:

1. **Accuracy:** Increased accuracy benefits patients and hospitals in non-emergency situations because the medical resources can be distributed to those in need more efficiently.
2. **Precision/PPV:** This metric measures the number of people that were actually dead for those predicted dead. High precision benefits hospitals but don't necessarily benefit patients.
3. **Recall:** This metric measures the fraction of dead patients being correctly predicted among all dead patients. A high recall benefits patients.
4. **FPR/FNR:** Analyzed when meaningful or necessary.

Due to the imbalanced data, the baseline accuracy (a model that only predicts survivability) is around **91%.**

```
General baseline accuracy: 0.9116
```

The following are the overall accuracy metrics of the ADS **(before test time data augmentation)**:

```
Model Accuracy (Overall): 0.9274
Model Precision (Overall): 0.7992
Model Recall (Overall): 0.2393
Model FNR (Overall): 0.7607
Model FPR (Overall): 0.0058
```

The ADS slightly outperformed the baseline model by 1%. This is good considering that the winning predictor has an AUC of 0.916. The precision of the model is 80%—the rate that the patient actually dies given the positive outcome by the model.

Due to the imbalanced dataset and high precision, the false positive rate of the model is at 0.56%. The false negative rate of the model is very high—at 0.76. A high FNR can be particularly concerning for hospitals since they will have difficulty identifying the patients in need.

Following are the overall accuracy metrics after test time augmentation:

```
Model Accuracy (Overall): 0.9277
Model Precision (Overall): 0.8038
Model Recall (Overall): 0.2412
Model FNR (Overall): 0.7588
Model FPR (Overall): 0.0057
```

For all the parts below, the test time augmentation technique slightly improves the accuracy and fairness of the model, in accordance with the paper published by Cohen. However, it does not result in any major improvement in the overall accuracy and fairness of the model. So we only display the results after test time augmentation below and left results before test time augmentation in the appendix.

Figure 11 shows the reliability diagram for the ADS, using the sklearn library. It compares how well the probabilistic predictions adhere to the actual situation after applying the test time augmentation technique. The ADS generally makes good predictions.

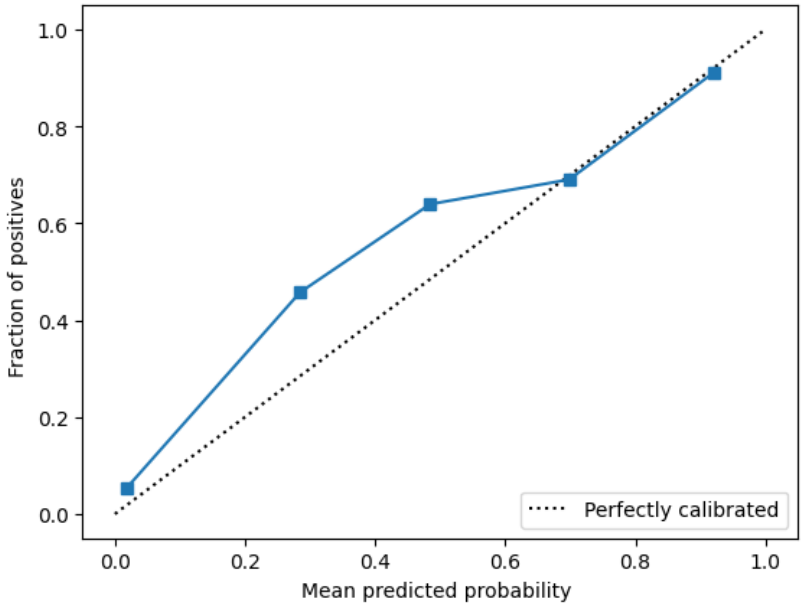

Figure 11: Calibration Curve for ADS

4.2 ADS Accuracy across Genders

Figure 12 shows the actual and predicted death count by gender. The majority of the dataset consists of males. However, the death rate is almost the same across males and females in reality as in appendix Figure 30. Thus, we are hoping to see fair outcomes on gender unless the ADS introduces some technical bias.

Figure 13 shows the accuracy metrics breakdown by gender after the test time augmentation. As shown in the figure, the accuracy is almost the same for both males and females. However, the precision and recall of the ADS are slightly higher for females than males. The result is that the predicted death rate of females is slightly higher than males, which disadvantages males slightly.

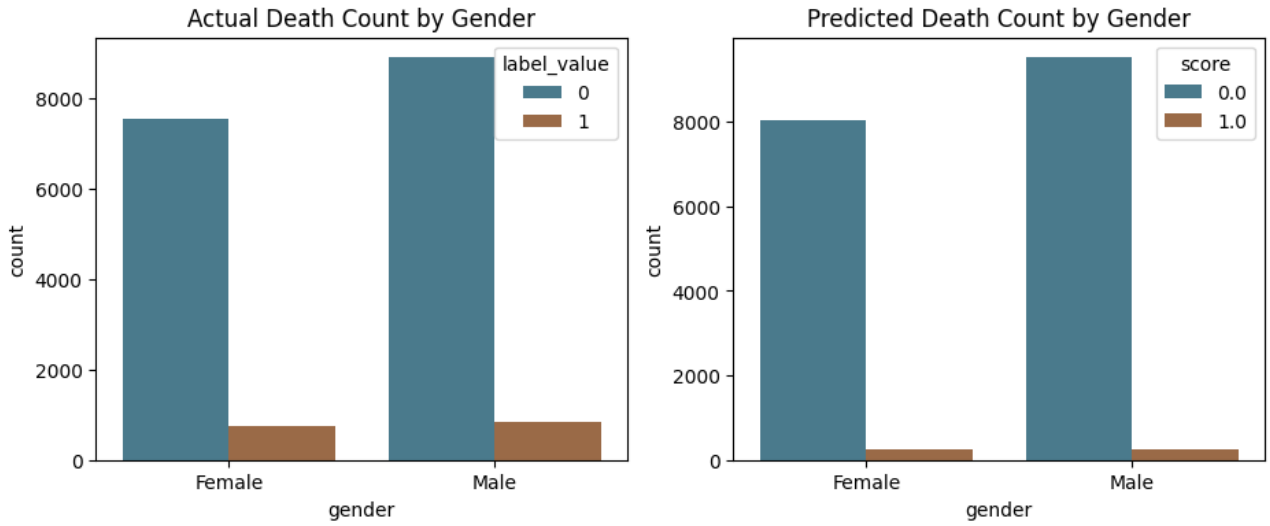

Figure 12: Death Count by Gender (Actual and Predicted)

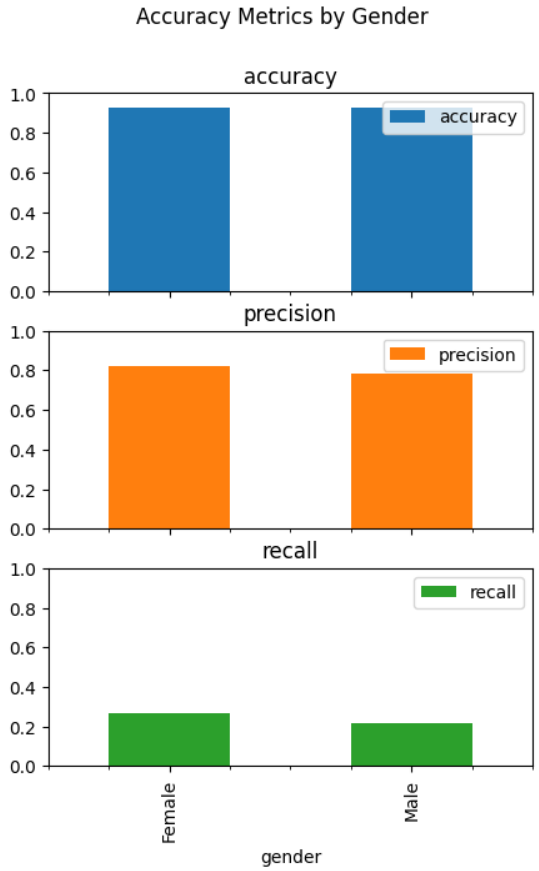

Figure 13: Accuracy Metrics By Gender (After Test Time Augmentation)

4.3 ADS Accuracy across Different Ethnicities

Again, here the test time data augmentation did improve the fairness metrics by a mostly negligible amount—less than 1 percent. It did bring Hispanics' data closer to other groups though by 1 to 2 %.

The ground truth data in Figure 4 shows that every other group except for Hispanic shows similar survival and death rate. This might not be the result of discrimination though, since the dataset contains large developing countries with a Hispanic majority like Brazil. The difference can be most likely accounted to global inequality.

Figure 16 shows the accuracy metrics breakdown by ethnicity after the test time augmentation.

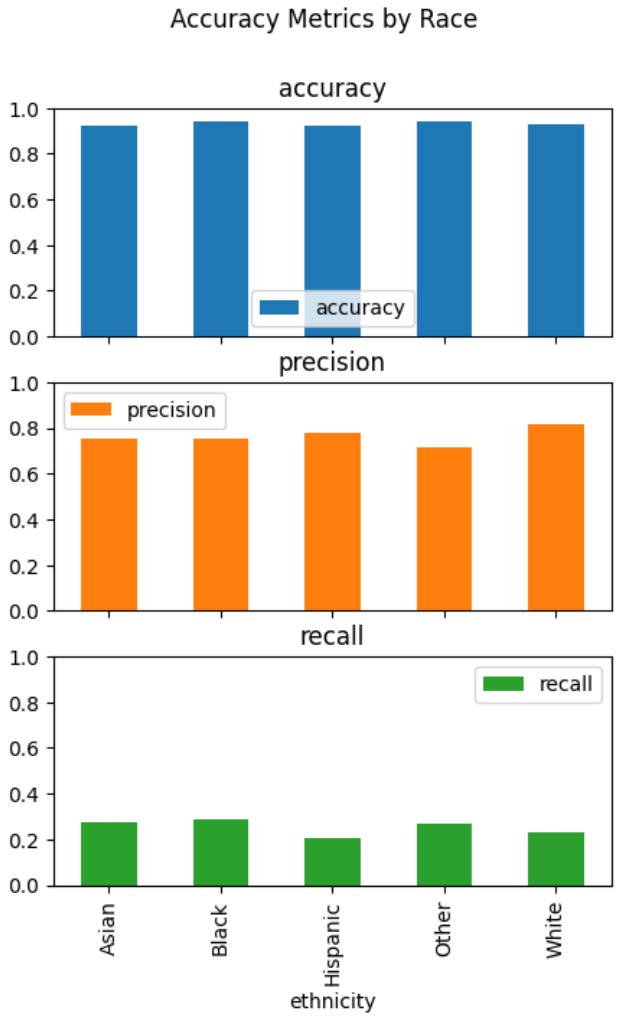

Figure 14: Metrics by Ethnicity

As shown in the figure, the accuracy is almost the same for all races. The disparities in precision and recall are however noticeable to some extent. White and Hispanics' precision is noticeably higher than other races while their recall is noticeably lower. Higher recall and lower precision mean that our model is more cautious at labeling white and Hispanics as dead compared to other races. As with gender-based analysis, the model has a high tendency for predicting false negatives and a very low false positive rate.

4.4 ADS Accuracy across Subpopulations based on Intersection of Ethnicity and Gender

When we break down gender and race by intersectionality, the imbalance of the dataset worsens and might further result in unstable predictions for minorities.

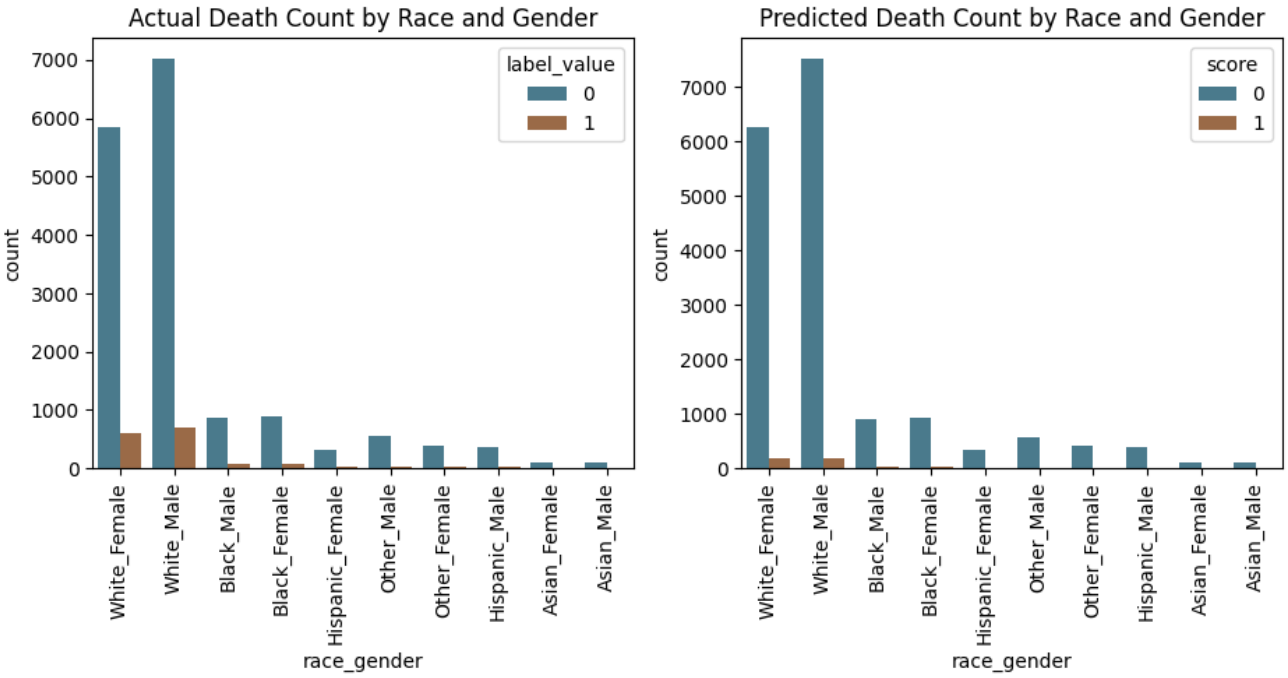

Figure 15: Actual and Predicted Death Count by Race and Gender

Figure 16 shows the accuracy metrics for each intersection group. As shown in the figure, the accuracy is similar across all groups. However, the discrepancies in precision are very noticeable—especially among minorities. The Asian male group has the highest precision, whereas the Asian female group has the lowest. The actual death rate, however, is almost similar for both groups. After Asian males, white females have the highest precision. Similarly, recall is lowest for Hispanic males but highest for Black females.

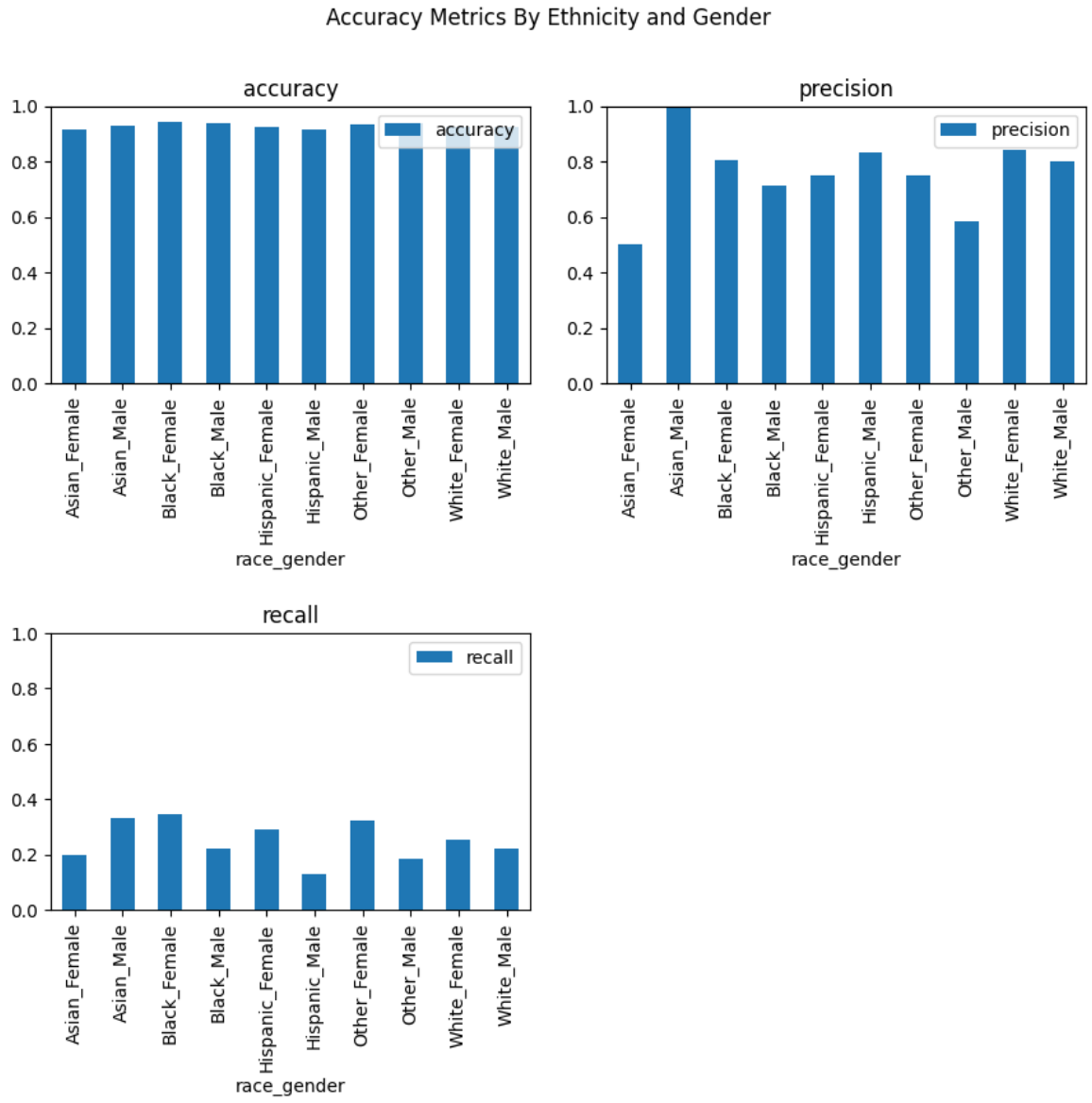

Figure 16: Accuracy Metrics By Ethnicity and Gender

4.5 Conclusion for ADS Accuracy

Gender-wise, the ADS seems to make more correct predictions for females and predicts more female death in proportion. This doesn't necessarily mean discrimination against males but would perhaps be indicative of different behavior patterns of men and women.

Ethnicity-wise, the model demonstrates similar accuracy but is less likely to predict White and Hispanics as dead. This could result in relatively more medical resources being allocated to

whites and Hispanics.

The model shows high variability when it comes to small intersectional groups of minorities. This is hard to avoid as the overwhelming majority of data are White. Still, there are minority groups that are clearly disadvantaged. The reality that the precision and recall of Asian females is only half of that of Asian males is a matter of concern whether there is fundamental discrimination on the ground.

(b) Analyze the fairness of the ADS, with respect to different fairness metrics. Carefully justify your choice of fairness metrics.

4.6 ADS Fairness

Assuming a situation with sufficient medical resources, high FNR hurts patients while high FPR deprives other truly dying patients of resources. We measure the following fairness metrics for each demographic group:

1. Demographic Parity Difference/Selection Rate Difference: Measures the difference between selection rates between groups.
2. Demographic Parity Ratio: Ratio of selection rates between groups.
3. FNR Difference: Measures the difference of false negative rate between groups.
4. FPR Difference: Measures the difference of false positive rate between groups.
5. Selection Rate: Measures the rate of being positive.
6. Equalized Odds Ratio: Stricter than demographic parity because it also takes into account equal FPR and FNR.

4.7 ADS Fairness across Genders

Figure 17 shows the FPR, FNR, and selection rate after test time augmentation for both subgroups. The model has 0 FPR for both males and females. The selection rate is almost the same for both groups. However, FNR for males is slightly higher than for females, which agrees with our earlier observation of females having higher precision and recall.

Following are the values of other fairness metrics based on gender as the sensitive feature:

```
Demographic parity difference (Gender): 0.0049
Selection rate difference (Gender): 0.0049
Demographic parity ratio (Gender): 0.8324
False negative rate difference (Gender): 0.0492
False positive rate difference (Gender): 0.0000
Equalized odds ratio (Gender): 0.8162
```

As shown above, the demographic parity difference, FPR difference, and difference are negligible. Males have 5% higher FNR and a lower selection rate. Also, the demographic parity ratio and equalized odds ratio further shows that the selection rate of men is only 80% that for women.

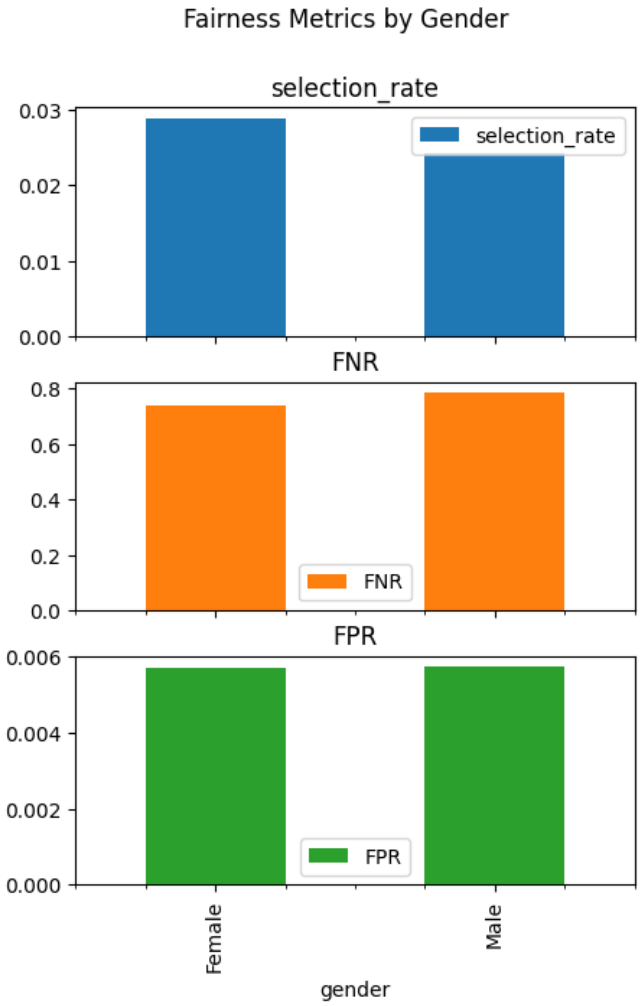

Figure 17: Fairness Metrics By Gender (After Test Time Augmentation)

4.8 ADS Fairness across Ethnicities

Figure 18 shows the FPR, FNR, and selection rate after test time augmentation for subgroups based on ethnicity. FPR for all races is almost smaller than 1%. The selection rate is low and similar for all subgroups based on ethnicity.

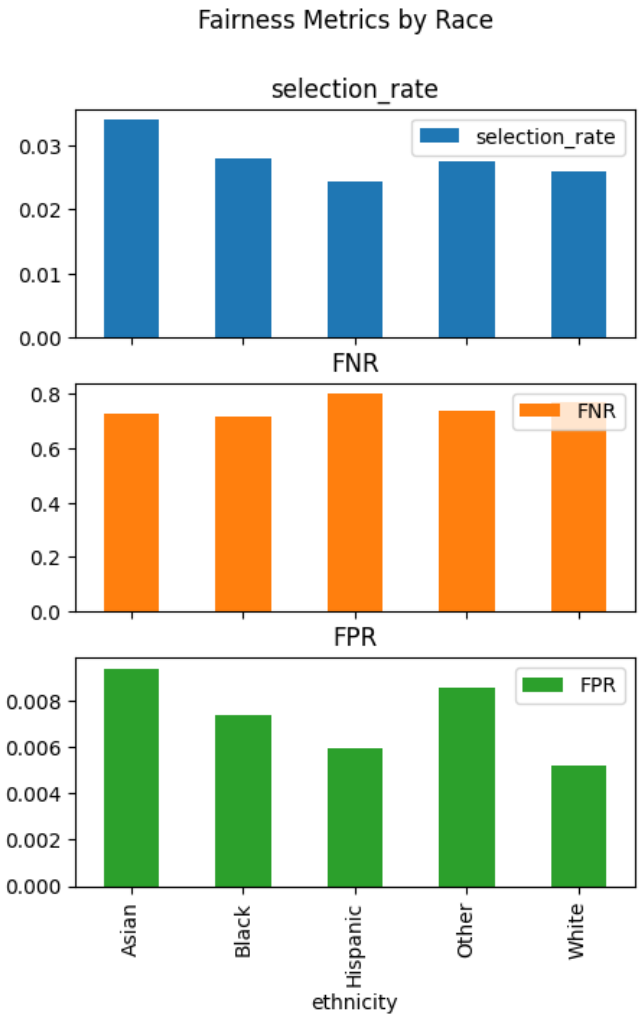

Figure 18: Fairness Metrics By Ethnicity (After Test Time Augmentation)

Following are the values of other fairness metrics based on the largest ethnicity difference:

```
Demographic parity difference (Ethnicity): 0.0100
Demographic parity ratio (Ethnicity): 0.7189
False negative rate difference (Ethnicity): 0.0903
False positive rate difference (Ethnicity): 0.0047
Equalized odds ratio (Ethnicity): 0.5273
Selection rate difference (Ethnicity): 0.0100
```

The demographic parity difference and selection rate difference are very low. However, the FNR difference is around 0.09, which is between white and black. It's of a large magnitude but the difference as a percentage is small. However, FPR as a difference of percentage is very large. For example, Asians and others are almost 50% more likely to be predicted as dead compared

to white or Hispanics. The equalized odds ratio is also very low. This supports our earlier arguments that white are disadvantaged.

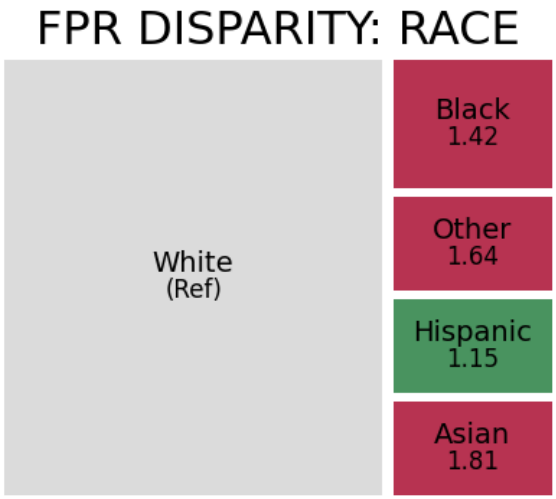

Figure 19: FPR Race Disparity

4.9 ADS Fairness across Subpopulations based on Intersection of Ethnicity and Gender

Figure 20 shows the FPR, FNR, and selection rate after test time augmentation for each intersection group based on ethnicity and gender. As shown in the figure, FPR for Asian males is 0. FPR for Hispanic males is very low. The rest of the groups (except Asian females) have slightly higher FPR, but almost the same. The Asian female group has the highest FPR among other groups and a high FNR. The least FNR is for Black females. Both groups have very little data. Following are the values of other fairness metrics based on ethnicity and gender as the sensitive feature:

```
Demographic parity difference (Race and Gender): 0.0163
Demographic parity ratio (Race and Gender): 0.5543
False negative rate difference (Race and Gender): 0.1630
False positive rate difference (Race and Gender): 0.0182
Selection rate difference (Race and Gender): 0.0163
```

The demographic parity difference and selection rate difference are low, but the FNR difference is high. For example, Hispanic males are 16% more likely to be falsely labeled as negative than black females. The selection rate difference is small in absolute terms but large in relative terms —which most likely comes from the difference between Other females and Hispanic males. Perhaps due to a lack of data, Asian females have an FPR of 3.27 that of white while for Asian males that rate is 0.

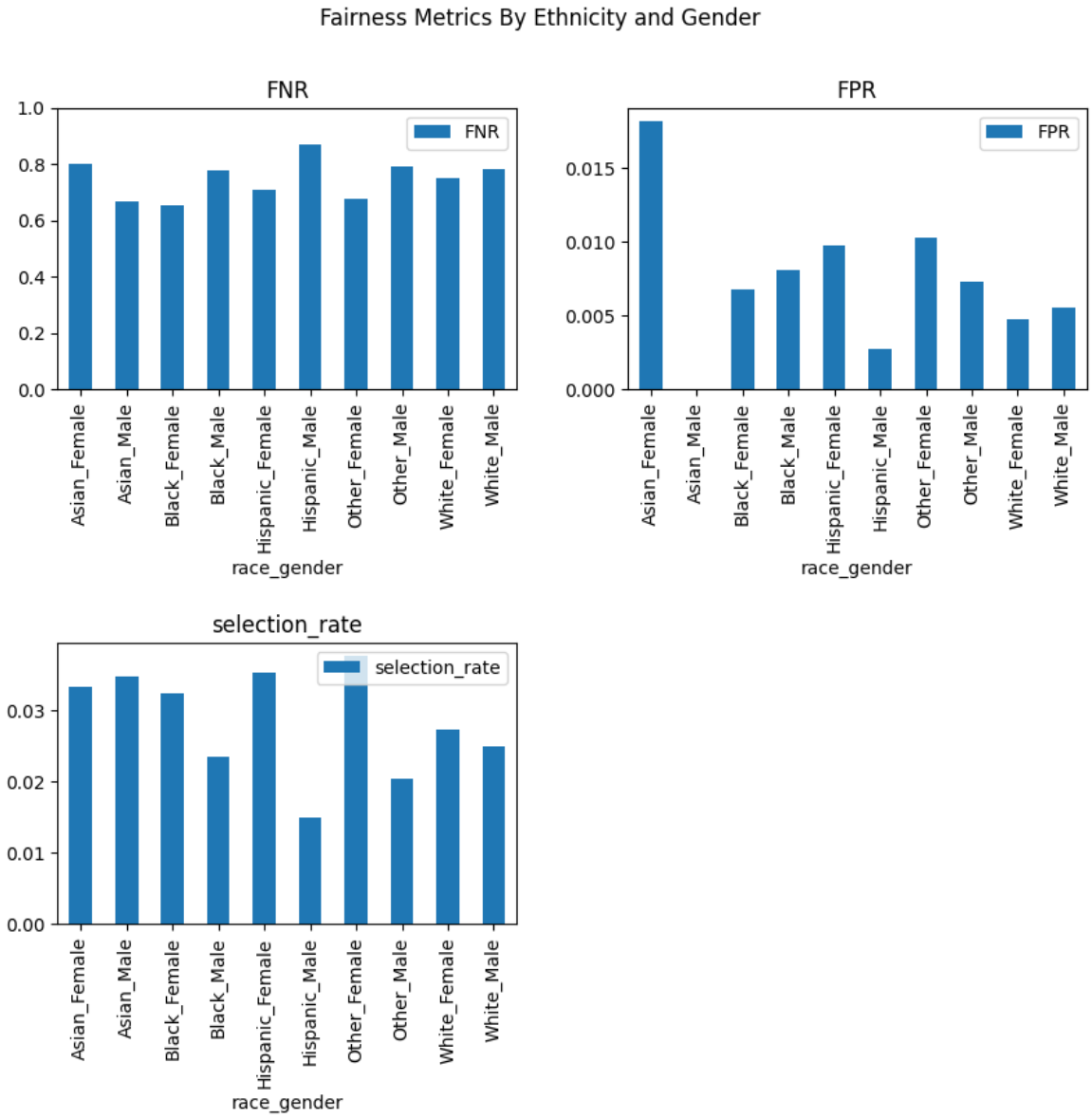

Figure 20: Fairness Metrics By Ethnicity and Gender (After Test Time Augmentation)

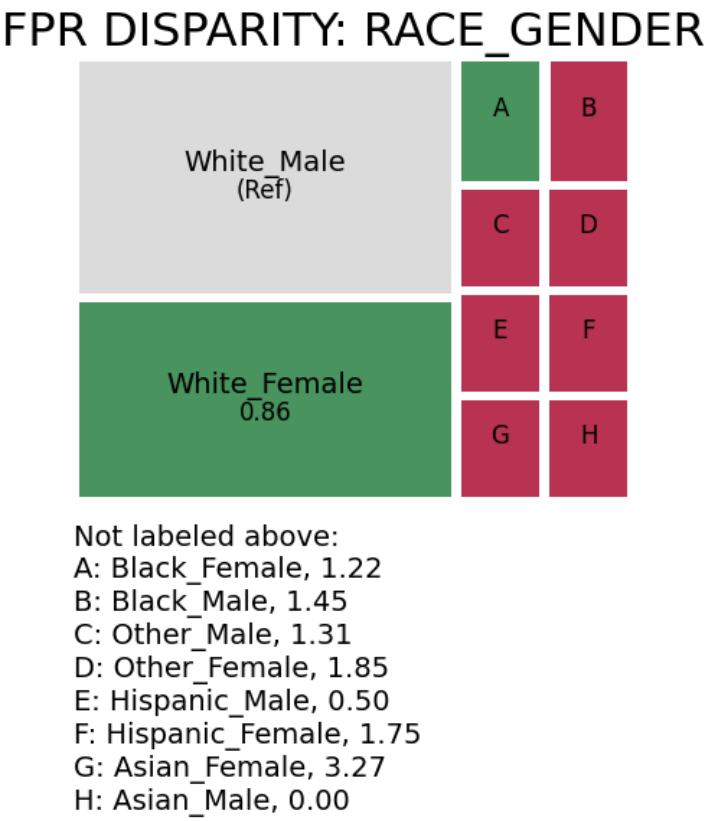

Figure 21: FPR Race Disparity

4.10 ADS Fairness Summary

Gender-wise, the model is slightly biased toward women but mostly fair.

Ethnicity-wise, the model shows fluctuating behavior toward Asians, and systematically under-predicts the death of Hispanics and whites. It's unclear whether imbalanced data, lack of data for minorities like Asians, and inequality between countries—which is immense, are the primary reason.

Intersection-wise, although an absolute pattern between races seems to be hard to draw, it's clear that the tool is especially unfair against Hispanic, white, and Asian males by assigning them very low FPR. The very high demographic parity difference and equalized odds difference suggest unfair prediction.

(c) Develop additional methods to analyze ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME or SHAP), or any other property that you believe is important to check for this ADS. Carefully justify your methodology.

4.11 Explainability of this ADS

First, we can compare the Shapley value of the most important attributes in the train and test set. Since the training and testing SHAP values are very similar, the top 20 Shapley values attribute from the testing set is displayed below:

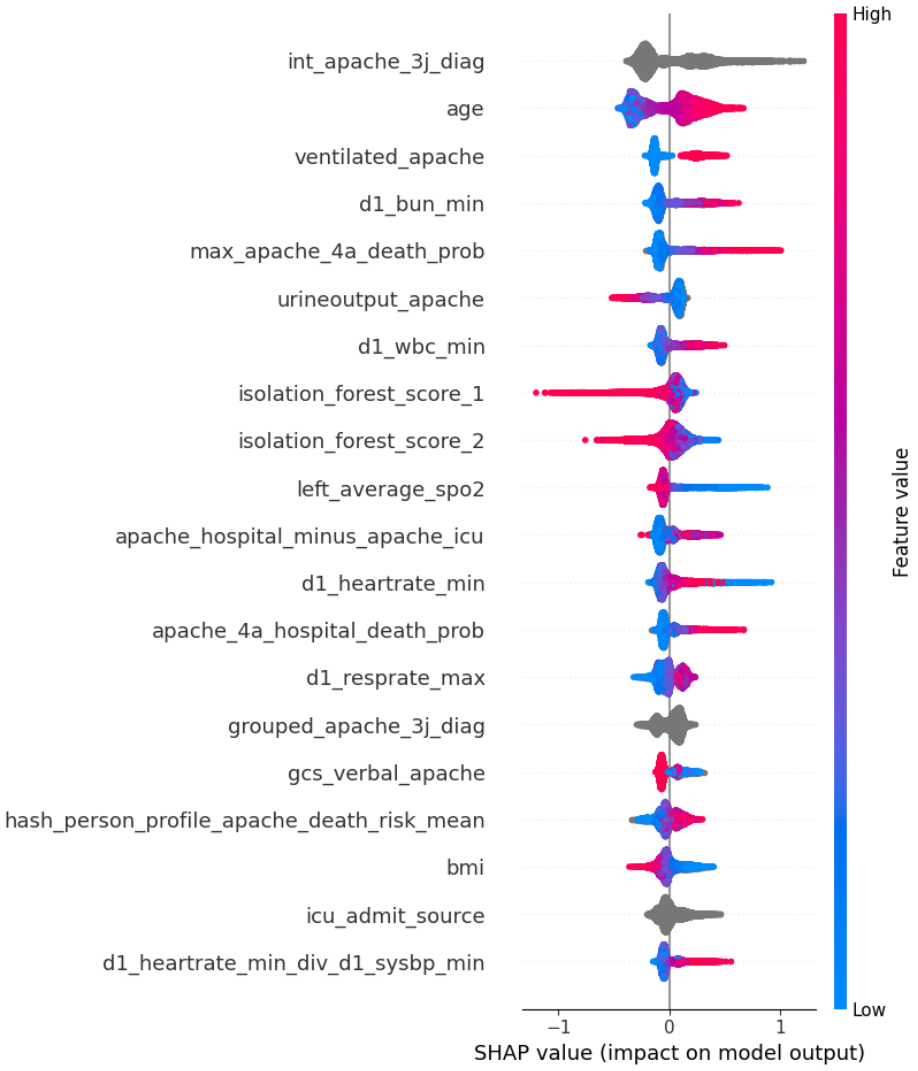

Figure 22: Shapley Value for Testing for CatBoost model

Interestingly, the same team led by Cohen that produced the Stack-Net also conducted their own Shapley value analysis and yielded quite a similar result (See Appendix Figure 34). The team explained some of these important features in their paper, and other data profiling has been done extensively on Kaggle. First, both the Catboost model and the StackNet model from those developers agree that the following attributes are the most important predictors of death:

1. ***ventilated apache*** indicating whether the person was mechanically ventilated.
2. ***age*** which directly correlates to health.
3. ***apache 3j diagnosis*** which categorizes the type of condition based on a mixture of other attributes.

Heart rate, Glasgow Coma Scale (GCS) which measures a person's consciousness, and **d1_spo2_min** which measures a person's minimal oxygen saturation also show up in the top 20 of both algorithms. **Urine output, WBC (white blood cell), BMI, and bun**(the lowest blood urea nitrogen concentration of the patient in their serum or plasma) are also in the top 20. Most of these factors are also acknowledged by the domain experts as good indicators of survivability themselves [Coh+21].

There are also some ICU admission examples that show the importance of these attributes via their Shap values. For example, here the most important attributes ranked roughly similar to the above-listed attributes starting with age, bun_min, and mechanical ventilation.

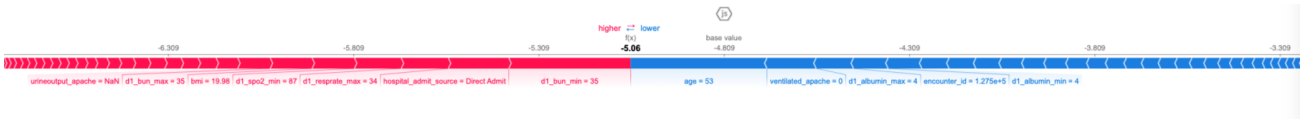

Figure 23: Shapley Explanation for One Patient in the test set.

Due to the project length limit, variability, and stability are covered earlier without running extra explicit experiments here.

5 Summary

5.1 Do you believe that the data was appropriate for this ADS?

I believe the data is somewhat appropriate for this ADS after cleaning because it contains primarily objective medical measures from the real world. Caveat: different countries' level of development varies by a lot, and ethnicity is correlated with country. This might be a confounding variable. For example, I certainly wouldn't expect Sri Lanka and Brazil to provide the same level of healthcare to Asians and Hispanics as those provided by hospitals in New

Zealand or the U.S. Hospital ID might be indicative of country but even hospitals within each country can vary a lot by quality of care. Hospital ID and ICU ID are in the dataset but are treated as useless columns by the author and removed. That's why when the macroscopic international development gap is not taken into account, it's hard to measure discrimination based on ethnicity and gender.

5.2 Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.

The implementation is good regarding accuracy. Even a simple CatBoost model can achieve an accuracy of 92.6%, higher than the baseline of 91%. The hospital would be happy to use it instead of the Apache index. But due to the original data being extremely imbalanced—death rate being only 8% and the white population consisting of more than 80% of the patients, it's hard to keep the fairness of different ethnicity balanced. In particular, we see high fluctuation with minorities like Asians, especially those intersectional groups with even fewer populations like Asian males. The robustness of these groups cannot be ensured once new data are introduced. The model consistently is better very slightly at predicting female over male's death rate but it's unclear if there is a biological basis. The difference is small enough that both genders would find it fair.

Consistently, we also observe that white and Hispanics landed much higher FNR and could therefore receive less attention from the healthcare facility as a result. Overall, the relatively similar accuracy between ethnic and gender groups can help hospitals allocate resources more efficiently, but the high FPR, FNR, low demographic parity, and high selection rate differences make the tool unreliable for minorities and may consistently underestimate the risk for white and Hispanics. It's unclear if there are inherently biological and cultural differences that affect different ethnic groups' true ICU death rates because of the difficulty of conducting controlled human experiments across the country. Test-time data augmentation did improve fairness but by a marginal amount only.

5.3 Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?

We certainly wouldn't feel comfortable deploying this tool for minorities at this early point because the robustness is too low and variability is too high. We need more data for minorities and better-labeled data with information about the country or a reliable way to take into account hospitals' physical and resource differences.

Also, we aren't entirely sure if the white population—which we have the most data for—are discriminated against by the model. While they have generally slightly lower FPR and higher FNR, it's unclear whether that is actually incorrect or only incorrect relative to the fluctuation

due to the limited minority data.

5.4 What improvements do you recommend to the data collection, processing, or analysis methodology?

The following solutions are recommended based on the observations above:

1. Gather more data, particularly for non-white patients.
2. Find metrics to control for hospital and country-wide wealth and development gaps.
3. Keep hospital ID and Country ID in the model.
4. Investigate with the domain experts whether the ethnic and gender differences are due to biology, culture, or discrimination.
5. Use Fairlearn preprocessing or in-processing tools to adjust the outcome if the difference is a product of discrimination.

References

[Coh+21] Seffi Cohen et al. "ICU survival prediction incorporating test-time augmentation to improve the accuracy of ensemble-based models". In: *IEEE Access* 9 (2021), pp. 91584–91592 (cit. on pp. 1, 33).

[EES20] Eman Mohamed Ebrahim, Ghada Abd Elsalam Eldeeb, and Zahra Ahmed Sayed. "Ageism in ICU: Knowledge, Attitude and Advocacy toward Caring of Critically ill Elderly Patient". In: *Assiut Scientific Nursing Journal* 8.23 (2020), pp. 158–166 (cit. on p. 19).

[MA97] Diane J Mick and Michael H Ackermun. "Neutralizing ageism in critical care via outcomes research". In: *AACN Advanced Critical Care* 8.4 (1997), pp. 597–608 (cit. on p. 19).

[WSB18] Mary F Wyman, Sharon Shiovitz-Ezra, and Jürgen Bengel. "Ageism in the health care system: Providers, patients, and systems". In: *Contemporary perspectives on ageism* (2018), pp. 193–212 (cit. on p. 19).

6 Appendix

6.1 Data profiling for Labs Statistics

6.1.1 Datatype and Null Count for Features

feature	datatype	null_count	null_ratio	num_uniques
d1_albumin_max	numeric	49096	0.54	35
d1_albumin_min	numeric	49096	0.54	35
d1_bilirubin_max	numeric	53673	0.59	374
d1_bilirubin_min	numeric	53673	0.59	357
d1_bun_max	numeric	10514	0.11	495
d1_bun_min	numeric	10514	0.11	472
d1_calcium_max	numeric	13069	0.14	47
d1_calcium_min	numeric	13069	0.14	49
d1_creatinine_max	numeric	10169	0.11	1187
d1_creatinine_min	numeric	10169	0.11	1093
d1_glucose_max	numeric	5807	0.06	538
d1_glucose_min	numeric	5807	0.06	256
d1_hco3_max	numeric	15071	0.16	223
d1_hco3_min	numeric	15071	0.16	255
d1_hemaglobin_max	numeric	12147	0.13	105
d1_hemaglobin_min	numeric	12147	0.13	115
d1_hematocrit_max	numeric	11654	0.13	312
d1_hematocrit_min	numeric	11654	0.13	340
d1_inr_max	numeric	57941	0.63	481
d1_inr_min	numeric	57941	0.63	393
d1_lactate_max	numeric	68396	0.75	704
d1_lactate_min	numeric	68396	0.75	486
d1_platelets_max	numeric	13444	0.15	559
d1_platelets_min	numeric	13444	0.15	540
d1_potassium_max	numeric	9585	0.1	100
d1_potassium_min	numeric	9585	0.1	116
d1_sodium_max	numeric	10195	0.11	71
d1_sodium_min	numeric	10195	0.11	138
d1_wbc_max	numeric	13174	0.14	3098
d1_wbc_min	numeric	13174	0.14	2774
h1_albumin_max	numeric	83824	0.91	37
h1_albumin_min	numeric	83824	0.91	37
h1_bilirubin_max	numeric	84619	0.92	181
h1_bilirubin_min	numeric	84619	0.92	181
h1_bun_max	numeric	75091	0.82	258

6.1.2 Value Ranges for Features

feature	unit	min	max	mean	std
d1_albumin_max	None	1.2	4.6	2.97	0.67
d1_albumin_min	g/L	1.1	4.5	2.9	0.67
d1_bilirubin_max	micromol/L	0.2	51.0	1.14	2.13
d1_bilirubin_min	micromol/L	0.2	51.0	1.07	2.02
d1_bun_max	mmol/L	4.0	126.0	25.69	20.47
d1_bun_min	mmol/L	3.0	113.09	23.77	18.8
d1_calcium_max	mmol/L	6.2	10.8	8.38	0.74
d1_calcium_min	mmol/L	5.5	10.3	8.18	0.78
d1_creatinine_max	micromol/L	0.34	11.11	1.49	1.51
d1_creatinine_min	micromol/L	0.3	9.94	1.37	1.33
d1_glucose_max	mmol/L	73.0	611.0	174.64	86.69
d1_glucose_min	mmol/L	33.0	288.0	114.38	38.27
d1_hco3_max	mmol/L	12.0	40.0	24.37	4.37
d1_hco3_min	None	7.0	39.0	23.17	4.99
d1_hemaglobin_max	g/dL	6.8	17.2	11.45	2.17
d1_hemaglobin_min	g/dL	5.3	16.7	10.89	2.36
d1_hematocrit_max	Fraction	20.4	51.5	34.53	6.24
d1_hematocrit_min	Fraction	16.1	50.0	32.95	6.85
d1_inr_max	micromol/L	0.9	7.76	1.6	0.96
d1_inr_min	micromol/L	0.9	6.13	1.48	0.75
d1_lactate_max	mmol/L	0.4	19.8	2.93	3.08
d1_lactate_min	mmol/L	0.4	15.1	2.13	2.11
d1_platelets_max	$10^9/L$	27.0	585.0	207.11	89.63
d1_platelets_min	$10^9/L$	18.55	557.45	196.77	88.18
d1_potassium_max	mmol/L	2.8	7.0	4.25	0.67
d1_potassium_min	mmol/L	2.4	5.8	3.93	0.58
d1_sodium_max	mmol/L	123.0	158.0	139.12	4.82
d1_sodium_min	mmol/L	117.0	153.0	137.72	4.92
d1_wbc_max	$10^9/L$	1.2	46.08	12.48	6.8
d1_wbc_min	$10^9/L$	0.9	40.9	11.31	5.95
h1_albumin_max	None	1.1	4.7	3.03	0.73
h1_albumin_min	g/L	1.1	4.7	3.03	0.73
h1_bilirubin_max	micromol/L	0.2	40.4	1.1	2.03
h1_bilirubin_min	micromol/L	0.2	40.4	1.1	2.03
h1_bun_max	mmol/L	4.0	135.0	25.84	21.44

6.2 Data Profiling for Lab Blood Gas Results

6.2.1 Datatype and Null Count for Features

feature	datatype	null_count	null_ratio	num_uniques
d1_arterial_pco2_max	numeric	59271	0.65	831
d1_arterial_pco2_min	numeric	59271	0.65	674
d1_arterial_ph_max	numeric	60123	0.66	527
d1_arterial_ph_min	numeric	60123	0.66	624
d1_arterial_po2_max	numeric	59262	0.65	2428
d1_arterial_po2_min	numeric	59262	0.65	1822
d1_pao2fio2ratio_max	numeric	66008	0.72	5194
d1_pao2fio2ratio_min	numeric	66008	0.72	4990
h1_arterial_pco2_max	numeric	75959	0.83	779
h1_arterial_pco2_min	numeric	75959	0.83	768
h1_arterial_ph_max	numeric	76424	0.83	561
h1_arterial_ph_min	numeric	76424	0.83	579
h1_arterial_po2_max	numeric	75945	0.83	1737
h1_arterial_po2_min	numeric	75945	0.83	1729
h1_pao2fio2ratio_max	numeric	80195	0.87	3142
h1_pao2fio2ratio_min	numeric	80195	0.87	3122

6.2.2 Value Ranges for Features

feature	unit	min	max	mean	std
d1_arterial_pco2_max	Millimetres of mercury	18.4	111.0	45.25	14.67
d1_arterial_pco2_min	Millimetres of mercury	14.9	85.91	38.43	10.94
d1_arterial_ph_max	None	7.05	7.62	7.39	0.08
d1_arterial_ph_min	None	6.89	7.56	7.32	0.11
d1_arterial_po2_max	Millimetres of mercury	39.0	540.87	165.91	108.01
d1_arterial_po2_min	Millimetres of mercury	28.0	448.89	103.51	61.85
d1_pao2fio2ratio_max	Fraction	54.8	834.8	285.67	128.22
d1_pao2fio2ratio_min	Fraction	36.0	604.23	223.52	117.55
h1_arterial_pco2_max	Millimetres of mercury	15.0	111.5	44.67	14.63
h1_arterial_pco2_min	Millimetres of mercury	15.0	107.0	43.38	14.11
h1_arterial_ph_max	None	6.93	7.57	7.34	0.11
h1_arterial_ph_min	None	6.9	7.56	7.33	0.11
h1_arterial_po2_max	Millimetres of mercury	34.0	534.9	163.84	113.46
h1_arterial_po2_min	Millimetres of mercury	31.0	514.9	144.15	98.46
h1_pao2fio2ratio_max	Fraction	42.0	720.0	244.4	129.96
h1_pao2fio2ratio_min	Fraction	38.0	654.81	235.93	126.46

6.3 Value Distribution for Height

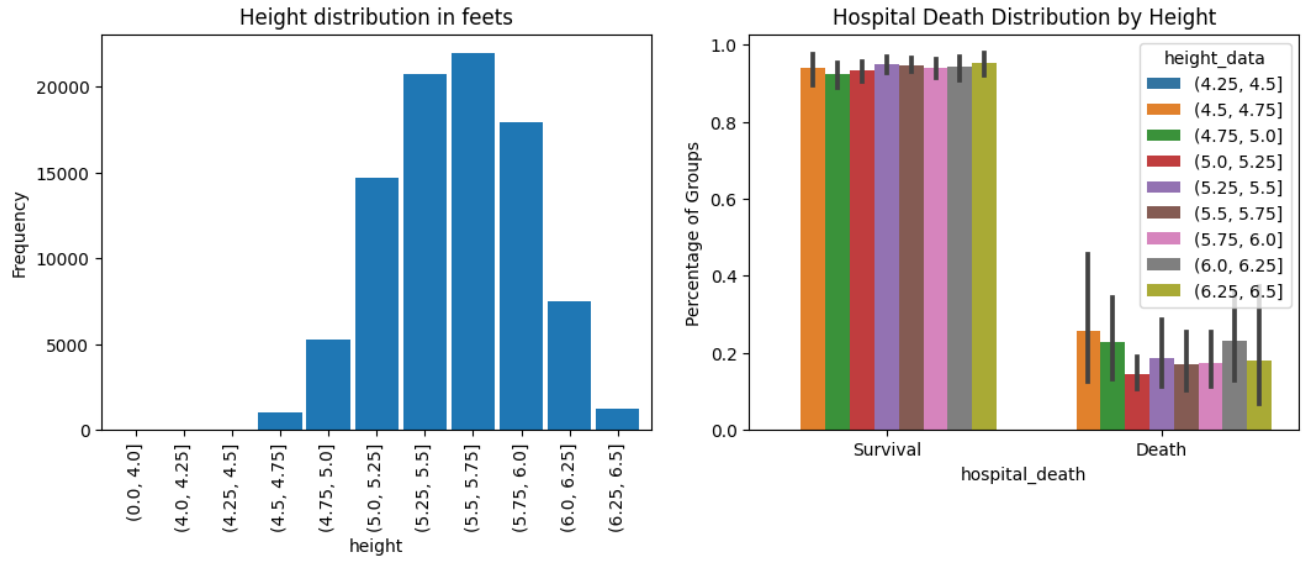

Figure 24: Height Distribution

6.4 Value Distribution for Weight

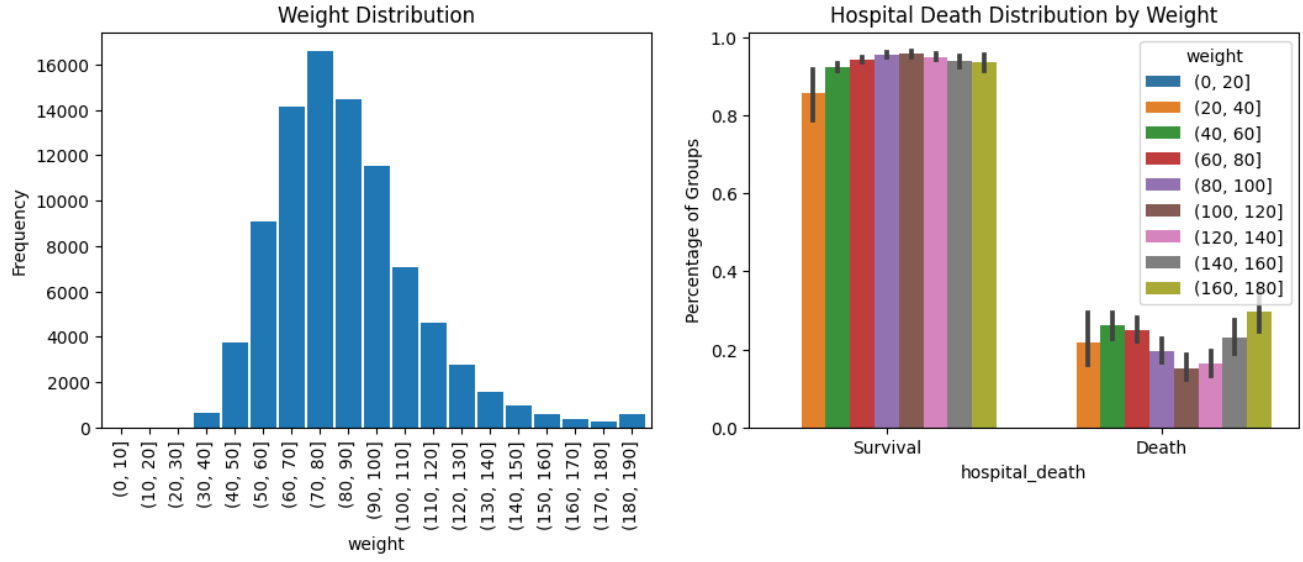

Figure 25: Weight Distribution

6.5 Value Distribution for Temperature (Celsius)

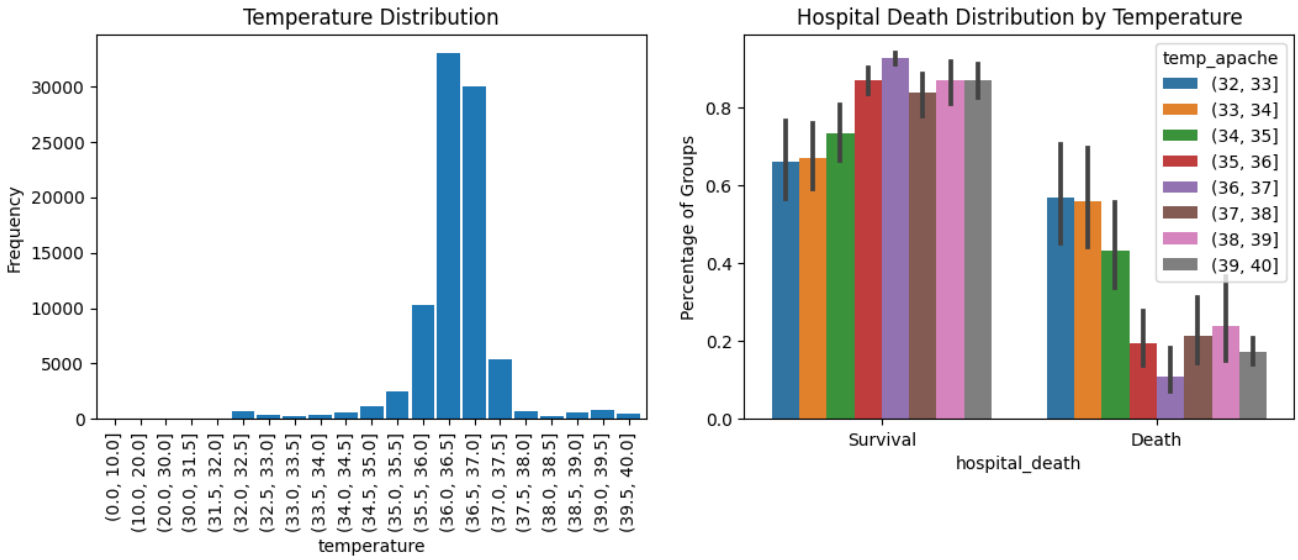

Figure 26: Temperature Distribution

6.6 Value Distribution for Respiratory Rate (Breaths per minute)

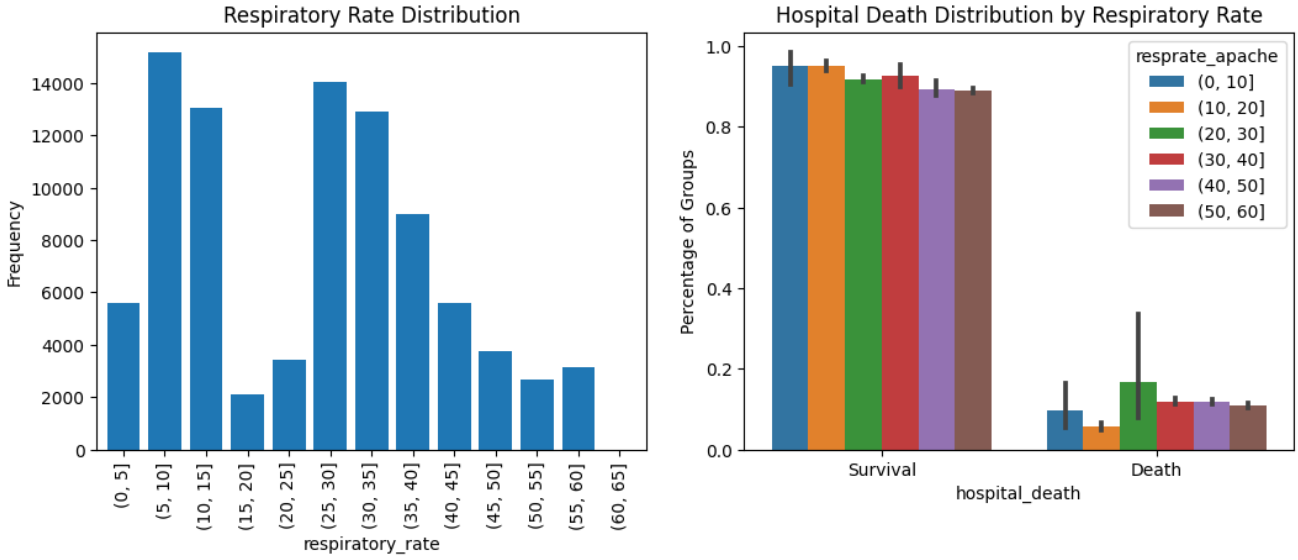

Figure 27: Respiratory Rate Distribution

6.7 Pairwise Correlations

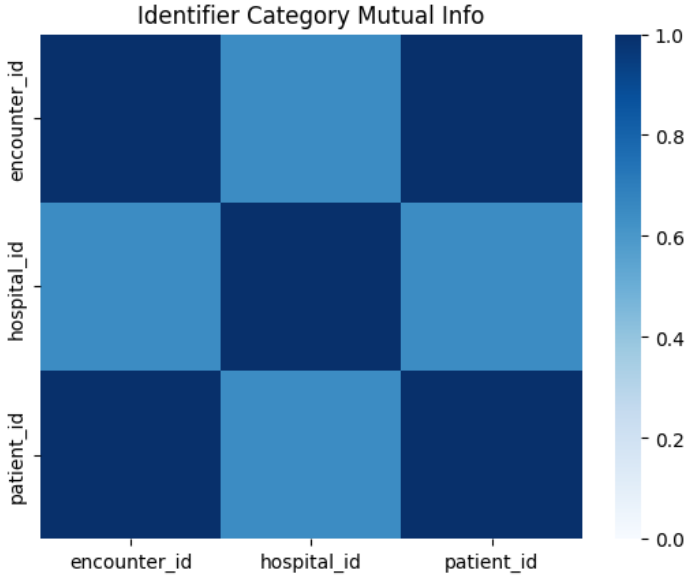

Figure 28: Pairwise Correlation between Identifier Attributes

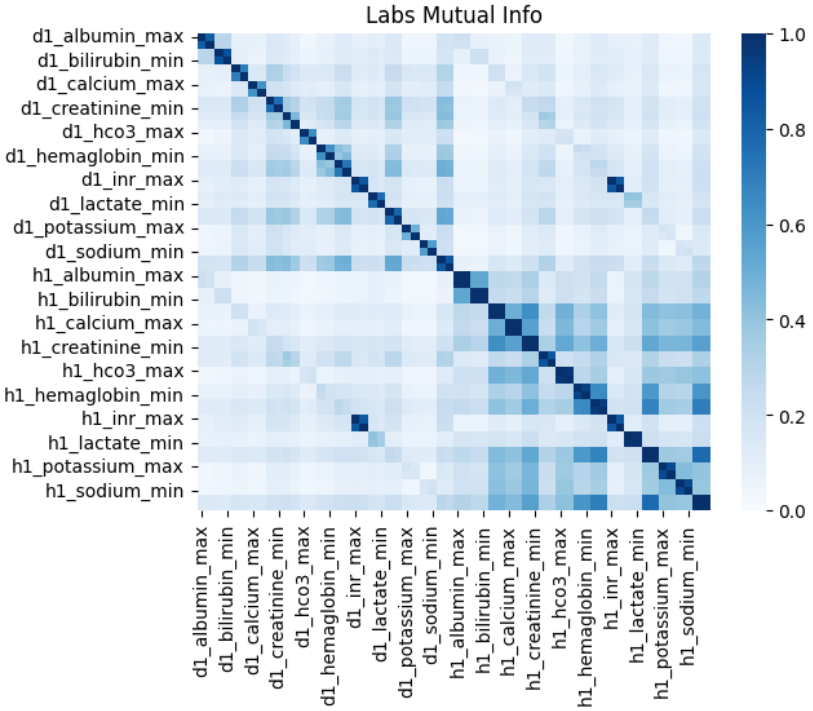

Figure 29: Pairwise Correlation between Lab Results

6.8 Actual Death Rate by Gender

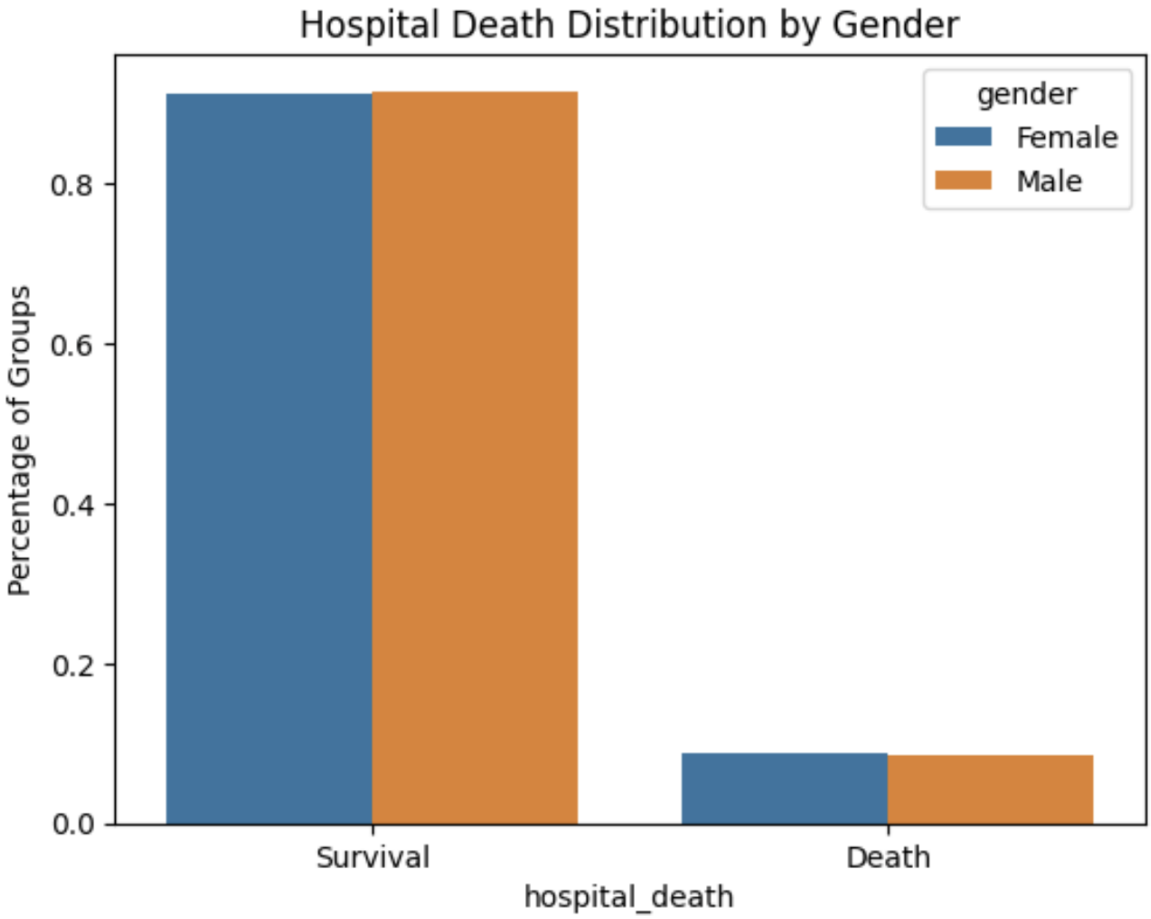

Figure 30: Actual Death Rate by Gender

6.9 Accuracy Metrics By Gender (Before Test Time Augmentation)

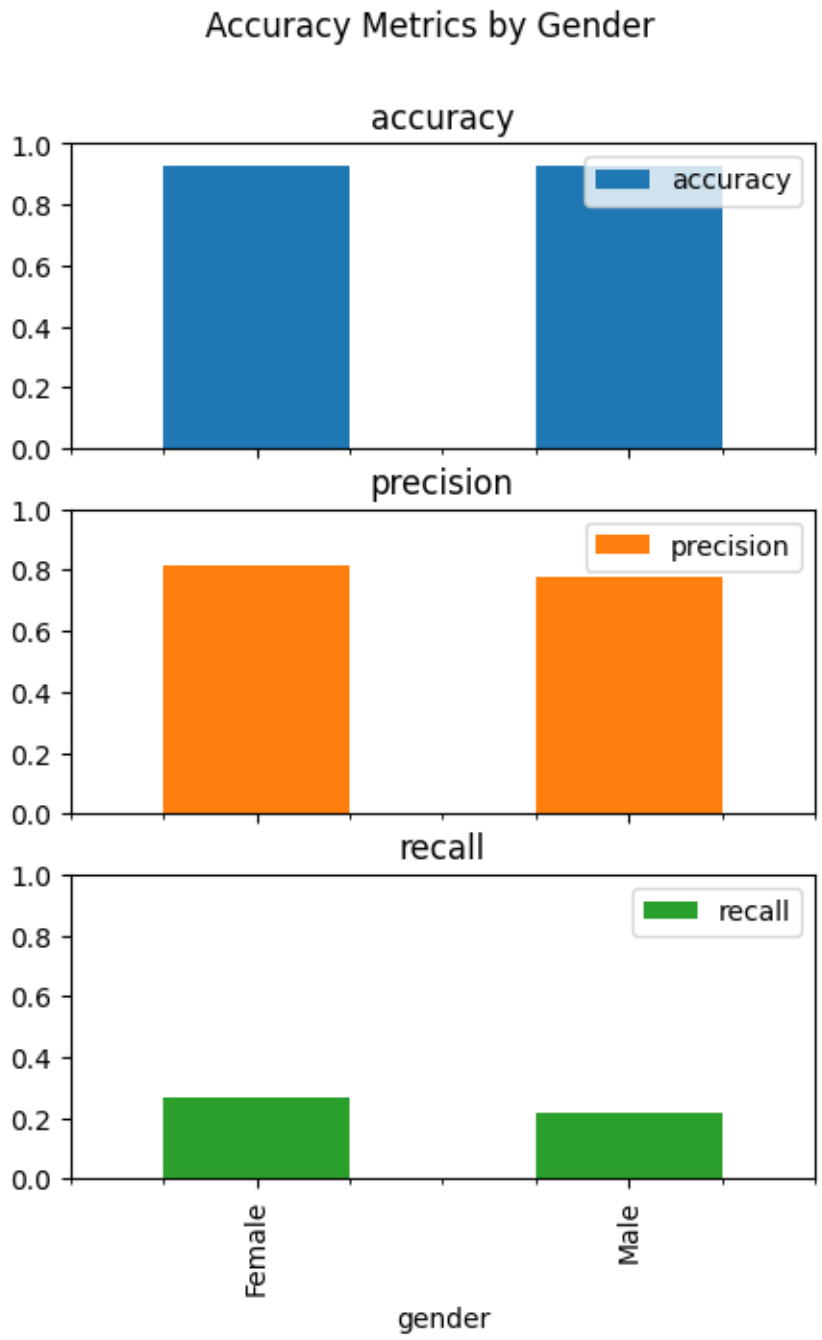

Figure 31: Accuracy Metrics By Gender (Before Test Time Augmentation)

6.10 Accuracy Metrics By Ethnicity (Before Test Time Augmentation)

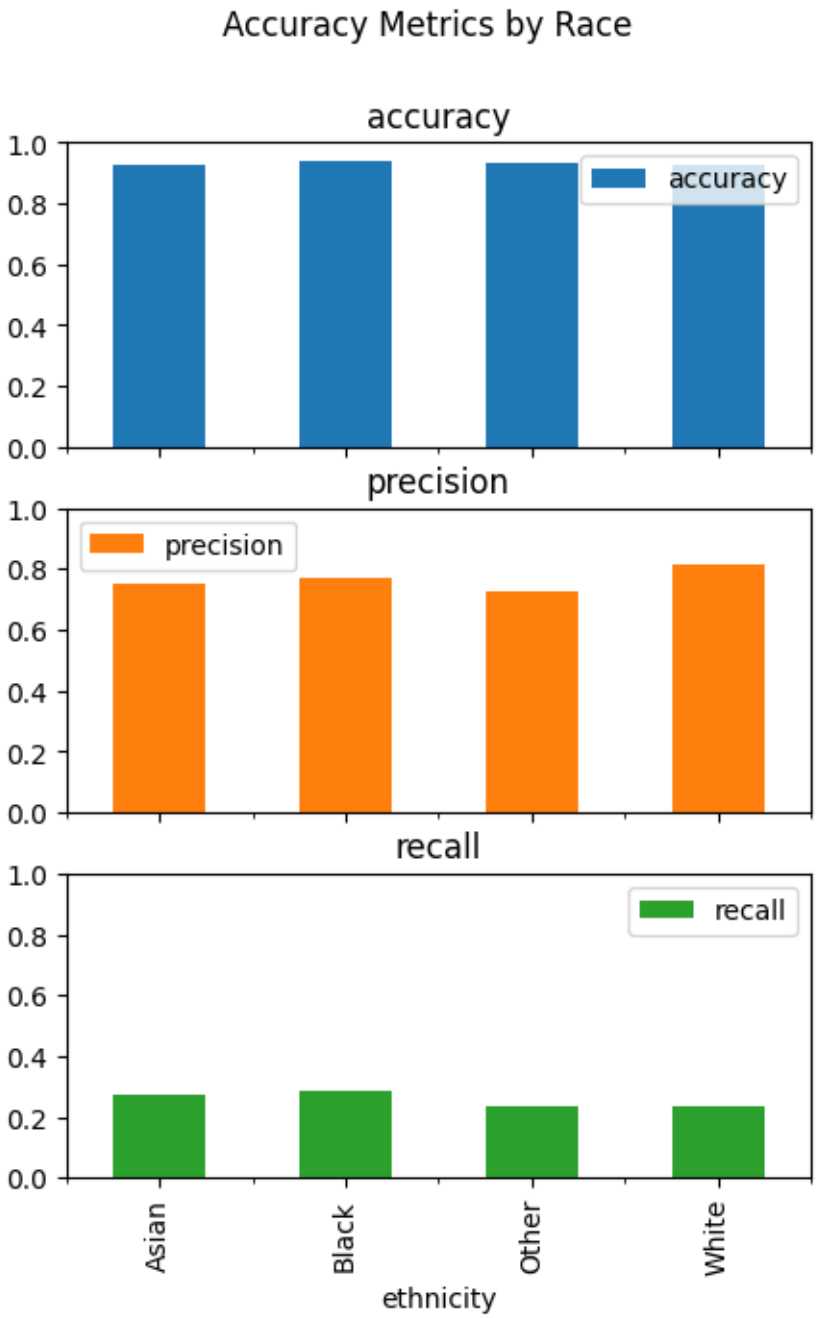

Figure 32: Accuracy Metrics By Ethnicity (Before Test Time Augmentation)

6.11 Fairness Metrics By Gender (Before Test Time Augmentation)

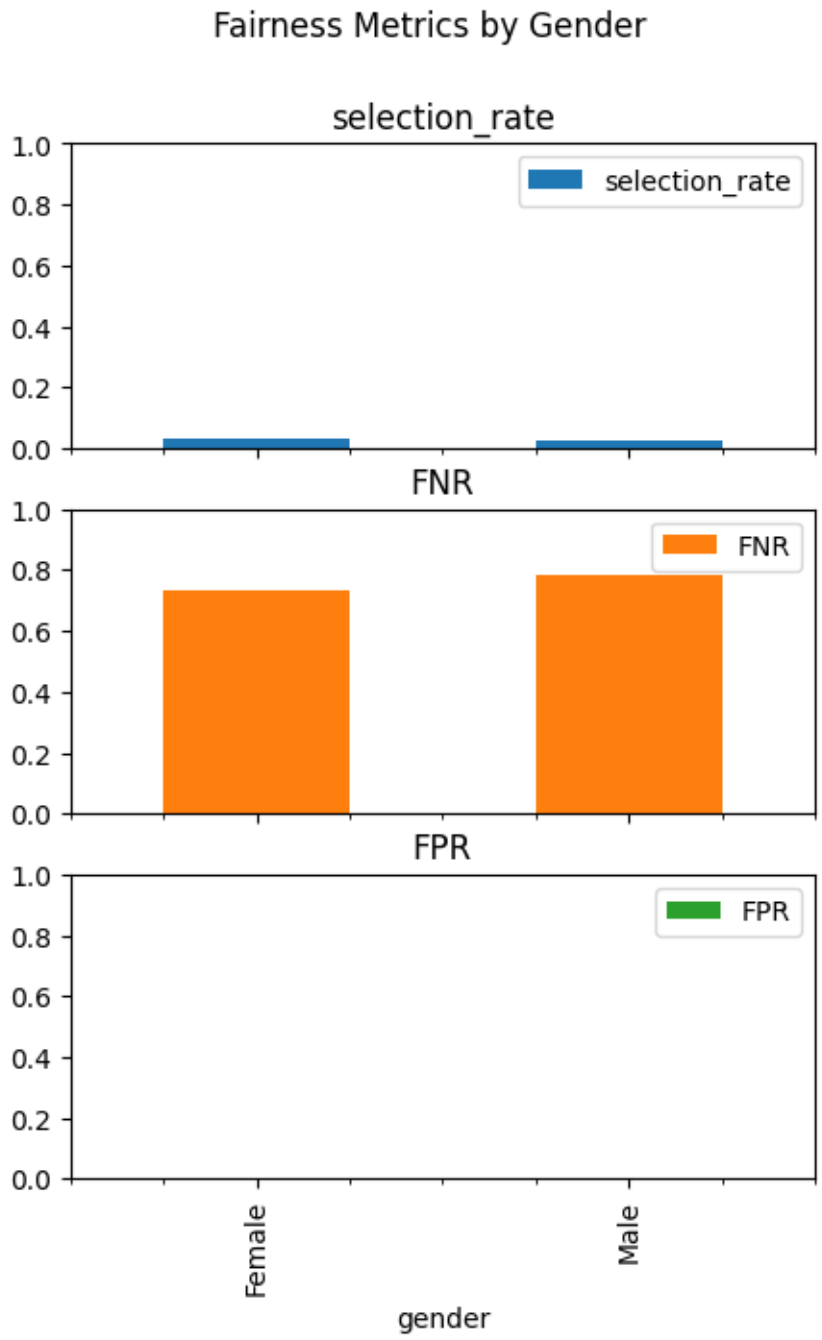

Figure 33: Fairness Metrics By Gender (Before Test Time Augmentation)

6.12 Team Cohen's SHAP Values

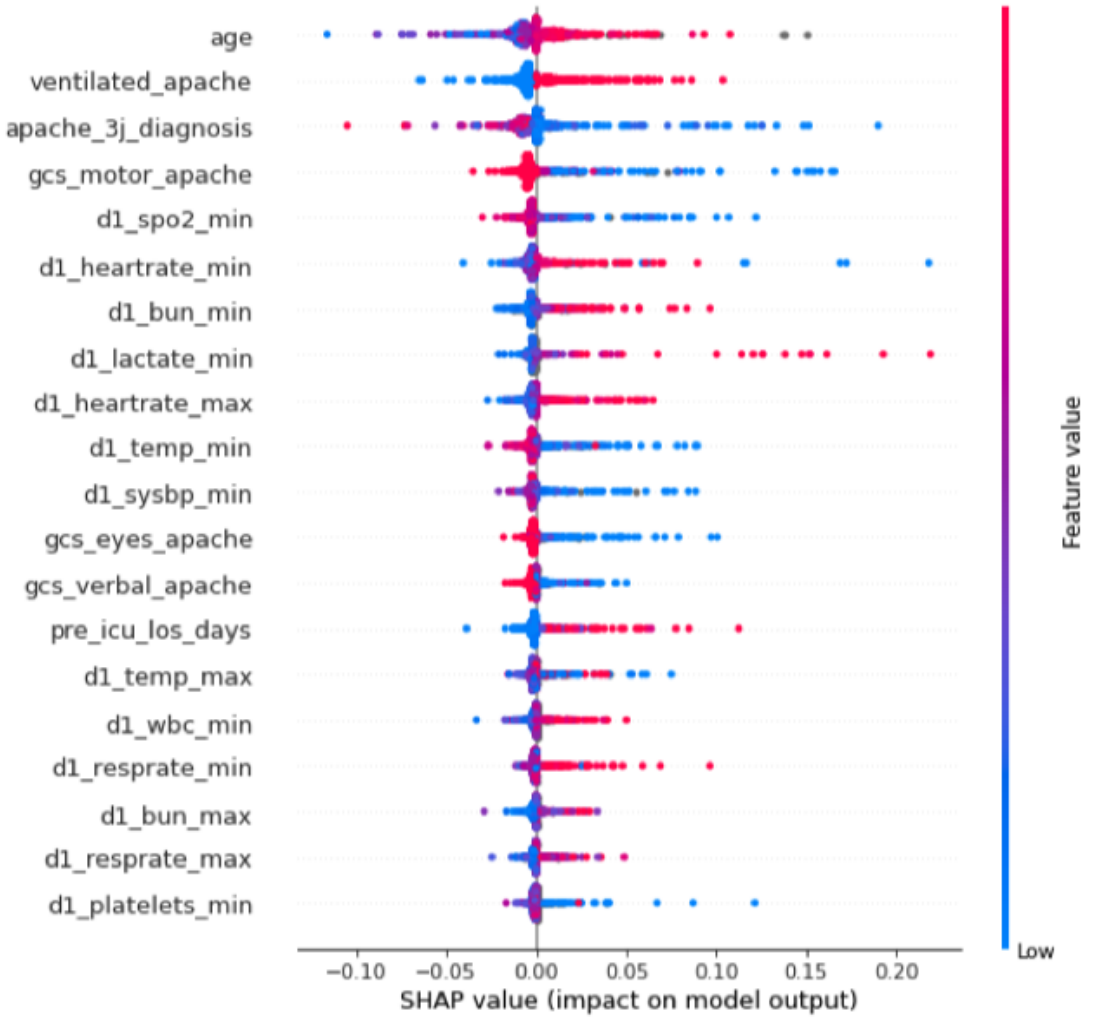

FIGURE 7. The top 20 SHAP values on the entire ensemble of models. The points on this plot represent the impact of the features on the prediction. For example, high values of 'age' caused higher probabilities, and low values caused low probabilities.

Figure 34: Team Cohen's SHAP value for their Stack-Net model