東海大學資訊工程研究所

碩士論文

Department of Computer Science

Tunghai University

急重症資料AI工具開發：使用XGBoost演算法預測COPD患者28天ICU致死率

An Interactive Medical Data Extraction Tool for MIMIC-III Database: 28-Day Mortality Prediction for COPD Patients Admitted in ICU using XGBOOST

指導教授：賴俊鳴

Advisor：Chun-Ming Lai, PhD.

研究生：馬可凡

Graduate Student：Frank Makowa

中華民國一一○年十月

Oct, 2021

Abstract

MIMIC-III has proven to be one of the commonest databases researchers use for their studies since it contains almost all the information one might need to perform real-world medical studies. However, most researchers find it is time-consuming to write codes and extract data from the database. With this challenge in mind, in this work, we present a tool that will enable researchers extract data from MIMIC-III as appropriate. From the data we extracted when testing the tool, a machine learning model is developed using eXtreme Gradient Boosting (XGBoost) to predict 28-day mortality in ICU for COPD patients and is explained using SHAP value. We further developed a much complex model by adding more features that were calculated on Elixhauser, Clinical Cut Point (CP) and ROC CP Scoring methods. Three other machine learning models were developed and compared with our main models. Both of the XGBoost models performed well based on AUROC, Accuracy and F1-score.

Table of Contents

Abstract i

Table of Contents ii

List of Figures iv

List of Tables v

1 Introduction 1

2 Related Work 2

2.1 MIMIC III 2

2.2 Data extractor 2

2.3 COPD 3

2.4 XGBOOST and SHAP 3

2.5 Logistic Regression 4

2.6 K-Nearest Neighbors 4

2.7 Random Forest 5

3 Methodology 6

3.1 Extraction Tool 6

3.2 Cohort recruitment and default features 6

3.3 Comorbidities 7

3.4 Additional Feature Selection 8

4 Experiments 11

4.1 Cohort 11

4.2 Statistical analysis 11

4.3 Inclusion Criteria 14

4.4 Experiment 1 16

4.4.1 Basic model 16

4.4.2 Model performance and Explanation 16
4.5 Experiment 2 20
4.5.1 Extra-Features Model 20
4.5.2 Model performance and Explanation 21
5 Discussion 24
6 Conclusion 26
Acknowledgement 27
References 28
Appendix A 32

List of Figures

Figure 1 Proposed System Architecture 6
Figure 2 ICD9_CODES search 7
Figure 3 Adding comorbidities 8
Figure 4 Search for item_id in the database 10
Figure 5 Searching item values (features) and adding feature column to the data frame 10
Figure 6 ROC curve for Scoring Models 14
Figure 7 Recruitment Inclusion Criteria 15
Figure 8 AUROC: Area Under Receiver Operating Curve for Basic model. 17
Figure 9 SHAP values for the single instance 18
Figure 10 SHAP Value impact on model output 19
Figure 11 Average SHAP Value Importance for XGBoost with basic feature 20
Figure 12 AUROC: Area Under Receiver Operating Curve for all models tested on extra featured dataset. (KNN. LR, RF, XGBoost) 21
Figure 13 Average SHAP Value Importance for XGBoost with more featues 22

List of Tables

Table 1Mimic 3 tables descriptions (source Mimic 3 website) 9
Table 2 Demographic data and outcomes regarding to mortality in COPD patients 12
Table 3 Univariable and Multivariable Regression Analyses of Mortality 13
Table 4 Model comparison results (AUROC, Accuracy, and F1-score) 23

1 Introduction

Having access and being able to easily get data from a medical database is one of the most desired things for medical researchers. There are several data sources from which researchers can get data for their various studies; some are public and others are private. Some sources have the data already grouped in different categories and diseases with all features related to it (i.e. vital signs, lab tests and medication) already predefined. For other sources, however, one has to search for the links between the disease features. In most cases, these databases offer more information than the predefined ones. Some of these data sources include the Healthcare Cost and Utilization Project (HCUP), data.gov, Kent Ridge Bio-medical Dataset, UCI machine learning repository, as well as MIMIC databases.

The Medical Information Mart for Intensive Care (MIMIC)-III database has data for over 40,000 patients from Beth Israel Deaconess Medical Center (BIDMC) [1] that were admitted to the intensive care units (ICU). Mimic-III database is one of the free databases which is widely used by many researchers globally. Despite having access to such, some researchers especially those that are not familiar with medical databases find it difficult to navigate through. Johnson et al. [2] in their research provided examples on how to do recruitment of cohorts and also emphasized on the idea of reproducible studies. This is a concern to many who are trying to do a me-too study as the code(s) may produce different results from the original study [3].

In this study, we develop a tool to help researchers extract enough information to get them started with their research. This tool is tested by extracting COPD data and developing a model to predict mortality in COPD patients admitted in the ICU up to 28 days.

2 Related Work

2.1 MIMIC III

MIMIC-III is relational databases which consist of health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center (BIDMC) between the years 2001 and 2012. A variety of information like demographics, bedside vital sign measurements, laboratory test results, procedures, medications, notes from caregiver, imaging reports, and mortality (together with post-hospital discharge) are included in this database. The data in MIMIC-III database was first de-identified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting before it was added to the database.

MIMIC-III has two information systems for critical care that were in place during the data collection period as described, in which are the Philips CareVue Clinical Information System (models M2331A and M1215A; Philips Health-care, Andover, MA) - just known as CareVue, and iMDsoft MetaVision ICU (iMDsoft, Needham, MA) - known as MetaVision.

Several researchers around the world conduct miscellaneous range of analytical studies spanning from epidemiology, clinical decision-rule improvement, and development of electronic tools on MIMIC-III [2]. One of the reasons MIMIC-III is popular and a good database for researches is that its data ranges from time-stamped, several physiological measurements verified by nurses, free-text interpretations of radiology department images for studies.

2.2 Data extractor

Data extraction is the process of retrieving data from a source or various sources into a more useful format for further processing or storage. There are several tools that were developed to do data extraction; some are open-source, while others are off-the-shelf. These data-extraction tools allow the retrieval of data from well-structured, poorly-structured and unstructured data. Many works have developed different data extraction tools for different sources. For example, Ferrara et al. [4] focused on web scrapping tools and techniques in their study. Examples of web-extraction tools include importio, myTrama, Octoparse, Hevo, and many more [5]–[7].

Others devoted their work on developing a PDF data-extraction tool. The PDF Data Extractor (PDE) - which is an R package - extracts data from full-text articles from PDF based on keywords that are defined by the user, without the requirement of a training set[8]. Some researchers[3], [9], [10] worked on data extraction tools for relational databases.

2.3 COPD

Chronic Obstructive Pulmonary Disease (COPD) accounted for nearly 3.2 million deaths globally in the year 2017 and ranked third among the leading causes of death worldwide. By the year 2020, COPD is believed to have caused an estimate of 6 million global deaths annually [11]. The WHO [12] expressed that COPD is mostly associated with risk factors ranging from poor nutrition, smoking, exposure to fumes and smoke, occupational hazards. Despite having several risk factors, COPD also has several comorbidities associated with it [13], [14].

2.4 XGBOOST and SHAP

Extreme Gradient Boosting popularly knowns as XGBoost, is an advanced machine learning algorithm (MLA) that was based on a gradient-boosting decision tree. It combines a set of algorithms to come up with a much better MLA as a whole. XGBoost offers a parallel tree boosting (GBDT, GBM) which is the reason many science problems are solved fast and with high accuracy. Recently, several studies have employed the use of XGBoost to solve many scientific problems [15]–[17]. It is one of the leading few machine learning libraries used for solving problems for classification, ranking as well as regression. Some of the advantages of XGBoost are that it is scalable, and fast.

It is noted that most of the machine learning models developed using XGBoost such as[18]–[20] are explained well using SHAP.

"SHAP (SHapley Additive exPlanations) *is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions"* [21]

Above is the definition of SHAP from the developers, but on their blog, Dario Radecic [22], they simplified the definition further by saying, “they are measures of contributions each predictor (feature) has in a machine learning model.”

In their study, Lundberg et al. [19] demonstrated how inconsistent popular techniques for feature attribution can be. This is to say that the methods can lower a feature’s assigned importance when the true impact of that particular feature is essentially high which casts doubt on any comparison between features. This is a fundamental problem was addressed using SHAP values, which are the unique, reliable plus locally precise attribution values. The study Lundberg et al. [19] compared SHAP with other feature attribution methods like Saabas, Gain, Permutation, and Split Count on two simple tree models.

2.5 Logistic Regression

Logistic Regression (LR) is a classification method in statistical model, which was borrowed into machine learning. It is a choice in many medical data classifications used to compute the likelihood of certain classes or events, and it also allows modeling and multivariate analysis of binary dependent variables. In the study to categorize type 1 and type 2 diabetes patients A. L. Lynam *et al.* [23] used logistic regression as their algorithm. The coefficients of predictors included in the final model are estimated using the multivariate analysis and are then adjusted based on the predictors of the model. The risk estimate of the outcome is quantified by the contribution of each predictor.

2.6 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple supervised learning machine learning algorithm that assumes similar things exists in close proximity and looks for a pattern in those occurrences. KNN can be used for both classification problems or regression problems. Despite being a simple algorithm, it can still give competitive results. However, in the industry, KNN is widely used in classification problems. Theerthagriri et al. in their study [24] used KNN to predict

COVID-19 possibilities with an 80.4% prediction accuracy compared with other models in their study.

2.7 Random Forest

Random Forest (RF) is an ensemble machine learning algorithm, which puts together multiple decision trees to predict the outcome based on the average probability of all the trees on each subset of data samples to obtain better predictive performance that can't be obtained by a single algorithm. In particular, Spathis et al. [25] used RF in diagnosing asthma and COPD. Apart from producing precise predictions, being fast and easy to implement, with even a large number of input variables RF can perform quite well without overfitting.

3 Methodology

In this section, we discuss the system architecture and methodology applied in this study.

3.1 Extraction Tool

This tool was developed using python 3.8, JupyterLab 3.0.14; the PostgreSQL database server is running on Ubuntu 20.04 server. Basically, it includes three main parts: Cohort recruitment with default features, Comorbidities, Additional Features. Figure 1 shows the main building blocks of our project.

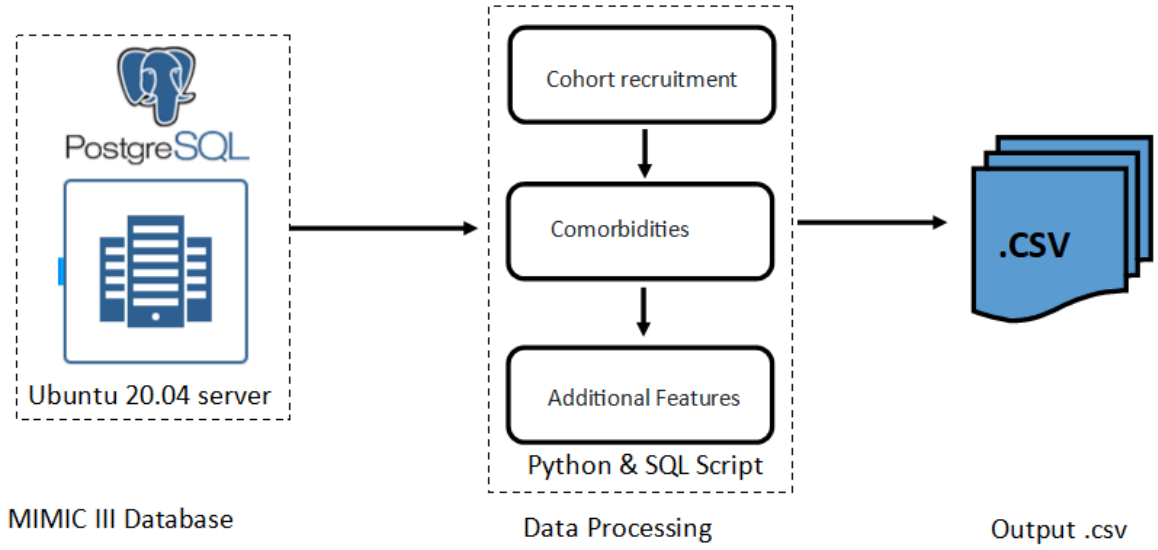
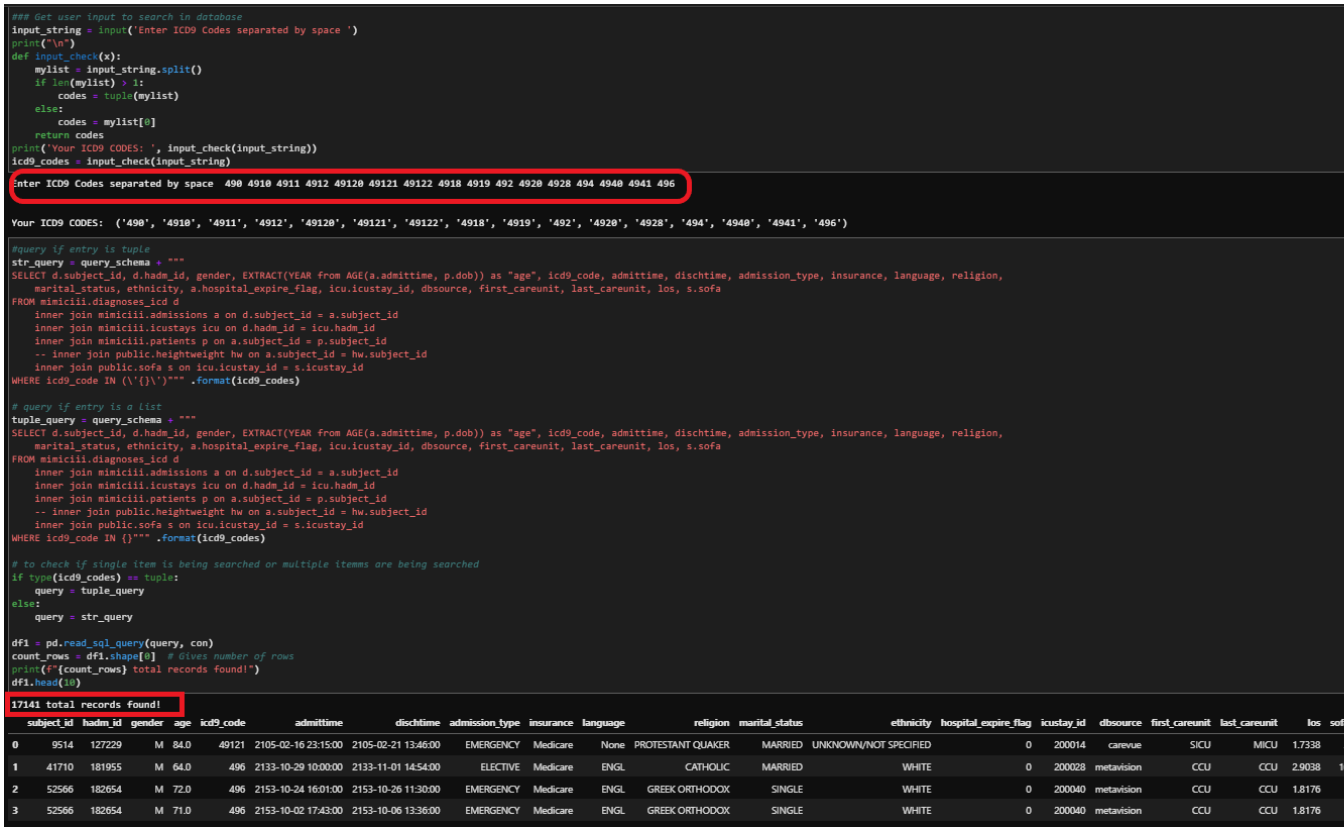


Figure 1 Proposed System Architecture

3.2 Cohort recruitment and default features

In MIMIC-III, getting patients with similar disease is easier when the International Classification of Diseases, Ninth Revision, Clinical Modification (icd9_cm) [26] are used to do the search than using name search. The first query when getting default features using icd9_code for the primary disease, we try to include as many patient demographic features as possible. The default demographic data extracted include: *age, gender, ethnicity, length of stay in hospital,*

length of stay in the ICU, hospital expire flag, admit time, discharge time, admission type, insurance, language, religion, marital status, Sofa score.



```
### Get user input to search in database
input_string = input('Enter ICD9 Codes separated by space ')
print("\n")
def input_check(x):
    mylist = input_string.split()
    if len(mylist) > 1:
        codes = tuple(mylist)
    else:
        codes = mylist[0]
    return codes
print('Your ICD9 CODES: ', input_check(input_string))
icd9_codes = input_check(input_string)
```

```
Enter ICD9 Codes separated by space  490 4910 4911 4912 49120 49121 49122 4918 4919 492 4920 4928 494 4940 4941 496

Your ICD9 CODES:  ('490', '4910', '4911', '4912', '49120', '49121', '49122', '4918', '4919', '492', '4920', '4928', '494', '4940', '4941', '496')
```

```
#query if entry is tuple
str_query = query_schema + """
SELECT d.subject_id, d.hadm_id, gender, EXTRACT(YEAR from AGE(a.admittime, p.dob)) as "age", icd9_code, admittime, dischtime, admission_type, insurance, language, religion,
    marital_status, ethnicity, a.hospital_expire_flag, icu.icustay_id, dbsource, first_careunit, last_careunit, los, s.sofa
FROM mimiciii.diagnoses_icd d
    inner join mimiciii.admissions a on d.subject_id = a.subject_id
    inner join mimiciii.icustays icu on d.hadm_id = icu.hadm_id
    inner join mimiciii.patients p on a.subject_id = p.subject_id
    -- inner join public.heightweight hw on a.subject_id = hw.subject_id
    inner join public.sofa s on icu.icustay_id = s.icustay_id
WHERE icd9_code IN (\'{}\')""" .format(icd9_codes)

# query if entry is a list
tuple_query = query_schema + """
SELECT d.subject_id, d.hadm_id, gender, EXTRACT(YEAR from AGE(a.admittime, p.dob)) as "age", icd9_code, admittime, dischtime, admission_type, insurance, language, religion,
    marital_status, ethnicity, a.hospital_expire_flag, icu.icustay_id, dbsource, first_careunit, last_careunit, los, s.sofa
FROM mimiciii.diagnoses_icd d
    inner join mimiciii.admissions a on d.subject_id = a.subject_id
    inner join mimiciii.icustays icu on d.hadm_id = icu.hadm_id
    inner join mimiciii.patients p on a.subject_id = p.subject_id
    -- inner join public.heightweight hw on a.subject_id = hw.subject_id
    inner join public.sofa s on icu.icustay_id = s.icustay_id
WHERE icd9_code IN {}""" .format(icd9_codes)

# to check if single item is being searched or multiple itemms are being searched
if type(icd9_codes) == tuple:
    query = tuple_query
else:
    query = str_query

df1 = pd.read_sql_query(query, con)
count_rows = df1.shape[0]  # Gives number of rows
print(f"{count_rows} total records found!")
df1.head(10)
```

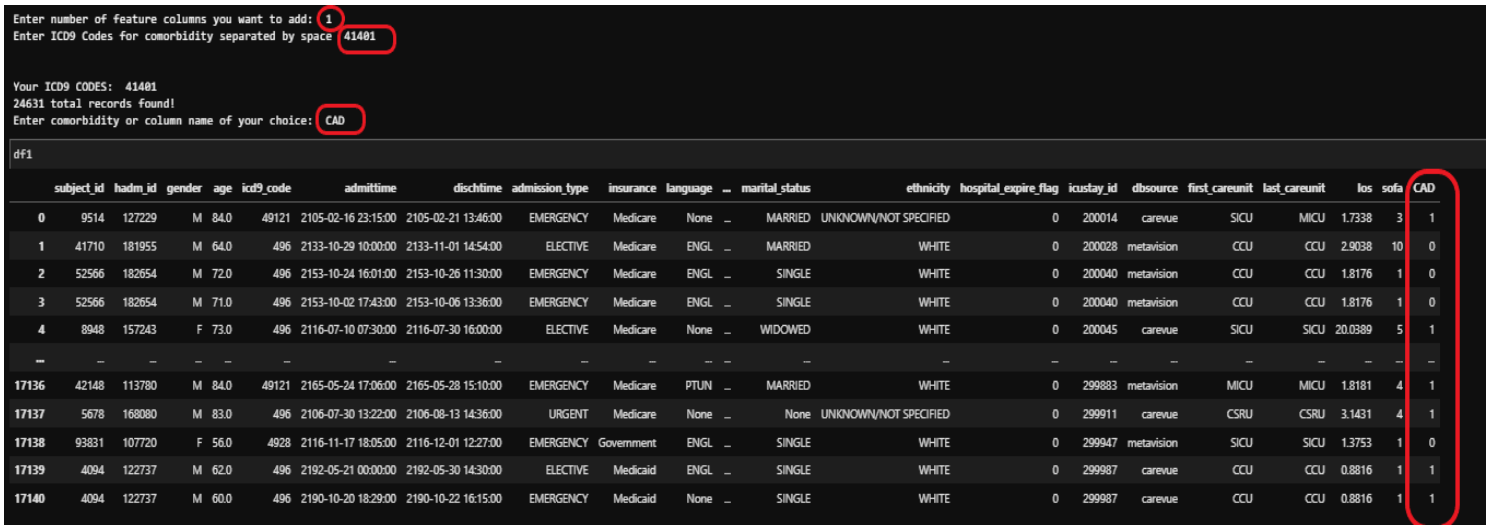
```
17141 total records found!
```

	subject_id	hadm_id	gender	age	icd9_code	admittime	dischtime	admission_type	insurance	language	religion	marital_status	ethnicity	hospital_expire_flag	icustay_id	dbsource	first_careunit	last_careunit	los	sof
0	9514	127229	M	84.0	49121	2105-02-16 23:15:00	2105-02-21 13:46:00	EMERGENCY	Medicare	None	PROTESTANT QUAKER	MARRIED	UNKNOWN/NOT SPECIFIED	0	200014	carevue	SICU	MICU	1.7338	
1	41710	181955	M	64.0	496	2133-10-29 10:00:00	2133-11-01 14:54:00	ELECTIVE	Medicare	ENGL	CATHOLIC	MARRIED	WHITE	0	200028	metavision	CCU	CCU	2.9038	1
2	52566	182654	M	72.0	496	2153-10-24 16:01:00	2153-10-26 11:30:00	EMERGENCY	Medicare	ENGL	GREEK ORTHODOX	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	
3	52566	182654	M	71.0	496	2153-10-02 17:43:00	2153-10-06 13:36:00	EMERGENCY	Medicare	ENGL	GREEK ORTHODOX	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	

Figure 2 ICD9_CODES search

3.3 Comorbidities

Some studies like [13], [14], [27]–[29] may require researchers to find some comorbidities [30] associated with the primary disease. As similar as the primary reason of admission, icd9_codes are used to identify comorbidities that a patient has during each admission. The tool searches for comorbidities by icd9_code and then creates a column with a name provided by the user marking '1' where the comorbidity is present, and '0' where it is not. Figure 3 shows how the comorbidities are searched and how the column is created.



```
Enter number of feature columns you want to add: 1
Enter ICD9 Codes for comorbidity separated by space 41401

Your ICD9 CODES:  41401
24631 total records found!
Enter comorbidity or column name of your choice: CAD
```

df1

	subject_id	hadm_id	gender	age	icd9_code	admittime	dischtime	admission_type	insurance	language	...	marital_status	ethnicity	hospital_expire_flag	icustay_id	dbsource	first_careunit	last_careunit	los	sofa	CAD
0	9514	127229	M	84.0	49121	2105-02-16 23:15:00	2105-02-21 13:46:00	EMERGENCY	Medicare	None	...	MARRIED	UNKNOWN/NOT SPECIFIED	0	200014	carevue	SICU	MICU	1.7338	3	1
1	41710	181955	M	64.0	496	2133-10-29 10:00:00	2133-11-01 14:54:00	ELECTIVE	Medicare	ENGL	...	MARRIED	WHITE	0	200028	metavision	CCU	CCU	2.9038	10	0
2	52566	182654	M	72.0	496	2153-10-24 16:01:00	2153-10-26 11:30:00	EMERGENCY	Medicare	ENGL	...	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	1	0
3	52566	182654	M	71.0	496	2153-10-02 17:43:00	2153-10-06 13:36:00	EMERGENCY	Medicare	ENGL	...	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	1	0
4	8948	157243	F	73.0	496	2116-07-10 07:30:00	2116-07-30 16:00:00	ELECTIVE	Medicare	None	...	WIDOWED	WHITE	0	200045	carevue	SICU	SICU	20.0389	5	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17136	42148	113780	M	84.0	49121	2165-05-24 17:06:00	2165-05-28 15:10:00	EMERGENCY	Medicare	PTUN	...	MARRIED	WHITE	0	299883	metavision	MICU	MICU	1.8181	4	1
17137	5678	168080	M	83.0	496	2106-07-30 13:22:00	2106-08-13 14:36:00	URGENT	Medicare	None	...	None	UNKNOWN/NOT SPECIFIED	0	299911	carevue	CSRU	CSRU	3.1431	4	1
17138	93831	107720	F	56.0	4928	2116-11-17 18:05:00	2116-12-01 12:27:00	EMERGENCY	Government	ENGL	...	SINGLE	WHITE	0	299947	metavision	SICU	SICU	1.3753	1	0
17139	4094	122737	M	62.0	496	2192-05-21 00:00:00	2192-05-30 14:30:00	ELECTIVE	Medicaid	ENGL	...	SINGLE	WHITE	0	299987	carevue	CCU	CCU	0.8816	1	1
17140	4094	122737	M	60.0	496	2190-10-20 18:29:00	2190-10-22 16:15:00	EMERGENCY	Medicaid	None	...	SINGLE	WHITE	0	299987	carevue	CCU	CCU	0.8816	1	1

3.4 Additional Feature Selection

In this part of the extraction, we have basic demographic features which we call default features, and then the other part consists of features from vital signs, lab measurements, and treatment interventions, which can be added to the data frame by the researcher. These are what we refer to as dynamic features in this study. To search for an item, we use item ID's to find the item value from different tables. This part has two functions, the first one is to search for the item ID using a substring so that item ID's can be identified. MIMIC-III has two tables that have item_id definitions, d_items and d_labitems [1], [2], these tables are regarded as dictionaries of local codes known as ITEMIDs which are in the MIMIC-III database. The second function is to use the item_id to get the values. Figure 4 shows how the first function works, and Figure 5 shows how to get the feature with the item_id. Once the item is found, a column is created and the values are mapped to each row base on its subject id. The data is exported to a csv file format. Table 1 shows table name and a brief description in the MIMIC-III database.

Table 1Mimic 3 tables descriptions (source Mimic 3 website)

Table	Description
ADMISSIONS	This table contains every unique hospitalization identified by hospital admission ID (HADM_ID) for each patient
PATIENTS	This table defines SUBJECT_ID which is for every unique patient.
ICUSTAYS	Contains every unique ICU stay which is defined by an ICUSTAY_ID.
CHARTEVENTS	All charted observations for patients are contained in this table.
INPUTEVENTS_CV	This stores the intake for patients monitored using the CareVue system during their stay in the ICU.
INPUTEVENTS_MV	This is for all the intake for patients monitored using Metavision system while in they were in the ICU.
OUTPUTEVENTS	This table, like the input events, contains output data for patients during ICU stay.
PROCEDUREEVENTS_MV	It contain all procedures for the subset of patients in the ICU who were monitored using MetaVision.
DIAGNOSES_ICD	This is considered as a dictionary for all the ICD9_CM codes .
LABEVENTS	All laboratory measurements for all patients in the database are stored in this table.
D_ITEMS	This is a dictionary of all the ITEMIDs in the database, with and exception to those that are related to the laboratory tests.
D_LABITEMS	Like the d_items, this table is a dictionary of ITEMIDs for laboratory tests only.
CALLOUT	When a patient was cleared for both ICU discharge as well as hospital discharge, the information is stored in this table.
SERVICES	The service under which a patient is currently or was previously registered. (clinical)
TRANSFERS	This provides information about patient bed to bed transfers within the hospital/ICU or discharged.
CAREGIVERS	it keeps details of the caregivers who were responsible for recording data in the database.
CPTEVENTS	It keeps codes for procedures done on various patients using CPT_codes.
DRGCODES	In order for the hospital to properly bill it keeps codes for Diagnosis Related Groups (DRG) in this table.
NOTEEVENTS	This table stores all the de-identified notes captured from different departments of the hospital.
DATETIMEEVENTS	For all the observations like dates and time are recorded in this table.
MICROBIOLOGYEVENTS	For items microbiology measurements and sensitivities recorded from the hospital are kept in this table.
PRESCRIPTIONS	This is for all patient medications that were ordered, and even if not administered.
PROCEDURES_ICD	Using ICD code all patient procedures details are stored in this table.
D_CPT	This is a dictionary of all the CPT codes used in MIMIC-III.
D_ICD_DIAGNOSES	Dictionary of ICD codes that are relating to diagnoses.
D_ICD_PROCEDURES	This defines all the ICD codes relating to procedures.

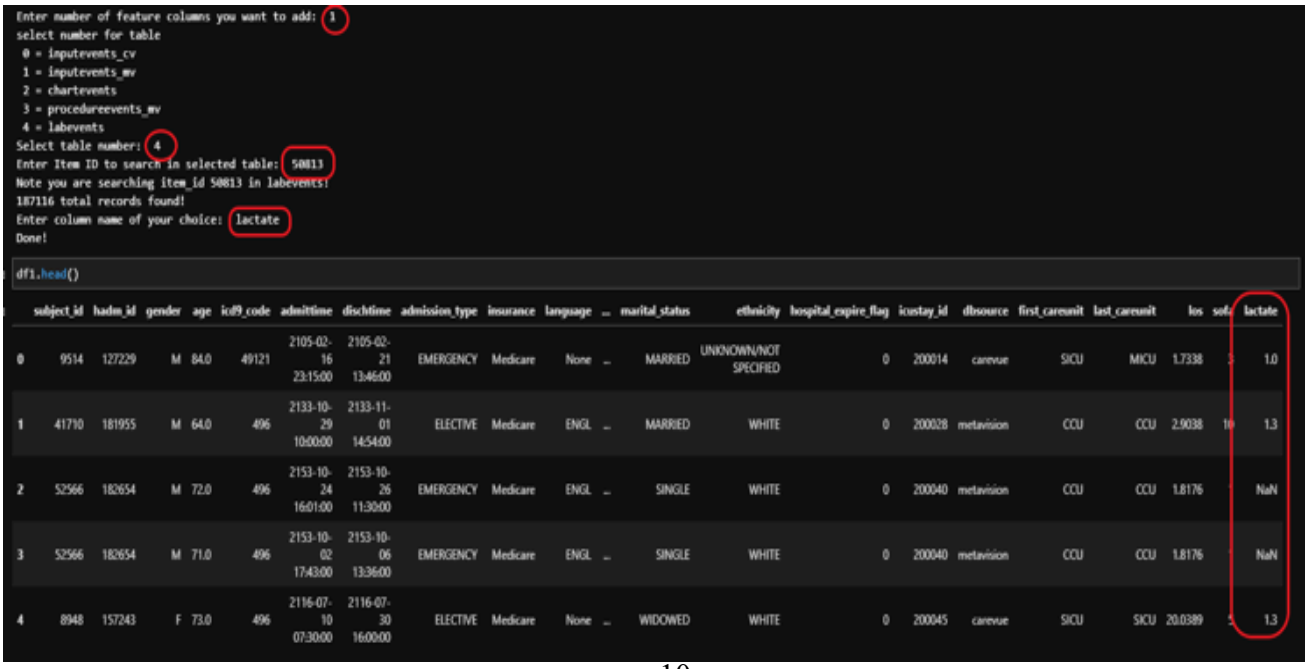
```
select number for table
 0 = d_items
 1 = d_labitems

Select table number:  1
Enter substring for the item to search: Lactate
Note you are searching item_id Lactate in d_labitems!
5 total records found!
```

	row_id	itemid	label	fluid	category	loinc_code
0	14	50813	Lactate	Blood	Blood Gas	32693-4
1	44	50843	Lactate Dehydrogenase, Ascites	Ascites	Chemistry	2531-2
2	155	50954	Lactate Dehydrogenase (LD)	Blood	Chemistry	2532-0
3	215	51015	Lactate Dehydrogenase, CSF	Cerebrospinal Fluid (CSF)	Chemistry	2528-8
4	254	51054	Lactate Dehydrogenase, Pleural	Pleural	Chemistry	2530-4

Figure 4 Search for item_id in the database

Figure 5 Searching item values (features) and adding feature column to the data frame



	subject_id	hadm_id	gender	age	icd9_code	admittime	dischtime	admission_type	insurance	language	...	marital_status	ethnicity	hospital_expire_flag	icustay_id	dbsource	first_careunit	last_careunit	los	sofa	lactate
0	9514	127229	M	84.0	49121	2105-02-16 23:15:00	2105-02-21 13:46:00	EMERGENCY	Medicare	None	...	MARRIED	UNKNOWN/NOT SPECIFIED	0	200014	carevue	SICU	MICU	1.7338	3	1.0
1	41710	181955	M	64.0	496	2133-10-29 10:00:00	2133-11-01 14:54:00	ELECTIVE	Medicare	ENGL	...	MARRIED	WHITE	0	200028	metavision	CCU	CCU	2.9038	10	1.3
2	52566	182654	M	72.0	496	2153-10-24 16:01:00	2153-10-26 11:30:00	EMERGENCY	Medicare	ENGL	...	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	1	NaN
3	52566	182654	M	71.0	496	2153-10-02 17:43:00	2153-10-06 13:36:00	EMERGENCY	Medicare	ENGL	...	SINGLE	WHITE	0	200040	metavision	CCU	CCU	1.8176	1	NaN
4	8948	157243	F	73.0	496	2116-07-10 07:30:00	2116-07-30 16:00:00	ELECTIVE	Medicare	None	...	WIDOWED	WHITE	0	200045	carevue	SICU	SICU	20.0389	5	1.3

4 Experiments

In this chapter, we will present a brief description of our sample dataset, and the models that were developed from the dataset. We used COPD for this study to predict mortality for between 1 to 28 days of ICU admission. We conducted two experiments with the same dataset. In experiment one, we used exactly the same features extracted using the tool. In experiment two, we added extra features based on some calculations on the extracted data.

4.1 Cohort

Data for the study was determined using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes for COPD. (COPD '490', '4910', '4911', '4912', '49120', '49121', '49122', '4918', '4919', '492', '4920', '4928', '494', '4940', '4941', '496',) [31], [32]. Basically, MIMIC-III contains a lot of records, in this study, the data was extracted for all the patients admitted in the ICU with COPD as morbidity, and from this cohort we also identified patients that were diagnosed with other underlying comorbidities.

Apart from comorbidities, COPD, medications, laboratory tests and other clinical variables were extracted or calculated: BMI calculated from weight and height, age calculated from the date of birth, first day of hospital admission, and the Elixhauser Comorbidity Index - calculated based on the same weights that were used in the study [33] using Van Walraven (VW).

4.2 Statistical analysis

We conducted statistical analyses on the data extracted. Continuous data were expressed as median (IQR) and were evaluated by two-tailed Student's t-test, while categorical data were expressed as frequencies (%) and were analyzed by the Chi-square tests. The t-test tells us how significant the differences between groups are. In other words, the t-test informs us if these differences (which are measured in means) could have occurred by chance. The bigger the t-value, the more likely it is that the results are repeatable. The demographic information and mortality status were presented in Table 1. We had the following scoring model-building process.

Table 2 Demographic data and outcomes regarding to mortality in COPD patients

Variable Median (IQR); (n, %)	Total (N=1,358)	Dead (N=310)	Survive (N=1048)	P value
Age	72.0 (63.0-80.0)	75.0 (67.0-82.0)	70.0 (62.0-79.0)	0.0192*
Sex (Male)[c]	748 (55.1)	163 (52.6%)	585 (55.8%)	0.3137
Comorbidity Status (Elixhausr>0)	145 (10.7%)	37 (11.9%)	108 (10.3%)	0.2405
Major Comorbidity				
CAD	539 (39.7%)	118 (38.1%)	421 (40.2%)	0.5053
HTN	755 (55.6%)	150 (48.4%)	605 (57.7%)	0.0036** *
CKD	354 (26.1%)	109 (35.2%)	245 (23.4%)	<0.0001 **
Cancer	083 (06.1%)	030 (9.7%)	53 (5.1%)	0.0029** *
DM	350 (25.8%)	80 (25.8%)	270 (25.8%)	0.9878
Clinical Values				
BMI	027.2 (22.7-32.5)	25.8 (21.1-30.8)	27.7 (23.4-33.1)	<0.0001 **
HbA1c	6.4 (6.4-6.4)	6.4 (6.4-6.4)	6.4 (6.4-6.4)	0.1954
FPG	139.4 (119.0-139.4)	139.4 (121.0-150.0)	139.4 (118.0-139.4)	0.0026** *
SBP	116.0 (102.0-133.0)	110.0 (97.0-126.0)	117.0 (104.0-135.0)	<0.0001 **
PIP (N=1335)	19.8 (16.0-24.0)	20.0 (16.0-27.0)	19.8 (16.0-23.0)	0.0001** *
MAPS (N=1332)	9.0 (8.0-11.0)	9.5 (8.0-12.0)	9.0 (7.0-10.0)	<0.0001 **
WBC	10.7 (8.0-14.6)	11.3 (8.0-16.3)	10.6 (8.1-14.3)	0.0424*
Neutrophil (N=1,199)	81.4 (73.4-88.0)	84.9 (75.9-90.4)	80.4 (73.0-87.0)	<0.0001 **
Lactate	1.5 (1.0-2.1)	1.7 (1.1-2.6)	1.4 (1.0-2.0)	<0.0001 **
CREAT	1.0 (0.7-1.4)	1.2 (0.7-2.0)	1.0 (0.7-1.4)	<0.0001 **
SOFA Score	5.0 (3.0-7.0)	5.5 (3.0-9.0)	4.0 (3.0-6.0)	<0.0001 **
Medication				
Norepinephrine	463 (34.1%)	184 (059.4%)	279 (26.6%)	<0.0001 **

Epinephrine	140 (10.3%)	035 (11.3%)	105 (10.0%)	0.5179
Vasopressin	150 (11.1%)	91 (29.4%)	059 (05.6%)	<0.0001**
Outcomes				
ICU length of stay	3.6 (2.0-6.5)	5.0 (2.5-8.5)	3.4 (2.0-6.0)	<0.0001**
Respiratory failure	1040 (76.6%)	260 (83.9%)	780 (74.4%)	0.0006**
Renal replacement therapy	69 (05.1%)	25 (8.1%)	44 (4.2%)	0.0065**

[c] Chi-square test. Mann-Whitney U test. *p<0.05, **p<0.01. CAD: Coronary artery disease; HTN: Hypertension; CKD: Chronic kidney disease; DM: diabetes mellitus; BMI: Body mass index; HbA1c: hemoglobin A1c; FPG: Fasting plasma glucose; SBP: Systolic blood pressure; PIP: Peak inspiratory pressure; MAPS: Mean airway pressure; WBC: White blood cell; CREAT: Creatinine; SOFA score: the sequential organ failure assessment score; ICU: intensive care unit

Variable	Univariable		Multivariable Model 1	
	OR (95% CI)	P value	OR (95% CI)	P value
Age	1.03 (1.02-1.04)	<0.0001*	1.04 (1.03-1.06)	<0.0001*
Sex (Male)	0.88 (0.68-1.13)	0.3139	0.76 (0.55-1.07)	0.1164
DM	1.00 (0.75-1.34)	0.9878	1.06 (0.72-1.57)	0.7638
HTN	0.69 (0.53-0.89)	0.0037*	0.73 (0.52-1.03)	0.0722
CKD	1.78 (1.35-2.34)	<0.0001*	1.36 (0.89-2.09)	0.1562
Cancer	2.01 (1.26-3.21)	0.0034*	2.35 (1.28-4.30)	0.0057*
BMI	0.97 (0.95-0.99)	0.0002*	0.95 (0.93-0.97)	<0.0001*
FPG	1.01 (1.00-1.01)	0.0002*	1.01 (1.00-1.01)	0.0009*
SBP	0.98 (0.98-0.99)	<0.0001*	0.99 (0.99-1.00)	0.0369*
PIP	1.05 (1.03-1.06)	<0.0001*	1.03 (1.01-1.05)	0.0145*

MAPS	1.14 (1.10-1.18)	<0.0001*	1.10 (1.05-1.16)	0.0002*
WBC	1.03 (1.01-1.05)	0.0005*	1.02 (0.99-1.04)	0.2249
Neutrophil	1.01 (1.00-1.02)	0.3334		
LACTATE	1.36 (1.25-1.48)	<0.0001*	1.20 (1.08-1.35)	0.0012*
CREAT	1.28 (1.16-1.42)	<0.0001*	1.11 (0.95-1.30)	0.1861
SOFA Score	1.14 (1.10-1.19)	<0.0001*	1.00 (0.95-1.06)	0.9498
Norepinephrine	4.03 (3.09-5.25)	<0.0001*	2.12 (1.46-3.09)	<0.0001*
Vasopressin	6.96 (4.86-9.97)	<0.0001*	2.25 (1.39-3.63)	0.0009*
ICU length of stay	1.06 (1.03-1.08)	<0.0001*	1.03 (1.00-1.06)	0.1072
Respiratory failure	1.79 (1.28-2.49)	0.0006*	1.66 (0.94-2.91)	0.0797
Renal replacement	2.00 (1.20-3.33)	0.0074*	1.20 (0.58-2.47)	0.6301

Table 3 Univariable and Multivariable Regression Analyses of Mortality

[c] Chi-square test. Mann-Whitney U test. *p<0.05, **p<0.01. CAD: Coronary artery disease; HTN: Hypertension; CKD: Chronic kidney disease; DM: diabetes mellitus; BMI: Body mass index; HbA1c: hemoglobin A1c; FPG: Fasting plasma glucose; SBP: Systolic blood pressure; SPO^2: oxygen saturation; PIP: Peak inspiratory pressure; MAPS: Mean airway pressure; WBC: White blood cell; CREAT: Creatinine; ICU: intensive care unit

Model 1 adjusted for age, gender & diabetes status.

First, the univariate analyses for all variables were performed and identified the variables with *p*-value <0.05. Second, all potential variables identified in the first step were entered into multivariate analysis simultaneously to adjust for the potential confounding effects, and only *p*-values <0.05 were retained in the multivariate model (Table 2). We only retained the variable with *p*-values <0.05 in the multivariate model in the scoring model. The scoring model was built according to clinical stand or receiver operating characteristic curve (ROC) detailed in Appendix

A. The ability of the scoring model in distinguish mortality status were performed via ROC and presented in Figure 2 below.

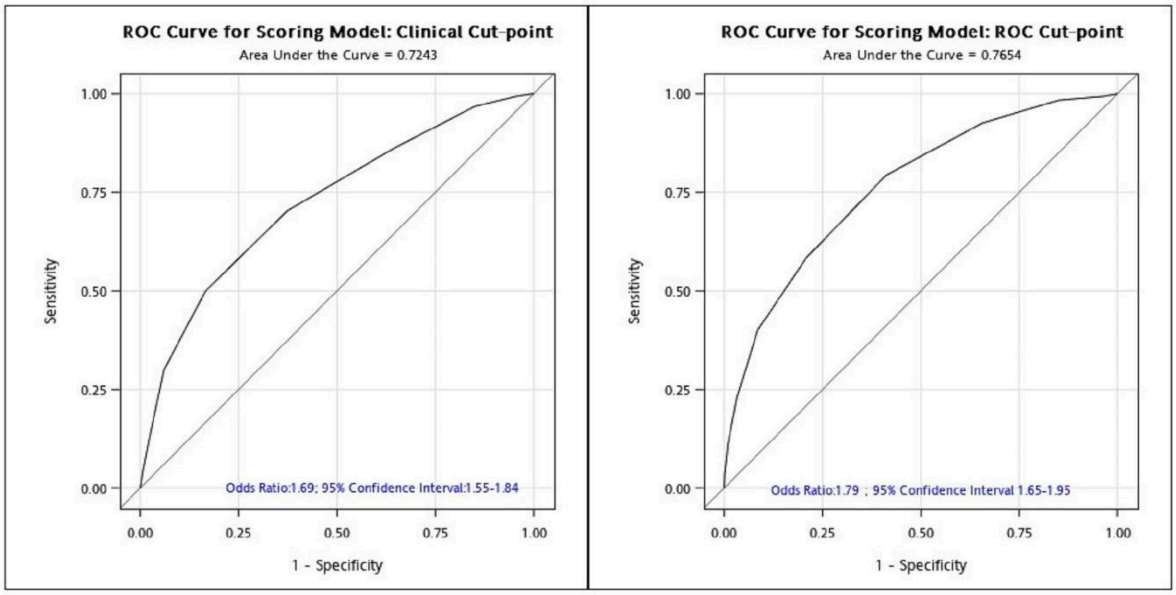


Figure 6 ROC curve for Scoring Models

4.3 Inclusion Criteria

Study participants were all the patients who were above 18 years and in this study, we only included the last record and the last day of admission to avoid having duplicated patient ID's and to retain the actual outcome of their survival status. The recruitment procedure was that we first selected patients based on their length of stay in ICU; we excluded patients who stayed for less than 1 day and longer than 28 days. We then removed all data with > 15% of missing values. A study by H. Zhang found that MIMIC-III database can have as high as 74% of missing

data [34]. In their study, N. Hou et al. [15] also used XGboost for predicting 30-days mortality for MIMIC-III ; it removed all data with more than 20% missing variables. In this study, we tried to keep as much real values as possible, hence removing all the records with more than 15% missing values. Figure 7 provides a detailed procedure on how we enrolled our participants for the study.

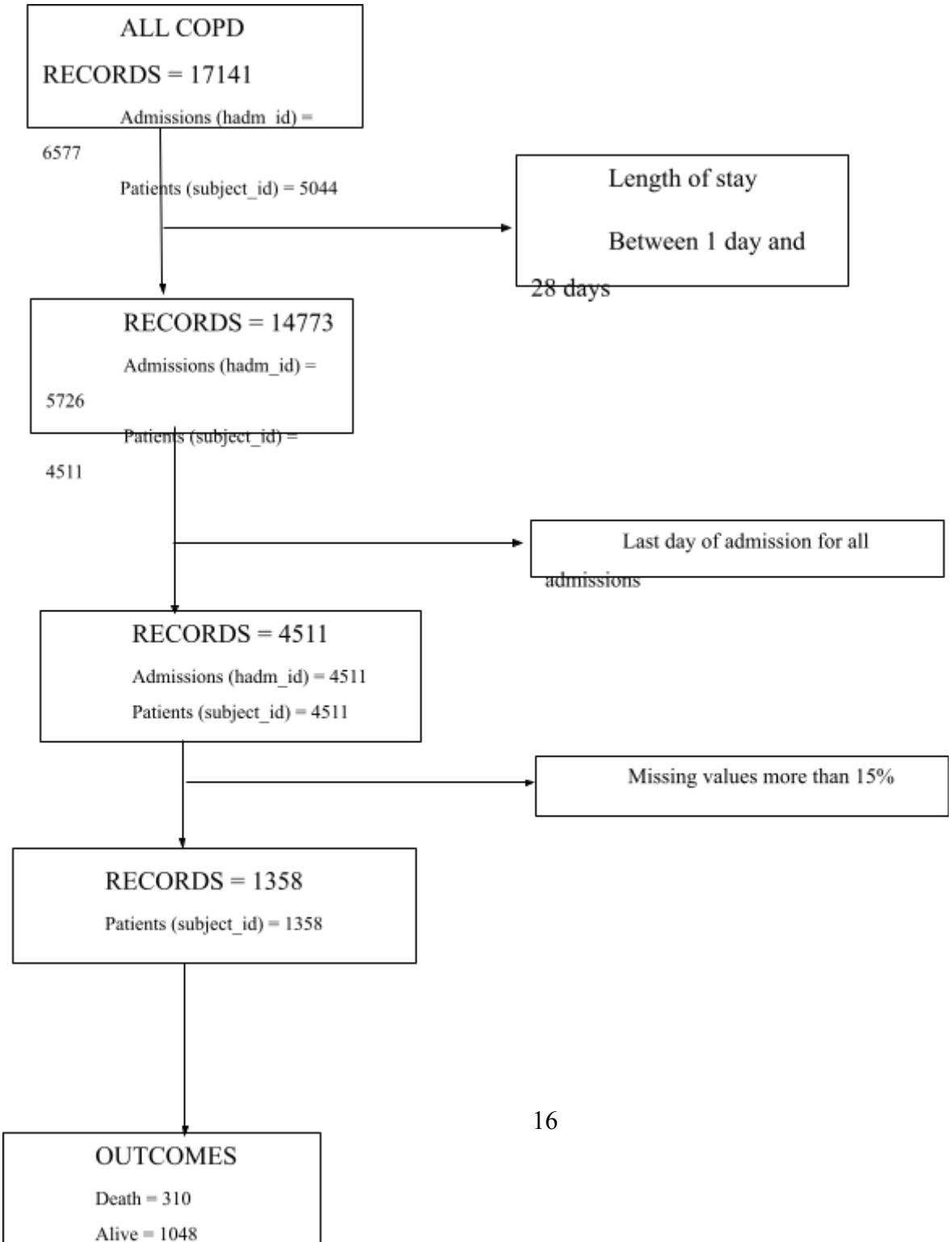


Figure 7 Recruitment Inclusion Criteria

4.4 Experiment 1

4.4.1 Basic model

In this experiment, we developed and trained an XGBoost model from the sample data we extracted. We extracted data for 5044 patients whose primary admission was COPD. The data was filtered based on length of stay (between 1 to 28 days in ICU) age greater than 18, and then we removed all those with more than 15% of missing values. The final cohort was 1358 patients and there are 19 features that were used for this training. The data was divided into 80% training data and 20% testing data. The model was interpreted by SHAP [21] which is currently being used in most XGBoost models as witnessed in [16], [17], [35], [36].

4.4.2 Model performance and Explanation

The model performance is good, with AUROC of 0.836 as depicted in Figure 8. Model explanation was done using SHAP; In Figures 9 to 11, we explain how we interpreted the model in deferent ways using SHAP. To validate the model performance further, we perfomed a 5-Fold cross validation. K fold cross validation is a statistical method used to compare and select a model for a given predictive problem. This method randomly divides the data into the number of folds specified (k-groups), one of the folds is used as validation while the other folds are combined and regarded as a training set. The process is repeated until all the groups are used as validation sets and also training set. The result for 5-fold was 83.05% accuracy with standard deviation of 2.05%.

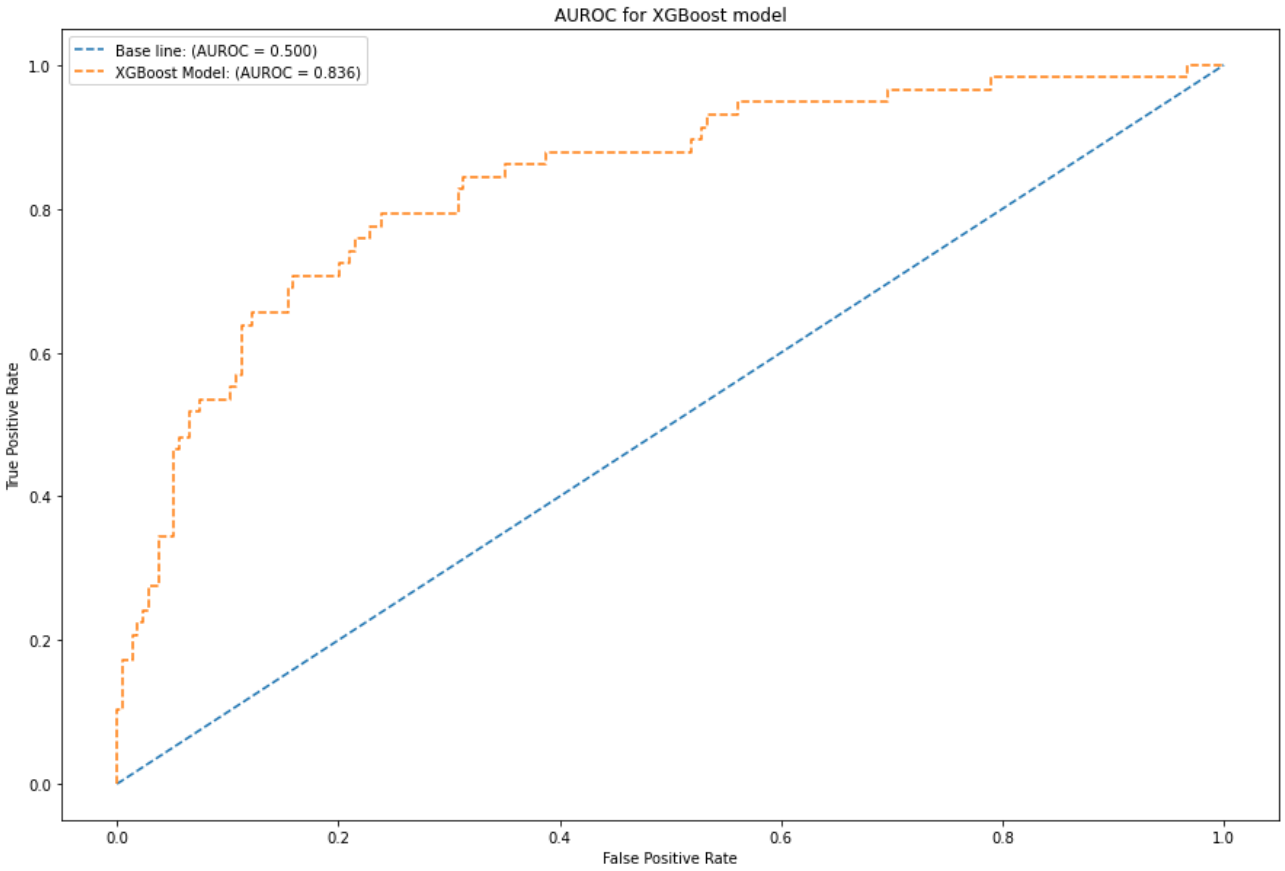
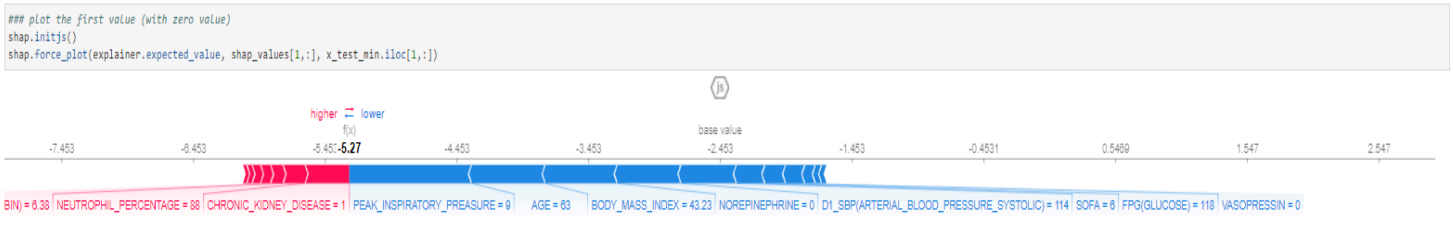


Figure 8 AUROC: Area Under Receiver Operating Curve for Basic model.

First, we look at a single feature explanation on how the features contributed to the decision in prediction of COPD mortality. Below is a figure that shows the impact each feature pushes the model output to a base value. Features pushing the prediction higher are shown in red, while features pushing the prediction lower are in blue. In the first plot, we can see how our top feature "PEAK_INSPIRATORY_PREASURE" (PIP) which is in blue, contributed towards the prediction of mortality of the patient. Its value pushed the risk lower while the same feature in the second plot, its values pushed the risk higher. In this case, we can say the higher the PIP, the higher the risk of not surviving in COPD and the lower its value, the lower the risk. The same can be said to the other features in the plot.



After looking into the single value impact, we visualize an accumulative impact of all the feature values using summary plot. The summary plot sorts features by the sum of SHAP value magnitudes over all samples. It then uses SHAP values to show the distribution of the impact that every single feature has on the output of model. Just like the single instance, colors shown in the plot represents the feature impact in which red is high and blue is low.

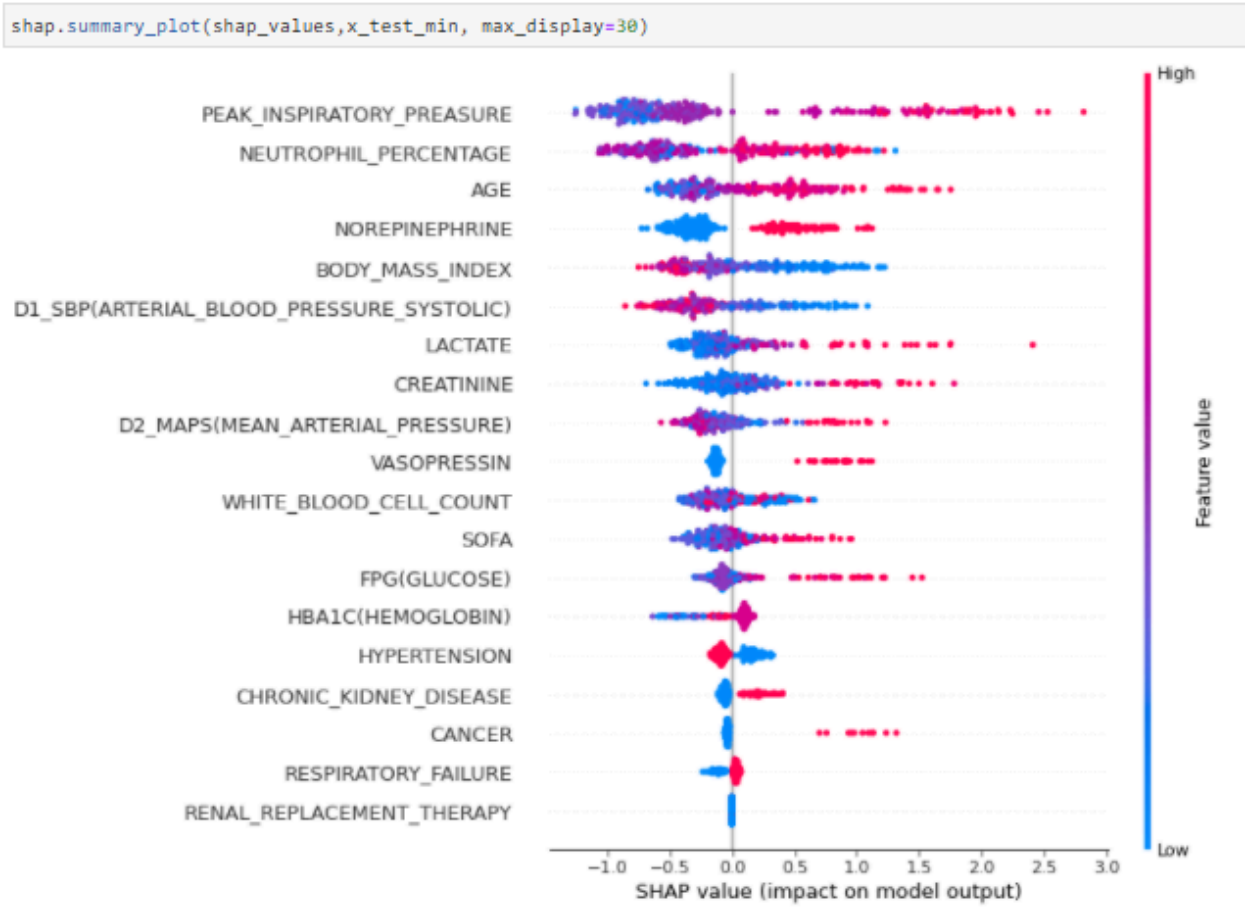


Figure 10 SHAP Value impact on model output

In an average feature importance plot, SHAP calculates the average impact of the feature values and plots the global importance of each feature. In Figure 8, the features were grouped into 6 domains, and the domains were also weighed based on the sum of the features and feature importance they had. The lab data domain was higher than other domains. Of all the features, Peak Inspiratory Pressure was the highest.

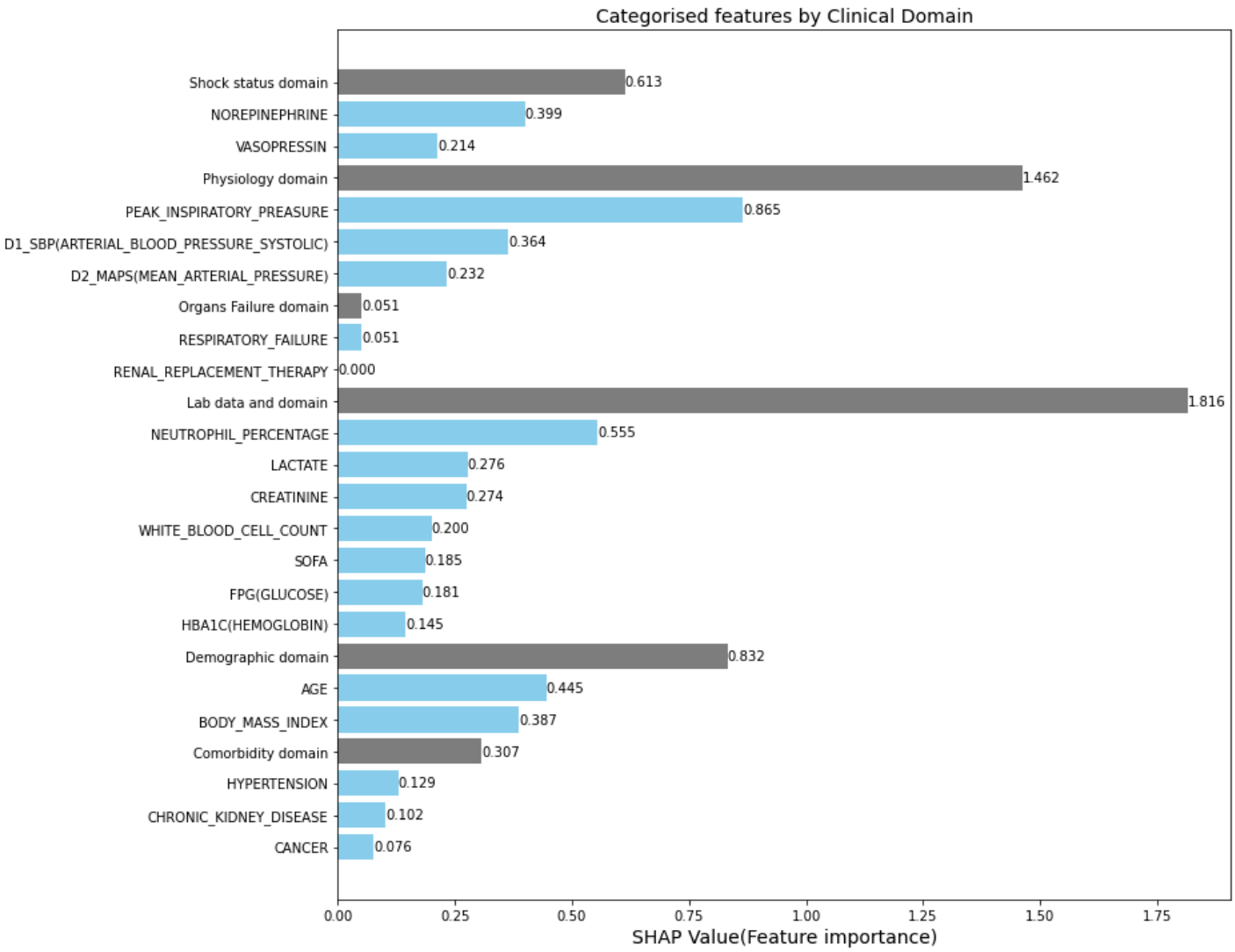


Figure 11 Average SHAP Value Importance for XGBoost with basic feature

4.5 Experiment 2

4.5.1 Extra-Features Model

In our basic model, we used 19 features that were extracted with our tool. For this experiment, we added more features to the existing features and we had a total of 27 features.

Some of the additional features include Elixhauser Comorbidity Index and other medications. We also included Clinical stand and Receiver Operating Characteristic curve (ROC) cut points - which are two scoring methods.

4.5.2 Model performance and Explanation

After adding new features, the models were assessed on performance using AUROC, Accuracy and F1-score. XGBoost outperformed the other models in both scores. Figure 9 shows the AUROC for all 4 models.

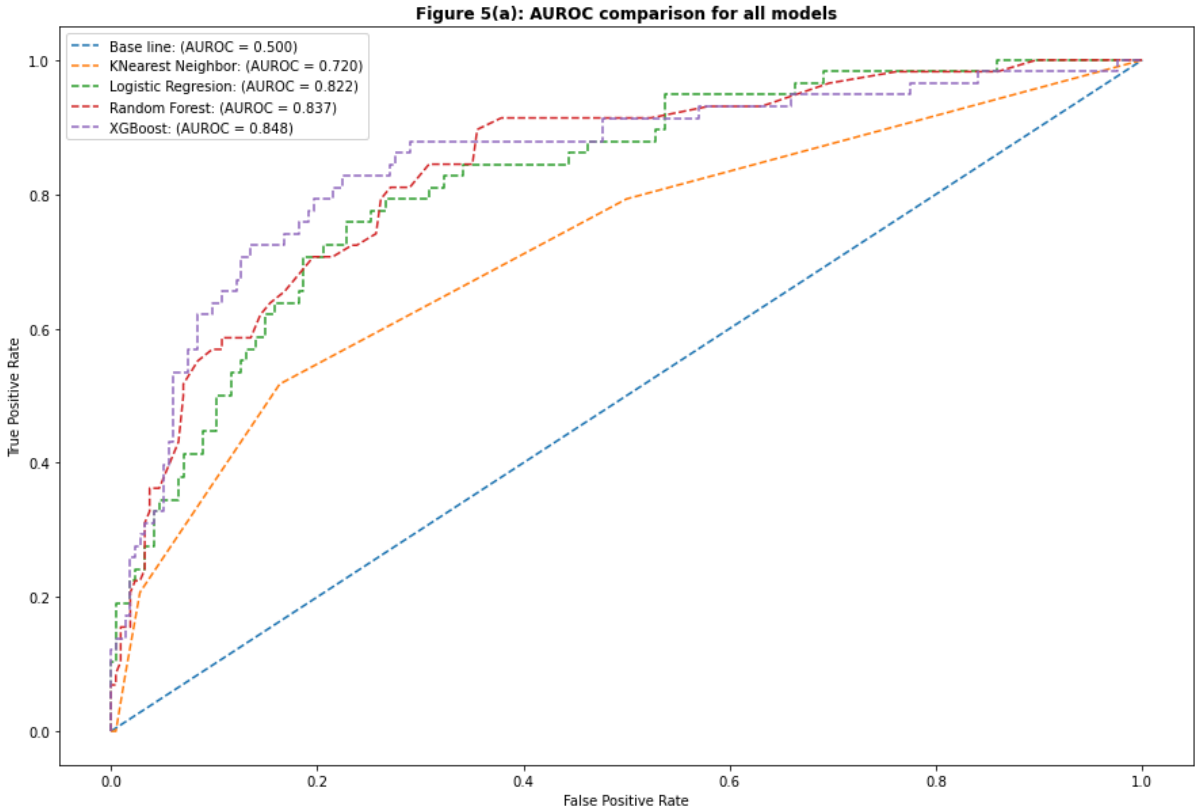


Figure 12 AUROC: Area Under Receiver Operating Curve for all models tested on extra featured dataset. (KNN. LR, RF, XGBoost)

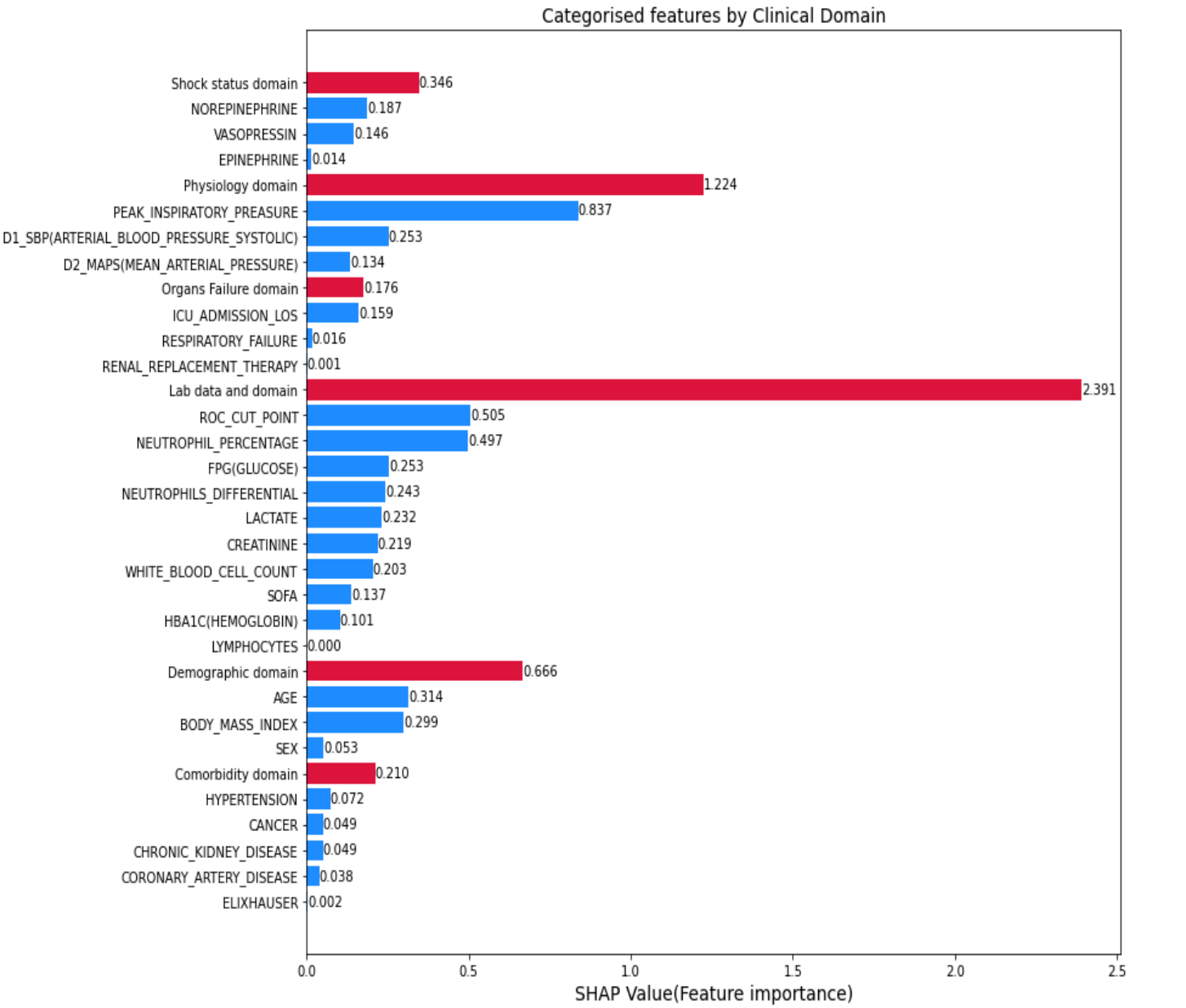

Categorised features by Clinical Domain
Shock status domain
NOREPINEPHRINE
VASOPRESSIN
EPINEPHRINE
Physiology domain
PEAK_INSPIRATORY_PREASURE
D1_SBP(ARTERIAL_BLOOD_PRESSURE_SYSTOLIC)
D2_MAPS(MEAN_ARTERIAL_PRESSURE)
Organs Failure domain
ICU_ADMISSION_LOS
RESPIRATORY_FAILURE
RENAL_REPLACEMENT_THERAPY
Lab data and domain
ROC_CUT_POINT
NEUTROPHIL_PERCENTAGE
FPG(GLUCOSE)
NEUTROPHILS_DIFFERENTIAL
LACTATE
CREATININE
WHITE_BLOOD_CELL_COUNT
SOFA
HBA1C(HEMOGLOBIN)
LYMPHOCYTES
Demographic domain
AGE
BODY_MASS_INDEX
SEX
Comorbidity domain
HYPERTENSION
CANCER
CHRONIC_KIDNEY_DISEASE
CORONARY_ARTERY_DISEASE
ELIXHAUSER
0.346
0.187
0.146
0.014
1.224
0.837
0.253
0.134
0.176
0.159
0.016
0.001
2.391
0.505
0.497
0.253
0.243
0.232
0.219
0.203
0.137
0.101
0.000
0.666
0.314
0.299
0.053
0.210
0.072
0.049
0.049
0.038
0.002
0.0
0.5
1.0
1.5
2.0
2.5
SHAP Value(Feature importance)

Table 4 Model comparison results (AUROC, Accuracy, and F1-score)

Model	Accuracy	AUROC	F1-Score
LR	0.81	0.821	0.65
RF	0.83	0.824	0.67
KNN	0.82	0.730	0.62
XGBoost (basic features)	0.82	0.833	0.71
XGBoost (More features)	0.84	0.847	0.75

5 Discussion

A MIMIC-III data extraction tool was successfully developed which can extract data and has some functionalities that make it easy for researchers to get information without spending a lot of time. As stated before, some researchers [3], [9], [10], [37] have done similar projects. But in their research [9], [10] developed web-based tools on MIMIC-II to visualize and extract data from the database, while our work is on the MIMIC-III database. S. Wang et al. [37] developed a data extraction tool on MIMIC-III, but this tool requires a lot of computation power with a minimum requirement of 50GM RAM, and it takes between 5 to 10 hours. This means that researchers with very minimal computation resources will not be able to utilize such an important tool. Our tool targets any researcher with a computer of at least 4GB RAM, and it takes less than 1 hour to produce basic cohort.

In the study, we selected patients who were admitted to ICU with COPD as a primary reason, and mortality prediction models were. Two experiments were done, the first one was with basic features extracted using our extraction tool. This model had a good performance; for all the performance matrics that we used to compare it with in the other models that were developed in the second experiment, it came second to our main model on AUROC and F1-score (AUROC: 0.833 ACCURACY: 0.820, F1-score: 0.71)

Comparison of the models was done and we ranked the performance based Area Under the Receiver Operating Characteristic (AUROC) curve analysis and we also checked accuracy as well as F1-score. We found that AUROC for predicting 28-day mortality in XGBoost (AUROC: 0.847 ACCURACY:0.84 F1-score: 0.75) was better than the other machine learning models. Comparison results are shown in detail in table 4 above.

On the machine learning models experiments, our main focus when comparing the models was on AUROC and F1-score. XGBoost performed better on both AUROC and F1-score. Using AUROC to measure model performance is considered to be the best way to measure a binary classification model than using accuracy. Accuracy in most cases is based on probability while AUROC is how well a model can classify the outcome. Using accuracy on imbalanced dataset is not a good practice because the model might just as well predict the

majority class only while missing the minor class which in case of mortality is dangerous. Since our dataset was unbalanced, we opted to use the AUROC to rank the model performance between our models developed.

Further more, we performed K-Fold cross validation on all the models, still our method come out with the best mean accuracy in comparison with the other models. We applied 5-Fold cross validation so that we could determine which model will come out to be the best. KNN had an Accuracy: 78.35% and standard deviation of (1.55%), LR Accuracy: 82.70% standard deviation (0.69%), RF Accuracy: 82.25% standard deviation (1.74%) and XGBoost scored an Accuracy: 82. 99% standard deviation (1.96%)

The following are the parameters used in XGBoost: learning_rate =0.025, n_estimators=400, scale_pos_weight=0.95, max_depth=12, min_child_weight=4, max_delta_step =10, subsample=0.9, colsample_bytree=1. Other parameters were left to be default parameters.

6 Conclusion

In this study, we proposed an extraction tool that can be used for extracting data from the MIMIC-III database effectively. This tool is very easy to use and will help to solve the problem of having unreproducible studies. By being able to easily pull data from MIMIC-III researchers will be able to reproduce results from other studies. This is not only for reproducing studies but also for new studies - which reduces the time that is spent on writing queries to pull data from the database. The tool developed in this research was able to extract data based on user inputs without them having to write any queries and, in the end, produced a data frame that can be worked upon for analysis and transformation.

Five models for predictions of 28-day mortality in ICU for patients with COPD were developed and were compared based on AUROC, Accuracy and F1-scores. Among the five models, XGBoost model which had more features added, outperformed other machine learning algorithms in both AUROC curve and Model Accuracy - with an AUROC of 0.847. With the aid of SHAP, important features were discovered that would help in alerting health service providers in time to save lives. The prediction of mortality is important in healthcare and proper care of the patient can be done soon after the prediction is made in good time. For timely clinical intervention decisions the models that were developed in this study will provide the much required help to health personnel for COPD patients.

Future work is to the tool compatible with the new MIMIC IV database, so that users can select which database the want to use.

Acknowledgement

We are grateful to the Department of Critical Medicine, Taichung Veterans General Hospital for the time and guidance given on health-related requirements during the study.

References

[1] Johnson, Alistair, Pollard, Tom, and Mark, Roger, "MIMIC-III Clinical Database." PhysioNet, 2015. doi: 10.13026/C2XW26.

[2] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

[3] S. Wang, M. B. A. McDermott, G. Chauhan, M. C. Hughes, T. Naumann, and M. Ghassemi, "MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III," *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, Apr. 2020, doi: 10.1145/3368555.3384469.

[4] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web Data Extraction, Applications and Techniques: A Survey," *ACM Computing Surveys (Under Review)*, vol. 70, Jul. 2012, doi: 10.1016/j.knosys.2014.07.007.

[5] "Web Scraping Tool & Free Web Crawlers | Octoparse." https://www.octoparse.com/ (accessed Sep. 26, 2021).

[6] "myTrama." https://www.mytrama.info/es/ (accessed Sep. 26, 2021).

[7] "Hevo Data | Automated Data Pipelines to Redshift, BigQuery, Snowflake." https://hevodata.com/ (accessed Sep. 26, 2021).

[8] E. Stricker and M. E. Scheurer, "PDF Data Extractor (PDE) - A Free Web Application and R Package Allowing the Extraction of Tables from Portable Document Format (PDF) Files and High-Throughput Keyword Searches of Full-Text Articles," Jul. 2021. doi: 10.1101/2021.07.13.452159.

[9] D. J. Scott *et al.*, "Accessing the public MIMIC-II intensive care relational database for clinical research," *BMC Med Inform Decis Mak*, vol. 13, p. 9, Jan. 2013, doi: 10.1186/1472-6947-13-9.

[10] J. Lee, E. Ribey, and J. R. Wallace, "A web-based data visualization tool for the MIMIC-II database," *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, p. 15, Feb. 2016, doi: 10.1186/s12911-016-0256-9.

[11] A. Cavaillès *et al.*, "Comorbidities of COPD," *European Respiratory Review*, vol. 22, no. 130, pp. 454–475, Dec. 2013, doi: 10.1183/09059180.00008612.

[12] WHO, "Chronic obstructive pulmonary disease (COPD)," Jun. 21, 2021. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (accessed Jun. 29, 2021).

[13] G. Hillas, F. Perlikos, I. Tsiligianni, and N. Tzanakis, "Managing comorbidities in COPD," *Int J Chron Obstruct Pulmon Dis*, vol. 10, pp. 95–109, 2015, doi: 10.2147/COPD.S54473.

[14] W. Huang, R. Xie, Y. Hong, and Q. Chen, "Association Between Comorbid Chronic Obstructive Pulmonary Disease and Prognosis of Patients Admitted to the Intensive Care Unit for Non-COPD Reasons: A Retrospective Cohort Study," *International Journal of COPD*, vol. 15, pp. 279–287, Feb. 2020, doi: 10.2147/COPD.S244020.

[15] N. Hou *et al.*, "Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost," *Journal of Translational Medicine*, vol. 18, no. 1, p. 462, Dec. 2020, doi: 10.1186/s12967-020-02620-5.

[16] C.-A. Hu *et al.*, "Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan," *BMJ Open*, vol. 10, no. 2, p. e033898, Feb. 2020, doi: 10.1136/bmjopen-2019-033898.

[17] J. Liu, J. Wu, S. Liu, M. Li, K. Hu, and K. Li, "Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model," *PLOS ONE*, vol. 16, no. 2, p. e0246306, Feb. 2021, doi: 10.1371/journal.pone.0246306.

[18] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Oct. 05, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[19] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," *arXiv:1802.03888 [cs, stat]*, Mar. 2019, Accessed: Sep. 27, 2021. [Online]. Available: http://arxiv.org/abs/1802.03888

[20] Batunacun, R. Wieland, T. Lakes, and C. Nendel, "Using SHAP to interpret XGBoost predictions of grassland degradation in Xilingol, China," Earth and space science informatics, preprint, Jun. 2020. doi: 10.5194/gmd-2020-59.

[21] "Welcome to the SHAP documentation — SHAP latest documentation." https://shap.readthedocs.io/en/latest/index.html (accessed Sep. 03, 2021).

[22] "SHAP: How to Interpret Machine Learning Models With Python," *Better Data Science*, Nov. 09, 2020. https://medium.com/@radecicdario/shap-how-to-interpret-machine-learning-models-with-python-2323f5af4be9 (accessed Sep. 27, 2021).

[23] A. L. Lynam *et al.*, "Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults," *Diagnostic and Prognostic Research*, vol. 4, no. 1, p. 6, Jun. 2020, doi: 10.1186/s41512-020-00075-2.

[24] P. Theerthagiri, I. J. Jacob, A. U. Ruby, and Y.Vamsidhar, "Prediction of COVID-19 Possibilities using KNN Classification Algorithm," In Review, preprint, Nov. 2020. doi: 10.21203/rs.3.rs-70985/v2.

[25] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning," *Health Informatics J*, vol. 25, no. 3, pp. 811–827, Sep. 2019, doi: 10.1177/1460458217723169.

[26] "ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification," Mar. 01, 2019. https://www.cdc.gov/nchs/icd/icd9cm.htm (accessed Sep. 03, 2021).

[27] Jaime Rosenberg, "Chronic Kidney Disease in COPD Negatively Impacts Mortality, Other Patient Outcomes," *AJMC*, Jul. 30, 2019. https://www.ajmc.com/view/chronic-kidney-disease-in-copd-negatively-impacts-mortality-other-patient-outcomes (accessed Jul. 07, 2021).

[28] P. Rogliani, G. Lucà, and D. Lauro, "Chronic obstructive pulmonary disease and diabetes," *COPD Research and Practice*, vol. 1, no. 1, p. 3, Aug. 2015, doi: 10.1186/s40749-015-0005-y.

[29] W. M. Chatila, B. M. Thomashow, O. A. Minai, G. J. Criner, and B. J. Make, "Comorbidities in Chronic Obstructive Pulmonary Disease," *Proc Am Thorac Soc*, vol. 5, no. 4, pp. 549–555, May 2008, doi: 10.1513/pats.200709-148ET.

[30] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining Comorbidity: Implications for Understanding Health and Health Services," *Ann Fam Med*, vol. 7, no. 4, pp. 357–363, Jul. 2009, doi: 10.1370/afm.983.

[31] R. A. Deyo, D. C. Cherkin, and M. A. Ciol, “Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases,” *J Clin Epidemiol*, vol. 45, no. 6, pp. 613–619, Jun. 1992, doi: 10.1016/0895-4356(92)90133-8.

[32] H. Quan *et al.*, “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data,” *Med Care*, vol. 43, no. 11, pp. 1130–1139, Nov. 2005, doi: 10.1097/01.mlr.0000182534.19832.83.

[33] N. R. Thompson *et al.*, “A New Elixhauser-based Comorbidity Summary Measure to Predict In-Hospital Mortality,” *Medical Care*, vol. 53, no. 4, pp. 374–379, Apr. 2015, doi: 10.1097/MLR.0000000000000326.

[34] H. Zhang, M.-F. Balcan, and D. P. Woodruff, “Medical Missing Data Imputation by Stackelberg GAN,” p. 13.

[35] Na Serere, “Black Box Model Using Explainable AI with Practical Example,” *Analytics Vidhya*, Oct. 24, 2020. https://www.analyticsvidhya.com/blog/2020/10/unveiling-the-black-box-model-using-explainable-ai-lime-shap-industry-use-case/ (accessed Jun. 29, 2021).

[36] X. Liu *et al.*, “Interpretable Machine Learning Model for Early Prediction of Mortality in Elderly Patients with Multiple Organ Dysfunction Syndrome (MODS): a Multicenter Retrospective Study and Cross Validation,” p. 33.

[37] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Michael C. Hughes, Tristan Naumann, Geeticka Chauhan, Michael C. Hughes, Tristan Naumann, and Marzyeh Ghassemi, *MIMIC-Extract:A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III*. Healthy ML, 2021. Accessed: Sep. 13, 2021. [Online]. Available: https://github.com/MLforHealth/MIMIC_Extract

Appendix A

Clinical Cut-Point

Factors	Cut-point
Age[1]	65
BMI[2]	18.5-24.9
FPG[3]	100-125
SBP[4]	≤120
PIP	≤40
MAPS	≤40
LACTATE	≤2
Creat	Male:0.74-1.35 mg/dL Female: 0.59 to 1.04 mg/dL
ICU LOS	7

Score range from 0-10

Total Score	Frequency	Percent
0	2	0.15
1	3	0.22
2	44	3.24
3	118	8.69

[1] https://www.who.int/healthinfo/survey/ageing_mds_report_en_daressalaam.pdf
[2] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
[3] https://care.diabetesjournals.org/content/diacare/suppl/2019/12/20/43.Supplement_1.DC1/Standards_of_Care_2020.pdf
[4] https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.120.15026

4	249	18.34
5	332	24.45
6	281	20.69
7	173	12.74
8	93	6.85
9	45	3.31
10	18	1.33

ROC Cut-point

Factors	Cut-point	AUC
Age	75	0.6029
BMI	22.8	0.5803
FPG	149	0.5546

SBP	104	0.5962
PIP	26	0.5729
MAPS	15	0.5956
LACTATE	1.9	0.6011
CREAT	1.6	0.5797
ICU LOS	4	0.5875

Score range from 0-11

Total Score	Frequency	Percent
0	1	0.07
1	39	2.87
2	121	8.91
3	222	16.35
4	297	21.87
5	278	20.47
6	187	13.77
7	108	7.95
8	60	4.42
9	33	2.43
10	10	0.74
11	2	0.15