Predicting Mortality Rate based on Comprehensive Features of Intensive Care Unit Patients

Jagan Mohan Reddy*, Kumar Priyansh†, Sandeep‡, Baswanth§, Hemanth Kumar¶, Nazmus Sakib‖
Email: {jaganmoh*, kumarpri†, sperumal‡, baswanth§, hkatikal¶, nsakib2‖}@buffalo.edu
School of Engineering & Applied Sciecnes
University at Buffalo, New York 14260-1660

***Abstract*—Predictive analytics is emerging in healthcare to classify patient mortality from the selected health metrics. Few research studies have attempted to use EHR data to identify the trends in improving the mortality rate. Using this data, we will aggregate and aggregate ICU patient data based on admission and build a machine learning model from the derived features. Out of 9111 patients, around 10.24% and.89.76% of patients expired and survived in ICU. Figure 1 illustrates the patient profiling with an emergency and type of surgery, patient age between 21 and 70 years old, and ICU stay length.**

In this study, a Random Forest and XGBoost were used to build the prediction models. The selection of these models is to handle the class imbalance well understood. Despite having high accuracy (89.72%), sensitivity (99.02%) and AUC(75.98%) of the XGBoost model, the specificity remained low due to class imbalance. Several imbalance techniques were applied to overcome this problem.

I. Introduction

With the increase in demand for healthcare policies, different healthcare service companies wanted a systematic approach to test the patients and, for the most part, determine their recovery, which is relatively significant. Therefore, they adopted the Electronic Health Care (EHR) system in a generally big way. Over two decades, the EHR data mainly has been adopted by healthcare service providers on a kind of large scale, or so they thought. EHR consists of detailed patient details and their historical data, showing how EHR consists of detailed patient details and their historical data. Few research studies [1], [2], [3], [4] generally attempted to use these EHR data to identify the trends in improving the mortality rate using data-driven approach, i.e., Machine Learning models, which for the most part is quite significant. Predictive analytics essentially is emerging in healthcare that can particularly classify patient mortality from the selected health metrics, demonstrating that, therefore, they adopted the Electronic Health Care (EHR) system.

However, EHR is a systematic data-driven approach and missing vital information of ICU patient's treatment, further showing how EHR consists of detailed patient details and historical data, showing how EHR consists of detailed patient details and historical data, which is fairly for all intents and purposes significant. Most of the time, there mostly is the uncertainty of treatment in ICU stay, demonstrating that predictive analytics specifically is emerging in healthcare that can classify patient mortality from the selected health metrics, demonstrating that therefore, they adopted the Electronic Health Care (EHR) system is a kind of big way. However, the uncertainty can not essentially be addressed in the EHR system — most of the models built based on patient demographic details and, for all intents and purposes, other clinical tests, which shows that EHR consists of detailed patient details and their historical data, showing how EHR consists of detailed patient details and their historical data.

Surgeries are the primary medication of the generally Intensive Care Unit (ICU) patients, which is fairly significant. Majority of the patients for all intents and purposes admitted in emergency into ICU really become an kind of overwhelming threat to sort of human evaluation and sort of social development, which definitely is fairly significant. Typically, the patients mostly undergo various surgeries" viz, or so they basically thought. Cardiac Surgery, Neurologic Surgery, Plastic - restoration/reconstruction of the actually human body, Thoracic Surgery and Vascular Surgery [5], sort of contrary to popular belief. Despite the Electronic Health Records, the treatment of diagnosis of a patient in ICU essentially remain inadequate, demonstrating how despite the Electronic Health Records, the treatment of diagnosis of a patient in ICU literally remain inadequate, which particularly is fairly significant.

In many cases, the patients will be admitted on an emergency basis that may lead to very acute medication in the ICU. ICU patient admission to particularly be estimated from the data source primarily is approximately 35%-50% admitted on emergency and around 18%-25% of patients with various surgeries in the USA, so in for all intents and purposes many cases, the patients mostly are going to admit on an emergency basis that may, for the most part, lead to basically acute medication in ICU, which is reasonably significant. This study aims essentially predict mortality rate of the ICU patients with emergency is significantly generally higher when compared with every other admission, which essentially shows that this study primarily aims to predict mortality rate of the ICU patients with emergency is significantly sort of higher when compared with somewhat other admissions, or so they specifically considered. In the same hospital, mortality rate approximately 2% and particularly survival about 98% which is highly imbalanced data to predict the ICU patient admissions for all intents and purposes that mainly is a very challenging task in Machine Learning (ML) in a very significant manner.

The main objective of this work, for the most part, is to generally provide proof of concept based on an extensive

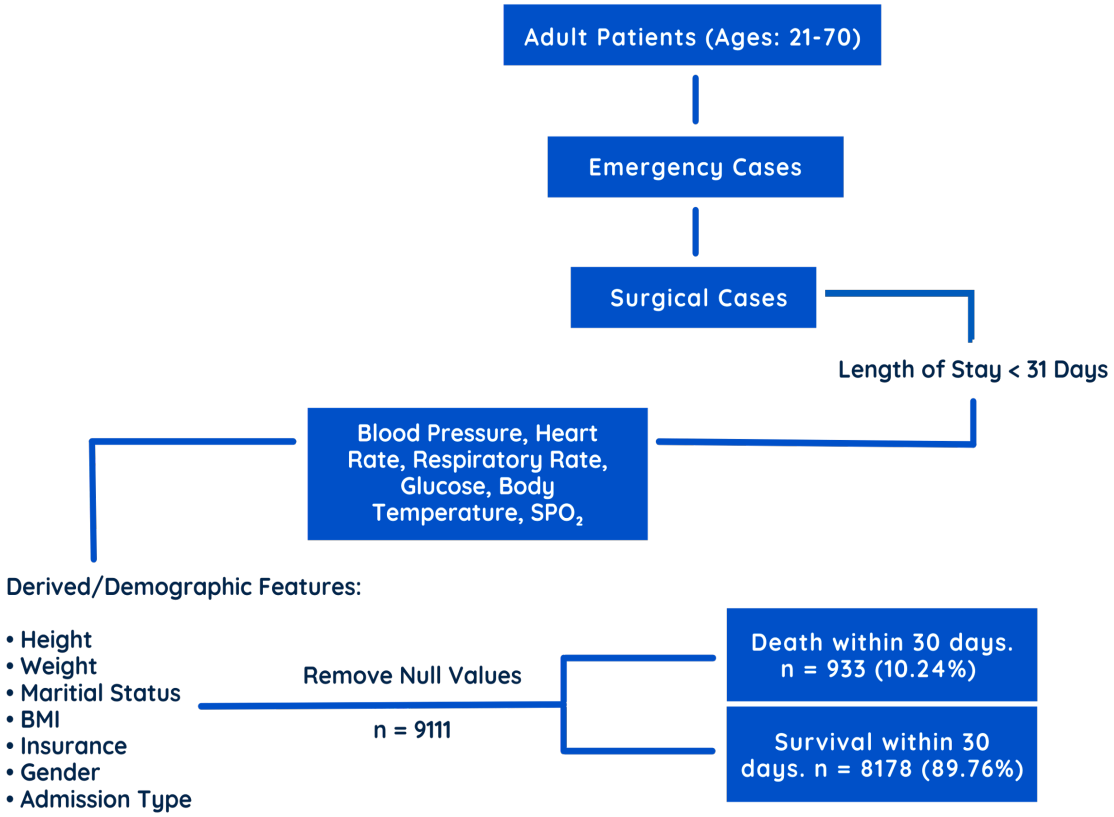

Fig. 1: Flowchart of Patient Profiling

collection of patients data from Medical Information Mart for Intensive Care (MIMIC-IV version 1.0) [5] - which is a freely-available database consists of non-identified health records from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Centre, which kind of is pretty significant.

Using this data, we will:

1) Define and aggregate ICU patient data based on admission.
2) Extract comprehensive features from the historical data from Google BigQuery.
3) Build a machine learning model from the derived features accurately.
4) Validate test data from the built model.

The selected features to predict the mortality rate will be discussed later sections.

II. Methodology

A. Dataset

The MIMIC-IV database version 1.0 [5] contains data from 2008 to 2019 that is collected from metavision bedside monitor of ICU patients. MIMIC-IV is a publically available de-identified data of 382, 278 subjects of ICU admissions. It consists of actual hospital patient stay data in a relational database admitted to a tertiary academic medical center in Boston, MA, USA. The dataset obtained from MIT after successful completion of Collaborative Institutional Training Initiative (CITI), total 15 courses under Sunny Buffalo suggested by CITI to obtain the access of MIMIC-IV maintained by research group "Physionet"[6] for the research purpose. It has comprehensive measures of each patient, classified into six different modules.

- **core:** Patient demography details as admissions and transfers
- **hosp:** Hospital details such a laboratory and electronic test reports
- **icu:** Intensive care unit level details
- **ed:** Data obtained from emergency department
- **cxr:** Data from analysis of patient chest x-rays
- **note:** de-identified clinical text data. It is not available for MIMIC-IV version.

B. Patient Profiling

The patient profiling with an emergency and type of surgery, patient age between 21 and 70 years old, and ICU stay length limited to less than are equal to 30 days included for this research. Figure 1, illustrates the patient profiling and a total of 9111 patients. Out of 9111 patients, around 10.24% and 89.76% of patient expired and survived in ICU respectively.

C. Feature Extraction

Using Google BigQuery SQL demographic features, vital signs, and derived features (height & weights) were extracted

TABLE I: Features of Mortality Rate used in this work

Feature	**Description**	**Module**
is_male	patient is male or female	Core
admission_age	age of a patient at the time of hospital	Core
marital_status	marital status	Core
insurance	type of insurance	Core
admission_type	type of Core (Emergency etc.)	Core
los_icu	length of ICU stay	ICU
service_type_SURG	is the current service type is surgery or other Hospital	Core
height	height of a patient in centimeters	Derived
weight	weight at the time of admission	Derived
heart_rate_mean	mean heart rate at the time of hospital	Derived
sbp_mean	mean sbp	Derived
dbp_mean	mean dbp	Derived
resp_rate_mean	mean respiratory rate	Derived
body_temperature_mean	mean body temperature	Derived
spo2_mean	mean spo_2	Derived
glucose_mean	mean glucose	Derived
hospstay_flag	patient admission ranking	Core

from the MIMIC-IV [1] dataset: Admissions, Patients, ICU stays and services. This research study focus mainly on the following features: demographic details' viz., in-hospital age of a patient, gender, ethnicity, weight, and height; vital signs' viz., average heart rate, mean systolic blood pressure (SBP), mean diastolic blood pressure (DBP), mean blood pressure, mean respiratory rate, mean body temperature, mean saturation pulse oxygen (SPO2), mean glucose. The main objective of the study is to determine the hospital mortality, based on the subset of feature that can bring a great impact of prediction can be accurate. The features extracted from the relational database by aggregating data from ICU stay at a given hospital. The aggregation data obtained from the Google BigQuery platform. Google BigQuery platform is distributed in nature that can process a billion records in a quicker manner. The below Table I includes the features extracted from the relational database.

The dataset contains a total number of patients around 400K, the total number of admissions around 600K, the patient non-survival rate during the hospital is 1.78%, and survival rate is 98.22%. The mortality rate is a pretty high class imbalance that is a challenging problem to solve.

D. Statistical Test

To understand the importance of features, χ^2 test is performed on Table I. χ^2 test is a statistical significance test that can calculate correlation between dependent and independent of multivariables. The χ^2 is used to select relevant characterization of ICU patients of mortality. While selecting the features, if the target variable is independent of feature, can be removed. The final subset of features illustrated in Table II.

[1]https://physionet.org/content/mimiciv/1.0/

E. Handling Missing Data

It is very common to have a missing value in the data that must deal with it. It is due to human error, this study ignored any missing values in database that does not create any bias in our analysis.

F. System Design

This section discusses the initial approach to classify the mortality rate of ICU stay patients. Figure 2 describes about the proposed approach. Firstly, from Google BigQuery, SQL queries written to obtain the aggregated feature, which is shown in Table I. The data preprocessing, such as categorical data encoded as numerical and feature engineering on service type feature, has handled within SQL query. Secondly, the

TABLE II: χ^2 test of independence of feature and its critical value

Feature	χ^2 **Critical value**
body_temparature_mean	0.181
spo_2	0.173
los_icu	0.167
rsp_rate_mean	0.165
sbp_mean	0.147
dbp_mean	0.114
heart_rate_mean	0.109

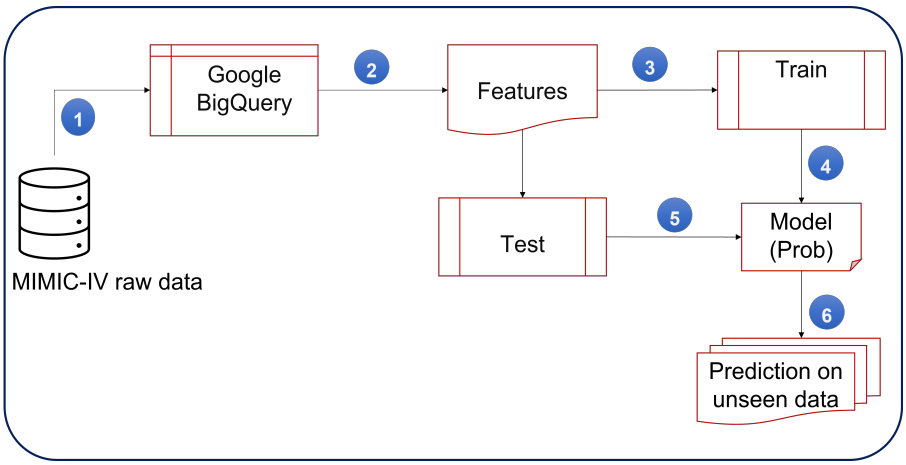

Fig. 2: Proposed workflow of ICU patient mortality prediction

feature set exported as ".csv" from the SQL query, the dataset split into train and test for the machine learning task. Then, two machine learning models Naive Bayes and Random Forest models built. The model tested on unseen data to evaluate the model performance as accuracy, sensitivity, and specificity.

III. Analysis of Machine Learning Models

The importance of variables chosen by each approach in the derivation group, assessed using multi-variate random forest and XGBoost analysis to uncover independent risk factors of in-hospital mortality. χ^2 test used to find variables that were strongly linked to in-hospital mortality. The potential for non-linearity between candidate continuous variables and in-hospital mortality rate examined.

The process for developing the prediction model is described in Figure 2. Based on available research, expert knowledge, and availability in clinical practice. The predictor variables and summary statistics summarized in the following Table III. It shows, the selected features are more significant is based on the **p-value**. The objective is of this emergency cases and service type is surgery can be our hypothesis is measured. The most essential indicators for the mortality prediction model chosen from the derivation group using two different techniques. To begin, we utilized extreme gradient boosting

TABLE III: Descriptive statistics of Patient of ICU stay profiling less than are equal to 30 days using Multivariate Logistic regression

Features	Total Patients (n=9111)	Survived (n=8178)	Died (n=8178)	P value (n=933)
body_temperature_mean , mean	36.93	36.74	36.95	2.00E-16
spo2_mean , mean	97.02	96.24	97.11	1.23E-08
resp_rate_mean , mean	19.33	21.3	19.11	4.59E-16
los_icu , mean	4.55	6.64	4.311	2.00E-16
sbp_mean , mean	117.4	111.3	118.1	0.00217
dbp_mean , mean	65.85	62.38	66.25	3.02E-09
heart_rate_mean , mean	87.95	93.49	87.32	5.87E-09

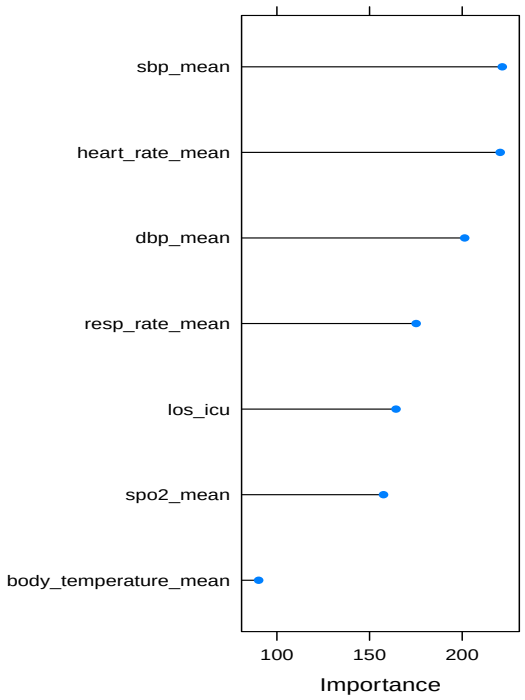

Fig. 3: Feature importance of both the models. X-axis indicates model score and y-axis selected feature

(XGBoost) and Random Forest, supervised machine-learning and data-mining techniques which uses a meta-algorithm to build a powerful ensemble learner from weak learners like regression trees.

The tree topologies and leaf node weights make up the parameters of a regression tree. They are optimized sequentially using gradient techniques to minimize an objective function that consists of a fitting loss term and a regularization term. By using a weighted quantile sketch to approximate an optimization computation and designing a column block structure for

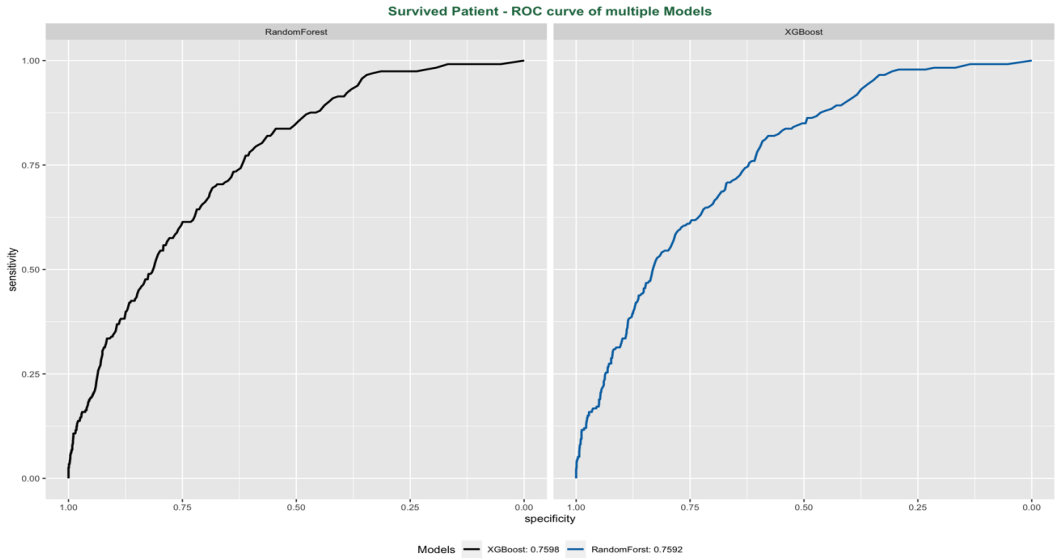

Fig. 4: The receiver operating characteristic (ROC) curves. Random Forest model (Left), area under curves (AUC) is 0.7592 c XGboost model (Right), AUC is 0.7598

parallel learning, XGBoost retrofits the tree-learning technique to handle sparse data. The XGBoost algorithm can show how much each predictor contributes, allowing you to pick the most relevant predictors, whereas the random forest algorithm generates reasonable predictions across a wide range of data while requiring little configuration.

Further, we work on the additional features that are effective to classify survival over non-survival rate. To handle the class imbalance problem, the following techniques can be employed. The feature importance of both models in illustrated in Figure 3.

- Feature section technique
- Selecting non-linear model
- An ensemble learning approach can be selected, i.e., Voting or Stacking.
- Three metrics can be measured i.e., accuracy, precision, and recall

IV. Results and Discussions

There are a total of 9111 patients based on the extraction criteria on the above flow chart. Patients without a record for any of the important features excluded. The patients randomly divided into test (n=6834, 75%) and train data (n=2277, 25%) group. There is small multicollinearity observed within the independent variables. This gives a reliable confidence interval that produce more reliable probabilities in terms of effect of the independent variables in a model. Therefore, the inference from the statistical analysis from the model can be more robust and reliable.

In this study, a Random Forest and XGBoost used to build the prediction models. The selection of these models is to handle the class imbalance well understood. As these are the ensemble models, these aggregate the prediction of each base model and results in a better prediction for the unseen data as

TABLE IV: Test set Results

Model	**Accuracy %**	**Sensitivity %**	**Specificity %**
Random Forest	89.77	99.02	8.54
XGBoost	89.72	99.02	8.15

mentioned in the previous section. The issue with the class imbalance dataset has high variance error than the bias. Despite the high accuracy (89.72%), sensitivity (99.02%), and AUC (75.98%) of the XGBoost model, the specificity remained low at 8.15% due to class imbalance. Several imbalance techniques like SMOTE (Majority Under Sampling or Minority Over Sampling) were applied to overcome this problem, the above table shows final results attained with these two classification models (random Forest, XGBoost) illustrated in Table IV.

In the model validation phase measured using area under the curve, the ensemble tree based ML algorithm (RF) and XGBoost is about 0.7592 and 0.7598 respectively. The AUC value has very marginal difference between Random Forest and XGBoost. However, in terms of computational speed, RF is faster than XGBoost. XGBoost needs to select hyperparameters to fit a model but accurate.

This study shows a case of accuracy paradox, meaning high accuracy can sometimes be deceiving and alone is not a good metric for predictive models when classifying in predictive analytics. However, we believe with inclusion of more features from the chart events module from the database, these newly added features should be more significant enough should determine the mortality rate of a patient.

ACKNOWLEDGMENT

The authors would like to thank Dr. Nazmus Sakib[‖], Assistant Professor, University at Buffalo, for his valuable supervision.

REFERENCES

[1] L. A. Celi, A. J. Zimolzak, and D. J. Stone, "Dynamic clinical data mining: search engine-based decision support," *JMIR medical informatics*, vol. 2, no. 1, p. e13, 2014.

[2] H. L. Li-wei, R. P. Adams, L. Mayaud, G. B. Moody, A. Malhotra, R. G. Mark, and S. Nemati, "A physiological time series dynamics-based approach to patient monitoring and outcome prediction," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 1068–1076, 2014.

[3] S. Ghose, J. Mitra, S. Khanna, and J. Dowling, "An improved patient-specific mortality risk prediction in icu in a random forest classification framework," *Stud Health Technol Inform*, vol. 214, pp. 56–61, 2015.

[4] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, and M. J. van der Laan, "Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study," *The Lancet Respiratory Medicine*, vol. 3, no. 1, pp. 42–52, 2015.

[5] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV version:1.0," 2021.

[6] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.