A Machine Learning Approach to Predict Survivability of Patients with Diabetes in Intensive Care Units

Chandreen Liyanage
Department of Computer Science
Lakehead University
Thunder Bay, ON, Canada
cliyanag@lakeheadu.ca

I. Introduction

Predicting the survivability or mortality of a patient in an intensive care unit (ICU) will be important to give proper and timely attention to save people's life. Diabetes is one of the main critical risk factors associated with these critically ill patients who are admitted into ICUs [1]. In addition, compared to other patients admitted into ICUs, patients with diabetes have shown higher mortality rates [2]. Moreover, diabetes increases the risk of organ failure, cardiovascular events, and infections, all of which may lead to ICU admission [2]–[4]. Therefore, immediate identification of the survivability of a diabetes patient will be crucial.

Many studies have predicted ICU mortality using different machine learning techniques, such as random forest, artificial neural network (ANN), decision tree, support vector machine (SVM), and Naïve Bayes [5], [7]–[9]. However, only limited studies have investigated the effect of diabetes on ICU mortality and the majority of these studies were based on statistical sampling techniques. Further, the risk factors associated with the mortality of diabetic patients have not been studied sufficiently.

Based on these considerations, this study conducted a method to predict the ICU survivability of patients with diabetes within the first 24 hours of their admission. A variety of data, such as patient demographic details, data related to ICU stays, laboratory test results taken within 24 hours of admission, and chronic health conditions were used in the analysis. Specifically, this study constructed the work toward achieving three research objectives:

- To predict the survivability/mortality of patients with diabetes in ICUs
- To find the best ML model for predicting ICU survivability under diabetes conditions
- To identify factors that are important in predicting survivability under diabetes conditions at ICUs.

The rest of the paper is structured as follows: in section 2, a brief review of the literature on techniques on ICU mortality prediction and the impact of diabetes on ICU admissions and survivability is presented. Next, the methodology section explains the techniques and approaches used in the study, including data preprocessing and model development. The results of the study and further discussion are presented in section 4. Finally, the conclusion and remark on future research directions will be included.

II. Background Study

Earlier researchers have used different ML techniques to predict deaths in ICUs. A study has used several data mining techniques; ANN, SVM, and decision trees to compare the performance of the statistical logistic regression model in predicting ICU mortality [5]. According to their results, the decision tree model gained the best ROC AUC value of 0.892 over others. Another study has used a test-time augmentation(TTA) technique on tabular patient data to improve the performance of ensemble learning models in ICU survival prediction [6]. This technique helped to improve the accuracy, and generalizability and reduce the bias of their results. They outperformed the commonly used survival prediction system; the APACHE IV scoring and recorded state-of-the-art results. The authors in [7] have used the random forest model to predict the mortality of acute kidney injury patients in ICUs and compared its performance with SVM and ANN. They proved the potential of using RF as an outperformed model in this domain. Other than these, Naïve Bayes [8], [9] and rule-based Projective Adaptive Resonance Theory (PART) [8] have been used as predictive models in this area of research. Additionally, some studies have focused on improving the feature selection in ICU mortality predictions, where [9] and [10] used a feature vector compaction (FVC) technique with ensemble classifiers on laboratory test results of patients at ICUs.

As diabetes increases the risk of patient mortality and is a primary cause of many other diseases, researchers have investigated the impact of diabetes on patients. Some authors have examined the mortality of patients in ICUs with and without prior kidney and heart problems through a population-based cohort study [2]. They used Kaplan-Meier and Cox regression methods during their tests and their findings proved the higher mortality of diabetes patients compared to non-

diabetes, especially who were with previous kidney diseases. While many studies have explored the association between diabetes and other chronic diseases, some studies investigated the impact of infectious diseases on the mortality of diabetic patients in ICUs. One of the previous works aimed to examine the relationship between diabetes and sepsis and predict those patients' 28 days after admission mortality at ICUs [3]. They also used the Cox regression model and propensity score method to statistically predict mortality and revealed that diabetes is not an independent risk factor causing mortality in patients with sepsis. Another study was researched to find the risk factors causing ICU admissions of Covid 19 patients with and without diabetes [4]. They conducted a binary regression analysis on patients' data and discovered that older age, higher respiratory rate, higher blood sugar levels, and higher AST levels are risk factors for such patients with diabetes.

Many researchers who conducted research in predicting ICU motility have commonly used patients' demographical data, physiological data, admission information, chronic health conditions, clinical test results, radiographic data, other impediments, and treatments [4], [5], [10], [11]. Some have used popular datasets, such as the MIT Global Open Source Severity of Illness Score dataset [6], the medical information mart for intensive care (MIMIC) III database [7], [9] and Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database [8], [9]. However, some of the researchers have also used datasets collected from hospitals specifically for their studies [3], [4].

III. Methodology

This work was conducted on google colab using Python language. Common python libraries, such as pandas, numpy, matplotlib and sklearn were used during the development. This implementation is a classification problem where the target is a binary variable stating whether a patient will die or not in ICU. The overall steps in the research methodology is shown in Figure 1.

A. The Dataset

This study used a labeled dataset recently published on Kaggle [12] for patient survival prediction in ICUs. The original dataset contained 85 attributes and 91714 records. However, for this work, the records of patients with diabetes were only extracted. Moreover, many lab test results were ignored to ensure the inclusion of test results only taken within 24 hours of patient admission. Hence, the initial dataset used for this study contained 37 features and 20492 records. A summary of the features is given in Table I.

B. Exploratory Data Analysis and Data preprocessing

The research work of the study was conducted in a few main stages. First, an Exploratory Data Analysis (EDA) was performed to discover the insights of the dataset. Here, the data types, NULL values, statistical distribution of the dataset, correlation between variables, and trends behind selected features were discovered. Next, during the data preprocessing stage,

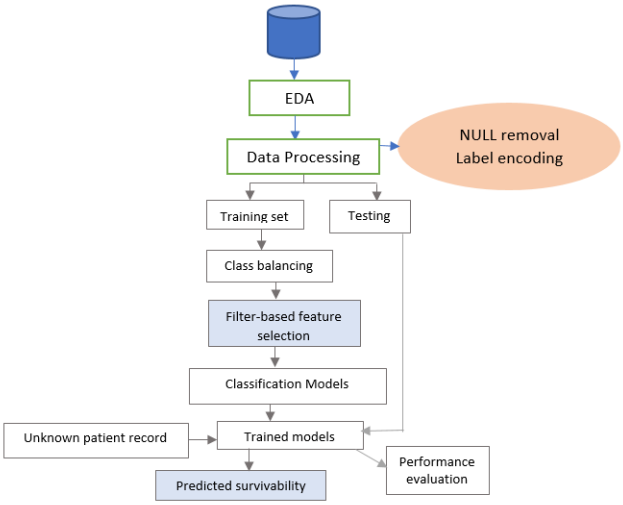

Fig. 1: Steps in model development: EDA, data preprocessing, class-balancing, feature selection and model development

TABLE I: List of features in the dataset

Independent variables	Features
Demographic data	Age, Gender, BMI, Ethnicity, surgery information
ICU stay data	Number of Pre-ICU days, ICU admit source, ICU stay type, ICU type, Pre ICU days
Lab test results (taken within 24hrs)	Heart rates, temperatures, respiratory rates, diagnosis of other illness, intabulation, glucose levels, apache2 diagnosis, ARF apache2, apache2 bodysystem
Chronic health items	Aids, Cirrhosis, Coma, Leukemia, Tumors

data type conversions, such as categorical to numerical encoding, removing features with high NULL value percentages, and removal of unnecessary features, such as patient id, and hospital id were performed. Moreover, the remaining NULL values in the dataset were handled by replacing categorical NULL values by their modes and numerical NULL values by their means.

C. Model development

In the next stage, the model development was conducted by training and testing data on three machine learning models; K nearest neighbor (KNN), decision tree, and random forest.

KNN is a supervised learning algorithm that can be used to solve both classification and regression problems. KNN algorithm assumes that similar things are close to each other. The algorithm stores all available data and classifies a new data point based on the similarity. Decision Trees are a non-parametric supervised learning method. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. Decision Trees require less computational power. Random Forest is an ensemble learning method where it aggregates the results of many decision trees which trained on different subsets of data.

The result of the Random Forest model is the average of prediction results of these individual decision trees. Further, Random Forests have proved to be good at handling larger datasets efficiently.

The dataset was split into 70:30 training and testing ratios respectively. As the majority of the dataset was from the hospital death "0" class (1: 18897 and 0:1595), a variant of the Synthetic Minority Over-sampling Technique (SMOTE) technique called "SMOTENC" [13] was used to perform the class balance in the dataset with both nominal and continuous data. The data balancing was performed only on the training set after separating the dataset to avoid possible information leakages.

D. Feature selection

As a requirement of one of the objectives and to avoid overfitting, a feature selection technique was implemented to filter the most important factors in predicting ICU mortality of diabetic patients. A filter-based feature selection technique was implemented using the inbuilt SelectKBest function in Python and ANOVA F-value was used to generate the scores. The higher scores represent the higher importance of features. scores for each feature returned from this method were used to rank the features in the dataset and these features. Next, these features were used to train the ML models by experimenting with different subsets of them. However, instead of using a random number of features, starting from the top-ranked attribute, each feature was added one by one and analyzed the performance. Finally, the most important feature subset was taken from the instance where the predictive model showed the highest performance.

E. Performance evaluation

The model performances were evaluated based on the area under the Receiver Operating Characteristic curve (ROC AUC) and 5-fold cross-validation (CV) accuracy. The CV was performed only on the training dataset. Apart from them, the classification reports and confusion matrices were generated to evaluate the results.

F. Hyper-parameter tuning

In the decision tree model, hyper-parameter tuning was performed to find the optimal hyper-parameter settings of the maximum depth of the tree and the function used to measure the quality of a split. For this GridSearchCV function of sklearn library was used. The model was retrained with these optimal settings; criterion as entropy and max depth as 92 for better performance.

IV. Results and Discussion

A. EDA

This dataset consisted of 20492 patients from 16 years to 89 years and the average age of patients is 64. Some of the features, such as hospital death probability, ICU death probability, and potassium levels in blood were contained with more than 1500 NULL values. However, 18 features in the initial dataset did not have any NULL values. The correlation between attributes in the dataset was discovered using the Pearson coefficient measurement and figure 2 shows the resulting heatmap. According to the results, the minimum and maximum heart rates, apache2 diagnosis and surgery status, maximum temperature, and maximum heart rate are highly correlated pairs of features.

Next, the survival rates of diabetic patients based on the places they were admitted into ICUs were analyzed. As Figure 3 explains, the patients who came from the floor and other hospitals are having higher average death rates. Comparatively, the patients admitted from operating rooms/ recovery are having fewer death rates. The accident and emergency wards are also showing higher to lower deaths.

B. Feature Selection and Model Development

The best performances of three ML models are recorded in Table II. According to the results, the random forest method has outperformed the other two models with 0.81 ROC AUC and 0.90 accuracies on the testing dataset. Also, the CV accuracy on the training dataset is 0.94. Among the other two models, the decision tree model showed better results than KNN, except for the ROC AUC value. Moreover, fine-tuning the decision tree model resulted in better performance in both CV and testing accuracy.

Figure 4 demonstrates how the ROC AUC values and cross-validation accuracy varies with the different number of features in the best model, random forest. As demonstrated, the performance of the model improves when the number of features are increasing.

As the answer for the third objective, the best number of features were selected from the best model. However, as the random forest gave the best ROC AUC at feature 29 and which is a considerably larger feature space, this study decide to select the best number of features as 10 where at random forest started to obtain the best results. This feature count was further justified by the CV accuracy at a similar feature count and the average number of features obtained by the decision tree model. These best 10 features are listed below. Many of the important factors are laboratory test results and no chronical diseases influencing the mortality of diabetic patients.

- maximum heart rate
- age
- minimum heart rate
- maximum temperature
- maximum respiration rate
- minimum temperature
- minimum respiration rate
- minimum glucose level in blood
- maximum glucose level in blood
- apache_2_diagnosis

V. Conclusion

This study predicts the mortality of patients with diabetes at ICUs using three machine learning models. The Random Forest model has performed well over the other two models,

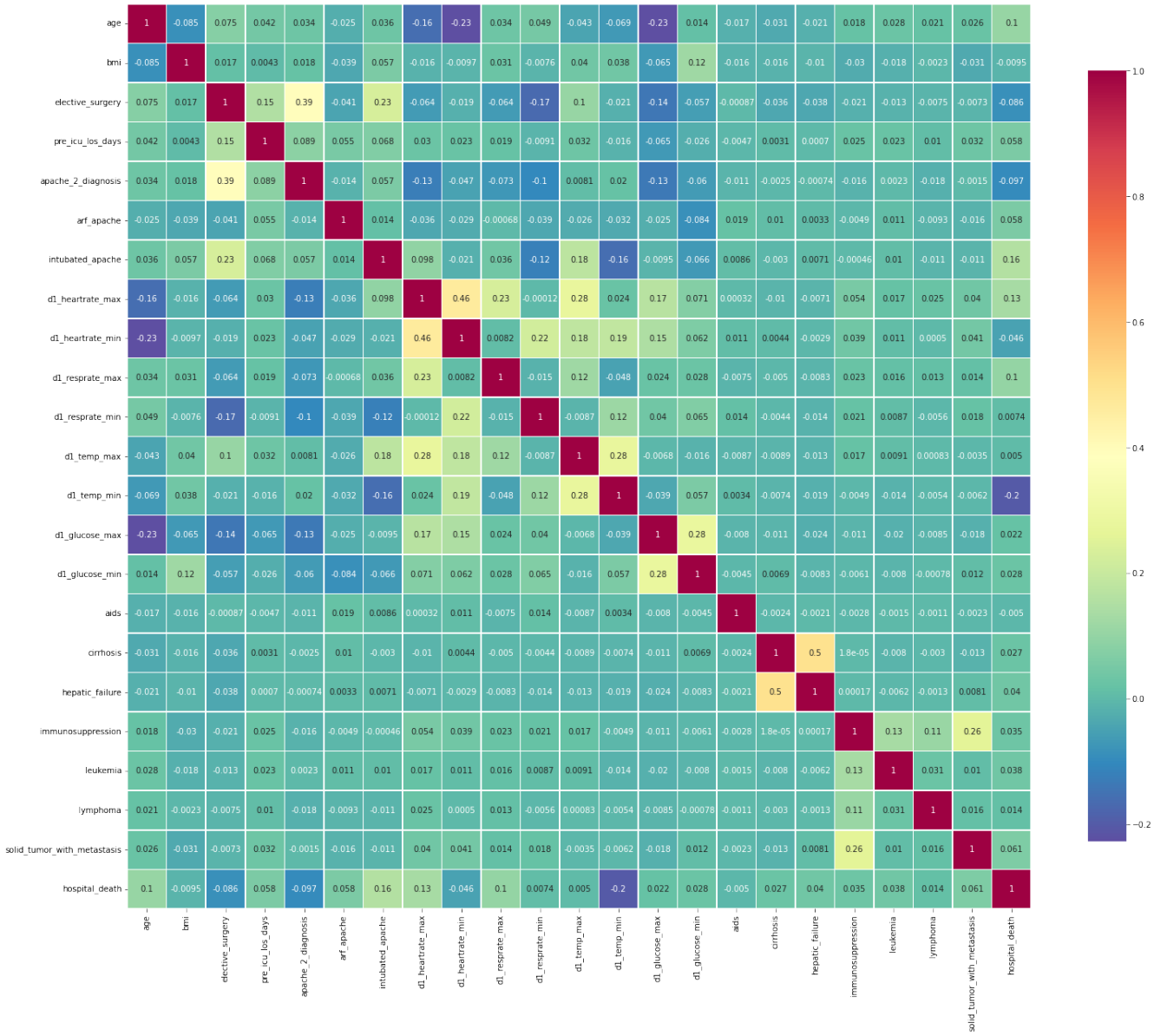

Fig. 2: The Person Correlation between features in the dataset

TABLE II: Results of classification models

Model	Best feature count from the filter-based method	Training Average CV accuracy	Testing ROC AUC	Testing accuracy
KNN	6	0.75	0.65	0.75
Decision Tree before fine-tuning	13	0.78	0.73	0.73
Decision Tree after fine-tuning	14	0.89	0.61	0.84
Random Forest	26	0.94	0.81	0.90

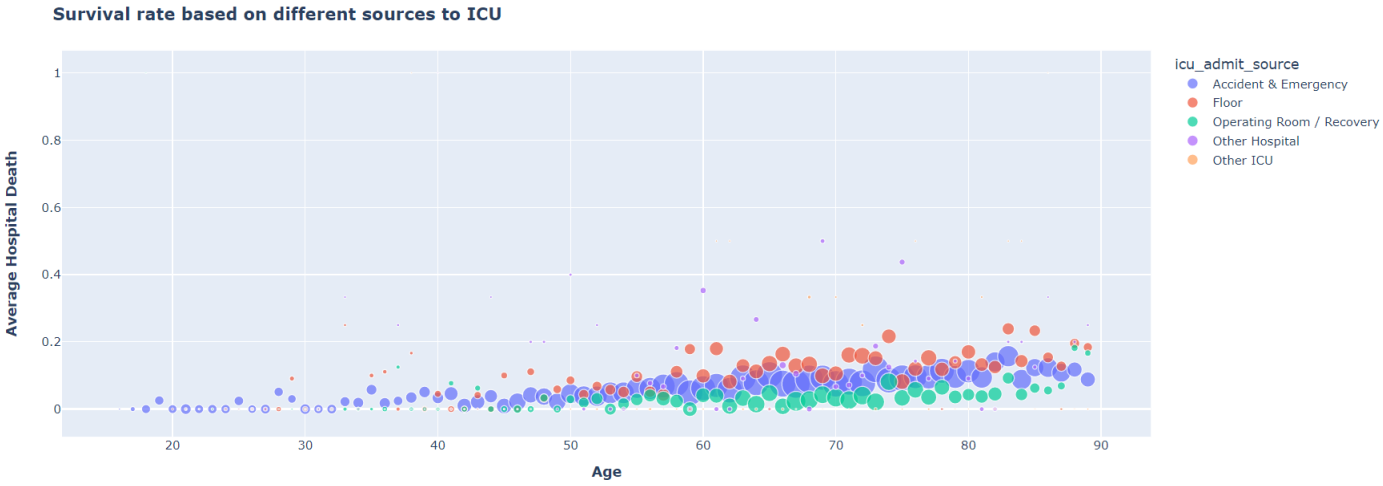

Fig. 3: patient survival rate based on different sources to ICU

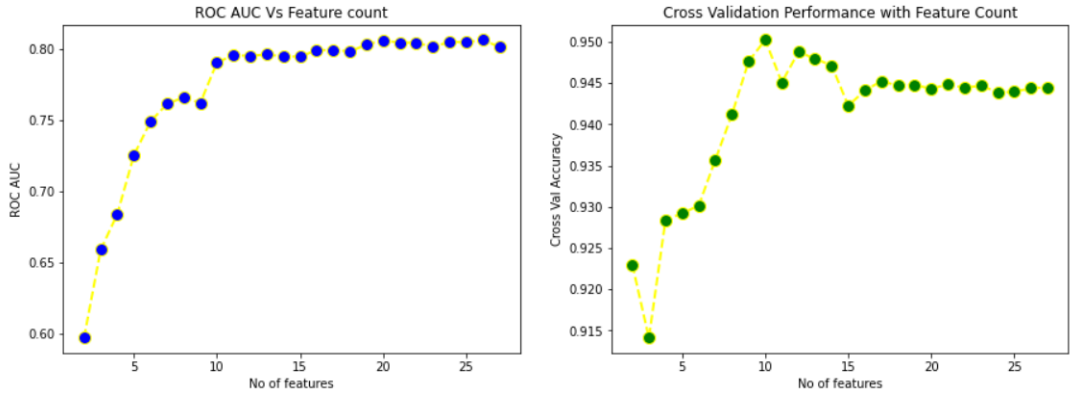

Fig. 4: Performance of random forest model left: ROU AUC right: CV accuracy against the number of features.

KNN and decision tree with better training cross-validation accuracies and testing ROC AUC and testing accuracy. Moreover, this study was able to find the best ten factors predicting diabetic patients' deaths in ICUs. Patients' heart rates, age, body temperatures, respiration rate, and glucose levels are some of them. This study will be significant due to the implementation of ML techniques to predict the mortality of patients with diabetes in ICUs, the ability to predict within 24hrs of a patient's admission, and using a fresh dataset to gain a variety of lab test measurements which has not been analyzed previously. The current study can be further improved by implementing many other ML and deep learning models. Moreover, advanced feature selection techniques can be used to enhance the results.

REFERENCES

[1] Siegelaar, Sarah E., et al. "The effect of diabetes on mortality in critically ill patients: a systematic review and meta-analysis." Critical Care 15.5 (2011): 1-12.

[2] Christiansen, Christian F., et al. "Type 2 diabetes and 1-year mortality in intensive care unit patients." European Journal of Clinical Investigation 43.3 (2013): 238-247.

[3] Lin, Shan, et al. "Association between comorbid diabetes mellitus and prognosis of patients with sepsis in the intensive care unit: a retrospective cohort study." Annals of Translational Medicine 9.1 (2021).

[4] Lei, Ming, et al. "Clinical features and risk factors of ICU admission for COVID-19 patients with diabetes." Journal of diabetes research 2020 (2020).

[5] Kim, Sujin, Woojae Kim, and Rae Woong Park. "A comparison of intensive care unit mortality prediction models through the use of data mining techniques." Healthcare informatics research 17.4 (2011): 232-243.

[6] Cohen, Seffi, et al. "ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models." IEEE Access 9 (2021): 91584-91592.

[7] Lin, Ke, Yonghua Hu, and Guilan Kong. "Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model." International journal of medical informatics 125 (2019): 55-61.

[8] Awad, Aya, et al. "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach." International journal of medical informatics 108 (2017): 185-195.

[9] Alam, Md Zahangir, et al. "Feature-ranking-based ensemble classifiers for survivability prediction of intensive care unit patients using lab test data." Informatics in Medicine Unlocked 22 (2021): 100495.

[10] Masud, Mohammad M., and Abdel Rahman Al Harahsheh. "Mortality prediction of ICU patients using lab test data by feature vector compaction classification." 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016.

[11] Smith, ME Beth, et al. "Early warning system scores for clinical deterioration in hospitalized patients: a systematic review." Annals of the American Thoracic Society 11.9 (2014): 1454-1465.

[12] Agarwal, M. "Patient Survival Prediction: Deep Learning." https://www.kaggle.com/code/mitishaagarwal/patient-survival-prediction-deep-learning/data (2022).

[13] SMOTENC. ”imbalanced-learn developers.” https://imbalanced-learn.org/stable/references/generated/imblearn.oversampling.SMOTENC.html (2022).