

Severity and Outcome Detection for the Cancer Patients with COVID-19 Using Machine Learning Models

Akbar Davoodi

Isfahan University of Medical Sciences

Shaghayegh Haghjooy Javanmard

Isfahan University of Medical Sciences

Golnaz Vaseghi

Isfahan University of Medical Sciences

Amirreza Manteghinejad (✉ manteghinejad@edc.mui.ac.ir)

Isfahan University of Medical Sciences

Research Article

Keywords: Machine learning, COVID-19, Cancer, Artificial Intelligence

Posted Date: November 5th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-989924/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background:

The COVID-19 pandemic challenges the healthcare system to provide enough resources to battle the pandemic without jeopardizing routine treatments. As a result, this is important that we can predict the outcomes of patients at the time of admission. This study aims to apply different machine learning (ML) models for predicting Intensive Care Unit (ICU) admission and mortality of Cancer Patients infected with COVID-19.

Methods:

This study's data were collected from a referral cancer center in Iran. The study included all patients with cancer and a confirmed diagnosis of COVID-19.

Different ML prediction algorithms like Logistic Regression (LR), Naïve Bayes (NB), k-Nearest Neighbours (kNN), Random Forest (RF), and Support Vector Machine (SVM) were used. Also, we applied the SelectKBest method to find the most important features for predicting ICU admission and mortality.

Results:

Three hundred thirty-nine patients enrolled in the study. One hundred fifteen were admitted to the Intensive Care Unit (ICU), and 118 patients died during the hospital admission.

The Area Under Curve (AUC) for predicting mortality is 0.61 for LR, 0.74 for NB, 0.61 for kNN, 0.6 for SVM, and 0.79 for RF. The AUC for predicting ICU admission is 0.61 for LR, 0.74 for NB, 0.56 for kNN, 0.55 for SVM, and 0.7 for RF.

C-reactive protein (CRP), Aspartate transaminase (AST), and Neutrophil-Lymphocyte Ratio (NLR) also are the most common features in predicting ICU admission and mortality.

Conclusion:

Our findings show the promise of different AI methods for predicting the risk of death or ICU in cancer patients infected with COVID-19, highlighting the importance of first laboratory results and patients' symptoms.

Background:

Since two years ago, the coronavirus disease 2019 (COVID-19) pandemic has remained a global public health emergency(1). The COVID-19 pandemic challenges the healthcare system to provide enough resources to battle the pandemic without jeopardizing routine treatments for other patients(2). Previous studies have shown a higher risk of COVID-19 infection (0.79–8.3%) as well as an increased risk of COVID-19-associated mortality in cancer patients (mortality ranging from 11.4–35.5%)(3).

Several predictive models have revealed COVID-19 severity and mortality determinants in the general population, but the clinical outcomes and interactions of COVID-19 with cancer and systemic anticancer drugs are poorly understood. Besides diagnosis and treatment, predicting prognosis is needed to reduce the pressure on healthcare systems and deliver the best possible care for patients. Several machine learning (ML) models have been created to help clinicians and improve the diagnosis of COVID-19(4).

ML models can also predict the COVID-19 outcomes such as Intensive Care Unit (ICU) admission and mortality(5). To date, various ML algorithms have been developed for the diagnosis of COVID-19 and prediction of severity and mortality risk, using clinical and laboratory data(6–10). In comparison, few studies surveyed ML algorithms in predicting COVID-19 outcomes in cancer patients(11, 12).

This study aims to apply different ML models for predicting Intensive Care Unit (ICU) admission and mortality in hospitalized cancer patients infected with COVID-19.

Materials And Methods:

Data Collection:

This study's data were collected from Omid hospital, a referral cancer center affiliated with Isfahan University of Medical Sciences and Isfahan COVID-19 Registry (I-CORE)(13). The study included all patients with active or previous cancer with confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection by RT-PCR from February 2020 to February 2021. Patients with a radiological or clinical diagnosis of COVID-19 without a positive RT-PCR test were not included in this analysis.

28 features from each patient were collected, including demographics data like age and sex, cancer characteristics including type of cancer, history of recent chemotherapy, COVID-19 symptoms including fever, cough, dyspnea, weakness, diarrhea, nausea, and vomiting. Comorbidities such as hypertension and diabetes were also recorded. All the laboratory data at the time of admission, including White Blood Cell (WBC), Absolute Neutrophil Count(ANC), Absolute Lymphocyte Count(ALC), Neutrophil to Lymphocyte Ratio (NLR), Hemoglobin (HB), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin Concentration (MCHC), Red cell Distribution Width (RDW), Platelets (PLT), Blood Urea Nitrogen (BUN), Creatinine (CR), Aspartate Transaminase (AST), Alanine transaminase (ALT), C-Reactive Protein (CRP) also were collected.

Defining patient outcomes:

ICU admission during hospitalization and patient status at the time of discharge (alive/dead) were defined as patients' outcomes. Machine learning models aim to predict these outcomes based on collected features and select the most important features that affect the outcomes.

Data Pre-processing:

The dataset is recorded very carefully; consequently, only less than 4 percent of all features in the whole dataset are detected as missing values. The missing values are all from the lab tests. For filling the missing values, we replace every missing value with the mean of the corresponding feature. After that, we randomly partition the dataset into two sets: the training set containing 305 inputs and the test set containing 34 patients' information. Then, as the third step, we apply a feature scaling algorithm to the training dataset, called standardization, to ensure that none of the features are dominated by other features. To do this, we used the formula $X_{Stand}(X) = \frac{X - Mean(X)}{\sqrt{Var(X)}}$ in which $Var(X)$ stands for the variance of X .

Prediction Algorithms:

We apply several machine learning prediction algorithms, namely Logistic Regression (LR), Naïve Bayes (NB), k-Nearest Neighbours (kNN), Random Forest (RF), and Support Vector Machine (SVM). In the following, we bring experimental settings of the reported prognostic models.

The first method for providing the results is binary logistic regression which is applied for data sets with "0" and "1" class labels.

The second technique is naïve Bayes which is a conditional probability model. Assume that x is the feature vector corresponding to input, and C_1 and C_2 are the two possible label classes. Using Bayes theorem in probability theory, the model computes conditional probabilities $Pr(C_1 | x)$ and $Pr(C_2 | x)$ and then, based on the comparison of these two values, decides for the label of x . Since there are binary-valued variables in our dataset, we use the Bernoulli version of naïve Bayes.

k-Nearest Neighbors (kNN) classifier is another algorithm that is applied to this dataset. In this algorithm, the new input label is determined based on the labels of k nearest samples in the dataset. The user specifies the input parameter k , which is typically small. We set this variable to be seven by applying some different values.

Random Forest (RF) method is based on the bootstrap aggregating technique because the prediction of a single tree is very sensitive to noise in its training set, while the average of many trees is not if the trees are not correlated. Consequently, this algorithm reduces the variance of the model in some sense. To run the algorithm, you can determine the maximum number of steps to go to the depth of the tree for each tree. We found the best value for this parameter based on our training dataset. Therefore, we set this tuning parameter to be 13; however, if it is not determined, nodes are expanded until all leaves are pure or until all leaves contain less than the minimum number of samples required to split an internal node.

As another method, we use SVM. Let us have an N -dimensional features vector. The Support Vector Machine (SVM) model approach finds a hyperplane in an N -dimensional space with the maximum margin that distinctly classifies the data points.

Prediction Performance Evaluation Metrics:

To explain these measures, first, we need some preliminary definitions. As we have a binary classification problem, each label can be seen as positive or negative. By a true positive (TP) value, we mean a correctly predicted positive label. Similarly, a true negative (TN) stands for an indeed detected negative label. A false positive (FP) value shows an input data point that is predicted by positive while its actual label is negative. Finally, a false negative (FN) value is used for a data point where its real class is positive, but the algorithm predicts its label as negative.

One of the most common evaluation measures is accuracy, which is defined as the ratio of correctly predicted items to the total number of items. If we have imbalanced label classes, then accuracy cannot be applied as a good evaluation criterion. For example, if 98 percent of the labels are positive in a dataset, then the trivial algorithm which assigns positive to all the inputs reaches an accuracy of 98 percent. In this case, to better understand the algorithm's performance, we should realize how much of each label class is predicted correctly.

The second parameter, precision, is the ratio of truly detected positive labels to the total number of predicted positive labels in the test set. In other words, $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$.

Also, recall is defined as the number of correctly predicted positive labels divided by the total number of positive labels. Therefore, $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$. In binary classification, it is worth noting that recall of the positive and negative classes are called sensitivity and specificity, respectively. Finally, F1-score, a mixture parameter, is the weighted average of precision and recall, i.e., $\text{F1-score} = (2 (\text{Precision} * \text{Recall})) / (\text{Precision} + \text{Recall}) = \text{TP} / (\text{TP} + 1/2(\text{FP} + \text{FN}))$. It is worth noting that this score is much better than accuracy if we have imbalanced classes. On the other hand, it is better to look at accuracy whenever FPs and FNs have the same weight in the framework. Also, if the cost of FPs and FNs are very different, then both precision and recall should be taken into account.

Feature Selection:

To find out the most important features for predicting outcomes, a list of the five most important features is obtained for each prediction outcome. We applied SelectKBest, a univariate feature selection method, which identifies the k highest scoring features.

Ethics Statement:

This study was approved by the Research Ethics Committees of Vice-Chancellor in Research, Medical University of Isfahan (Approval ID: IR.MUI.RESEARCH.REC.1399.004).

Results:

Three hundred thirty-nine patients enrolled in the study with a positive PCR from 26 February 2020 to 15 February 2021. One hundred fifteen patients from this cohort were admitted to the Intensive Care Unit (ICU), and 118 patients died during the hospital admission. Table 1 shows the demographic and characteristics of patients.

Table 1
Demographic and characteristics of patients

Variable	Values	Number	Percent
Sex			
	Male	189	55.8%
	Female	150	44.2%
Age			
Mean \pm SD 53.37 \pm 18.69	<20	27	7.96%
	20-30	7	2.06%
	30-40	28	8.26%
	40-50	56	16.52%
	50-60	77	22.71%
	60-70	80	23.60%
	>70	64	18.88%
Tumor Type			
	solid	161	47.5%
	Hematologic	178	52.5%
ICU Admission			
	No	227	66.4%
	Yes	115	33.6%
Outcome			
	Alive	221	65.2%
	Dead	118	34.8%

Tables 2 and 3 show the results for different algorithms, in which a test set is used to evaluate the performance of algorithms according to different evaluation metrics, namely accuracy, precision, sensitivity, specificity, and F1-score. In each experiment, we split each dataset into training and test sets, 90 percent for training and 10 percent for the test sets. We repeat each experiment 30 times by changing the training and test sets. The average results are tabulated in Tables 2 and 3.

Table 2
Comparing different ML Models when Mortality is the outcome

ML Models	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Reg	0.678	0.708	0.605	0.714	0.685
NB	0.71	0.712	0.502	0.814	0.705
kNN	0.665	0.653	0.256	0.869	0.628
RF	0.685	0.666	0.304	0.874	0.653
SVM	0.7	0.69	0.226	0.938	0.645

Table 3
Comparing different ML Models when ICU Admission is the outcome

ML Models	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Reg	0.655	0.731	0.59	0.678	0.676
NB	0.725	0.725	0.367	0.847	0.717
kNN	0.732	0.659	0.093	0.948	0.665
RF	0.726	0.654	0.09	0.939	0.663
SVM	0.747	0.578	0.008	0.996	0.644

Moreover, we analyze the results of Receiver Operating Characteristic (ROC) curve for all the methods. This curve is commonly used to explain the diagnostic ability of each while its discrimination threshold is varied. It is well-known that the Area Under Curve (AUC) is considered as an evaluation metric to comprise different algorithms. AUC curves of different algorithms for mortality and ICU are shown in Figures 1 and 2. It is inferred from the figures that; NB and RF have better performance than the others. In fact, the curves of these two methods dominate the other curves in most of the cases. Similarly, their AUC values are higher than others.

Moreover, a list of the five most important features is obtained for each outcome prediction. Table 4 shows the results of feature selection.

Table 4
The five most important features that affected outcomes

Predicting item	Highest scoring features
ICU Admission	CRP-AST-NLR-Cough-Dyspnea
Mortality	BUN-CRP-AST-NLR-Vomit

Our feature detection approach shows that in different outcomes, we have 3 common features, which are: C-Reactive Protein (CRP), Neutrophil to Lymphocyte Ratio (NLR), and Aspartate Aminotransferase (AST).

Discussion:

We used machine learning algorithms to identify clinical variables predictive of severe COVID-19 illness in cancer patients at time zero. We compared different AI models, and NB and Rf have high performance with AUC of 0.74 and 0.79 in predicting ICU and death, respectively.

Cancer patients are more susceptible to COVID-19 infection and COVID-19-related complications regardless of active cancer(14). Several studies have shown that cancer patients infected with COVID-19 have higher mortality risks. For instance, a study in the United Kingdom (UK) showed that 28 percent of cancer patients admitted to hospital died because of COVID-19, a higher rate than the normal population(15). The results of another study by The US COVID-19 and Cancer Consortium (CCC19) showed that 13% of cancer patients died, and 26% became severely ill (14, 16).

Prediction of which patients are at higher risk of progression and poor outcomes can help clinicians better decisions during the critical time of disease course. Up to now, many ML prediction models have been developed which are useful in the clinical setting. Among all of them, we focused on five different methods LR, NB, RF, kNN, and SVM. Since we have imbalanced data and a small sample size for data of positive label, it is clear that the methods on this class label data are not as good as those negative ones. If it is important to have high performance on positive class labels, i.e., high sensitivity value, it is recommended to use probabilistic algorithms like NB and LR. Otherwise, according to ROC curve results, we can apply the most dominating methods, i.e., RF and NB.

Around 30 percent of data have a positive class label, and the remaining have a negative one. Since we have the problem of a small sample size for data points having the positive class label, we cannot capture their structure properly; thus, it is expected that most of the methods perform weakly in terms of sensitivity. By analyzing the results of Tables 2 and 3, we can also observe these meaningful differences between sensitivity and specificity measures. However, the logistic regression and naïve Bayes are two probabilistic methods where they could consider the uncertainty which exists in the embedding of data points. For this reason, the results of these two methods are better than others in terms of sensitivity metrics.

The most frequently reported predictors for the prognosis of COVID-19 cases were age, CRP, lymphocyte, and LDH in the general population(17).

Our results showed that between the routine laboratory tests, increased CRP, NLR, and AST are the best predictors of the severity and mortality in cancer patients infected by COVID-19.

Several studies have shown that the increased level of CRP can predict disease severity and outcome in patients with COVID-19(10, 18, 19). It has been suggested that an excessive immune response or

'cytokine storm' plays a critical role in COVID-19 severity and outcome.

High AST value may suggest that SARS CoV2 may cause damage to multiple organ systems, including the liver, when emerging as a severe inflammatory disease. NLR plays a key role in maintaining immune homeostasis and the inflammatory response in the body. Several previous reports have shown the prognostic value of NLR in COVID-19 patients(20, 21).

Our study has some limitations:

1. We analyzed a medium-sized dataset of 339 cancer patients; larger, more comprehensive datasets of cancer patients are needed to test the power of generalization of our approach.
2. Other algorithms and features are needed to predict how clinical variables change over time may affect future outcomes.
3. We did not report how many days in advance our model can produce predictions.

Conclusion:

Our findings show the promise of different AI methods for predicting the risk of death or ICU in cancer patients infected with COVID-19, highlighting the importance of first laboratory results and patient symptoms.

Abbreviations

Absolute Lymphocyte Count: ALC

Absolute Neutrophil Count: ANC

Alanine transaminase: ALT

Area Under Curve: AUC

Aspartate transaminase: AST

Blood Urea Nitrogen: BUN

coronavirus disease 2019: COVID-19

C-reactive protein: CRP

Creatinine: CR

False Negative: FN

False Positive: FP

Hematocrit: HCT

Hemoglobin: HB

Intensive Care Unit: ICU

Isfahan COVID-19 Registry: I-CORE

k-Nearest Neighbours: kNN

Logistic Regression: LR

Mean Corpuscular Hemoglobin Concentration: MCHC

Mean Corpuscular Volume: MCV

ML: Machine Learning

Naïve Bayes: NB

Neutrophil-Lymphocyte Ratio: NLR

Platelets: PLT

Random Forest: RF

Receiver Operating Characteristic: ROC

Red cell Distribution Width: RDW

severe acute respiratory syndrome coronavirus 2: SARS-CoV-2

Support Vector Machine: SVM

True Negative: TN

True Positive: TP

White Blood Cell: WBC

Declarations

Ethics approval and consent to participate:

The study was carried out in accordance with the Helsinki Declaration (IV Adaptation). The study protocol was approved by the Research Ethics Committees of Vice-Chancellor in Research, Medical University of Isfahan (Approval ID: IR.MUI.RESEARCH.REC.1399.004). In view of the retrospective nature, the need for

individual patient consent was waived by the research ethics committee as data protection safeguard was in place.

Consent for publication:

Not applicable

Availability of data and materials:

The datasets of the study are not publicly available due to patient confidentiality, but a de-identified version will be made available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions:

AM, SH, AD, and GV performed the study concept and design. AM was involved in data acquisition. AD and AM are involved in data pre-processing. AD built the models and trained them. AM, AD wrote the first draft of the manuscript. All authors contributed to the critical revision and final approval of this manuscript.

Acknowledgments:

The authors wish to thank Dr. Mehran Sharifi, Saeedeh Arabzadeh, and Zahra Rezaeian for their assistance during data gathering and other associated healthcare providers and hospital staff of Omid Hospital for their pure services and cooperation during COVID-19.

References

1. Cruz MP, Santos E, Cervantes MV, Juárez ML. COVID-19, a worldwide public health emergency. *Revista Clínica Española (English Edition)*. 2021;221(1):55–61.

2. Pati S, Mahapatra P, Kanungo S, Uddin A, Sahoo KC. Managing multimorbidity (multiple chronic diseases) amid COVID-19 pandemic: a community based study from Odisha, India. *Frontiers in public health*. 2021;8:1026.
3. de Las Heras B, Saini KS, Boyle F, Ades F, de Azambuja E, Bozovic-Spasojevic I, et al. Cancer treatment and research during the COVID-19 pandemic: experience of the first 6 months. *Oncology and therapy*. 2020;8(2):171–82.
4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*. 2020;369.
5. Subudhi S, Verma A, Patel AB. Prognostic machine learning models for COVID-19 to facilitate decision making. *International Journal of clinical practice*. 2020;74(12):e13685.
6. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2020;1(ahead-of-print).
7. Yao H, Zhang N, Zhang R, Duan M, Xie T, Pan J, et al. Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in cell and developmental biology*. 2020;8:683.
8. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence*. 2020;2(5):283–8.
9. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, et al. COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in public health*. 2020;8:357.
10. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *European Respiratory Journal*. 2020;56(2).
11. Navlakha S, Morjaria S, Perez-Johnston R, Zhang A, Taur Y. Projecting COVID-19 disease severity in cancer patients using purposefully-designed machine learning. *BMC infectious diseases*. 2021;21(1):1–8.
12. Dienstmann R, Menezes M, Silva MCE, Cruz H, Paes R, Silva JAd, et al. Machine learning prediction of COVID-19 mortality in cancer patients. *Journal of Clinical Oncology*. 2021;39(15_suppl):1558-.
13. Javanmard SH, Nasirian M, Ataei B, Vaseghi G, Vaezi A, Changiz T. Isfahan COvid-19 REgistry (I-CORE): Design and methodology. *J Res Med Sci*. 2020;25:32.
14. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR, et al. Clinical impact of COVID-19 on patients with cancer (CCC19): a cohort study. *The Lancet*. 2020;395(10241):1907–18.
15. Lee LY, Cazier JB, Starkey T, Turnbull C, Team UCCMP, Kerr R, et al. COVID-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a prospective cohort study. *The Lancet*. 2020;395(10241):1919–26.
16. Desai A, Gupta R, Advani S, Ouellette L, Kuderer NM, Lyman GH, et al. Mortality in hospitalized patients with cancer and coronavirus disease 2019: A systematic review and meta-analysis of cohort studies. *Cancer*. 2021;127(9):1459–68.

17. Adamidi ES, Mitsis K, Nikita KS. Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review. *Computational and Structural Biotechnology Journal*. 2021.
18. Guan X, Zhang B, Fu M, Li M, Yuan X, Zhu Y, et al. Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. *Annals of Medicine*. 2021;53(1):257–66.
19. Qin S, Li W, Shi X, Wu Y, Wang C, Shen J, et al. 3044 Cases reveal important prognosis signatures of COVID-19 patients. *Computational and Structural Biotechnology Journal*. 2021;19:1163–75.
20. Javanmard SH, Vaseghi G, Manteghinejad A, Nasirian M. Neutrophil-to-Lymphocyte ratio as a potential biomarker for disease severity in COVID-19 patients. *Journal of global antimicrobial resistance*. 2020;22:862–3.
21. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA internal medicine*. 2020;180(8):1081–9.

Figures

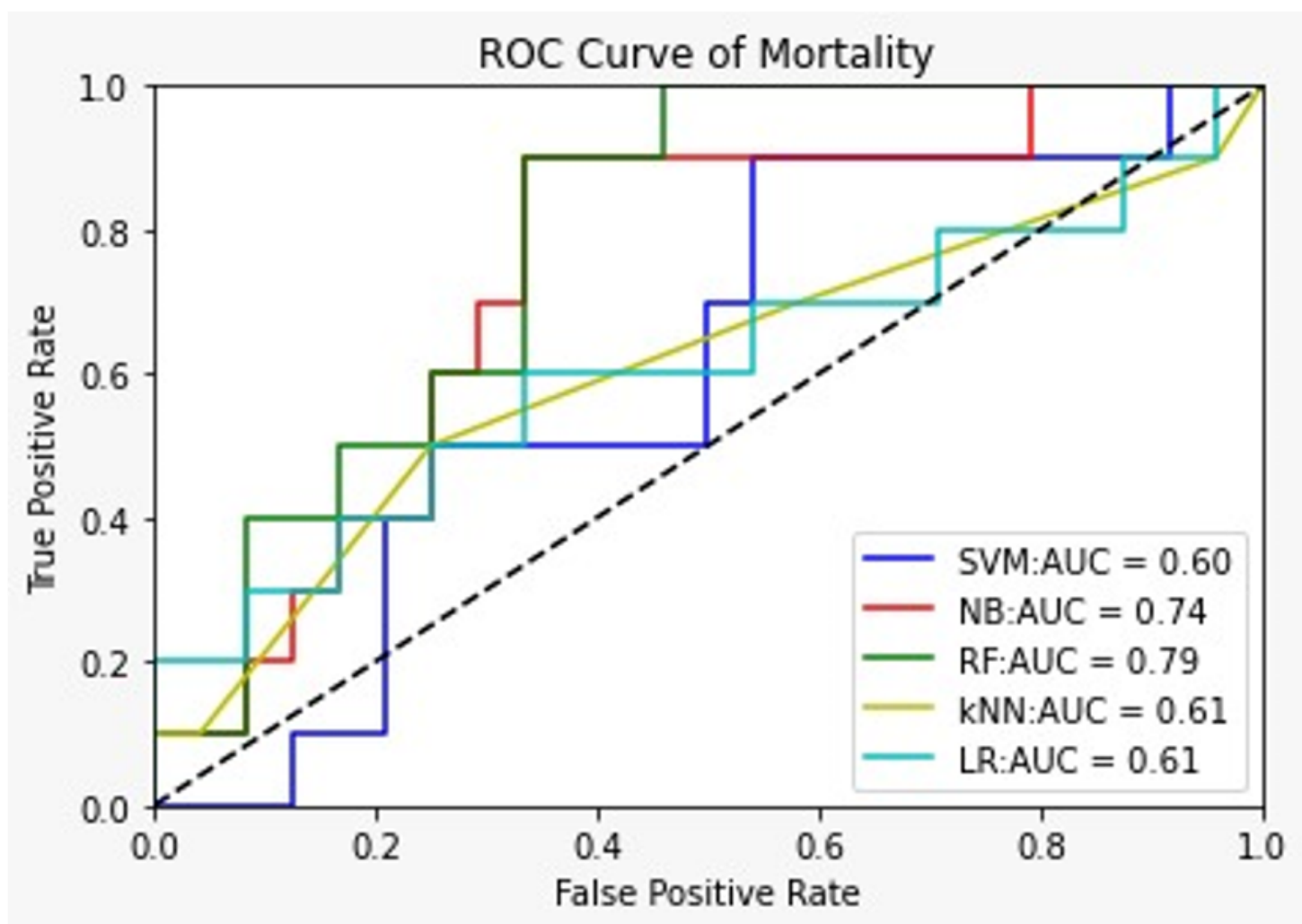


Figure 1

Mortality ROC Curves of different ML Models

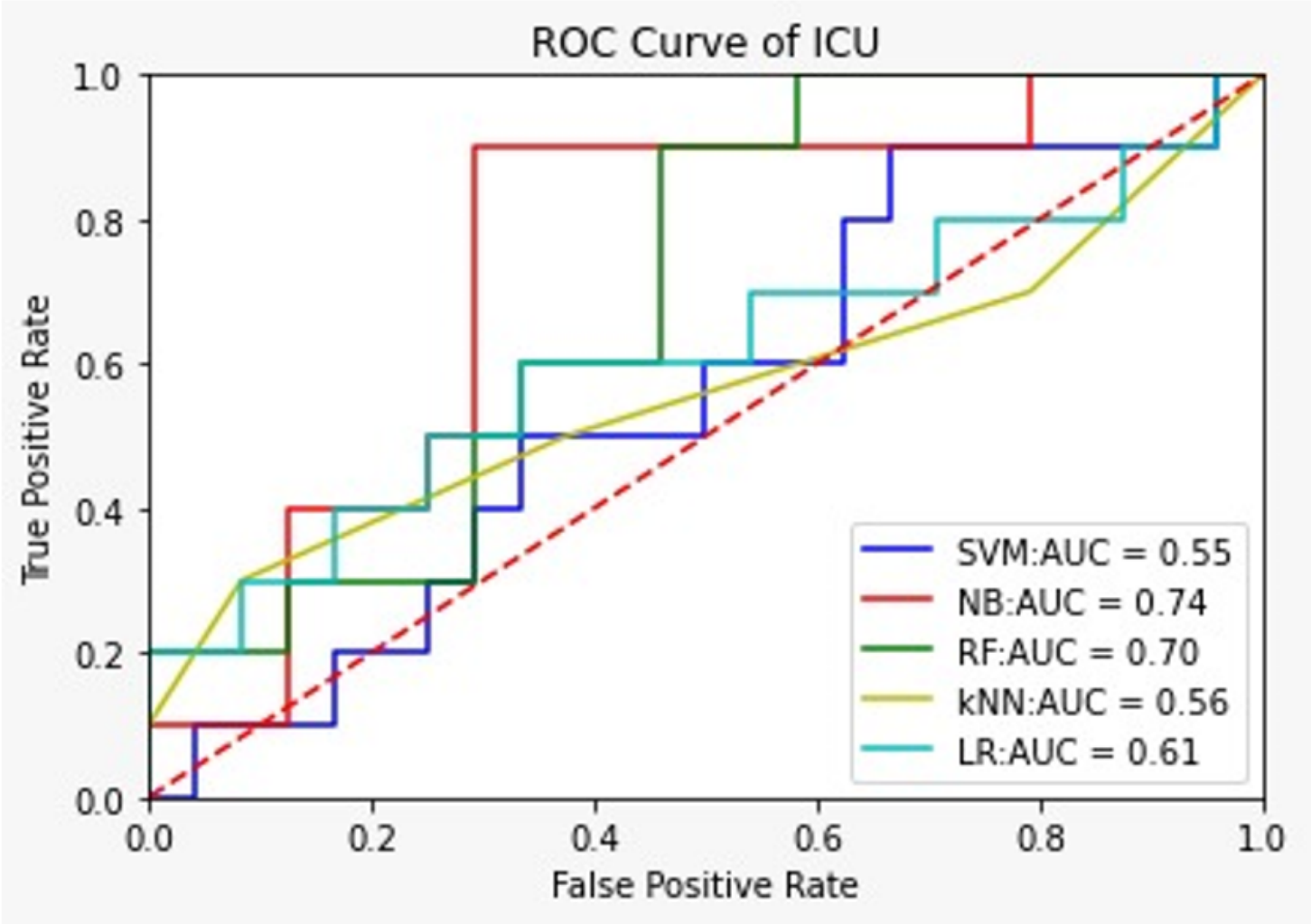


Figure 2

ICU Admission ROC Curves of different ML Models