Classification of Covid-19 Patients Requiring Intensive Care Unit

Fransiska A. Christiana Holy
Informatics Department
Sanata Dharma University
Yogyakarta, Indonesia
fransiska.annalisa.fa@gmail.com

Paulina H. Prima Rosa
Informatics Department
Sanata Dharma University
Yogyakarta, Indonesia
rosa@usd.ac.id

2021 25th International Computer Science and Engineering Conference (ICSEC) | 978-1-6654-1197-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSEC53205.2021.9684628

***Abstract*—As the world faced the covid-19 pandemic, there was a surge in the number of patients that overwhelmed many hospitals. Due to the limited number of Intensive Care Units (ICUs), some hospitals also find it difficult to meet ICU needs for covid-19 patients. So there is a need to set priorities for patients who really need to get treatment in ICU. In this paper, a classification modelling of Covid-19 patients requiring ICU was carried out using Support Vector Machine (SVM) algorithm. The data used to build the model was data from Mexican government obtained from the Kaggle website. Tests were carried out on 3 types of SVM kernels, namely Linear Kernel, Polynomial Kernel, and Gaussian RBF Kernel toward dataset before and after balancing process. From the results of validation testing using 3-fold and 5-fold cross validation, the best accuracy of 87.1055% was obtained using the three kernels toward dataset without balancing.**

Keywords—covid-19, intensive care unit, data mining, support vector machine

I. INTRODUCTION

To date, the world is still struggling to overcome the SARS covid-19 virus pandemic which has caused millions of people died. The virus emerged for the first time at the end of 2019 in Wuhan, China. It was identified as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS – Cov2) or better known as Covid - 19. This virus can attack children, adults and the elderly. Coronavirus is a collection of viruses that can infect the respiratory system. This virus causes the respiratory system to become severe such as a lung infection (pneumonia). Other symptoms include high fever (body temperature above 38^0C), dry cough, flu, runny nose, sore throat, and shortness of breath. The spread of the virus works within the last 2 weeks of contact in areas prone to Covid-19. Patients should undergo self-quarantine to minimize virus transmission. Patients with severe symptoms often have to be referred for hospitalization. Some of them even have to get special treatment in the ICU.

Continued increase number of patients who had to be hospitalized had overwhelmed the hospital. The limited number of ICU became problems when many patients require special treatment. Delays and inappropriate treatment can be fatal for the patient which leads to death. In such an emergency situation, ICU must be prioritized for those who really need it. It is necessary to have criteria for Covid-19 patients, so that the use of the ICU is optimal for patients who really need intensive care [1].

SARS-RAS Italian Society of Hypertension has conducted a study on Covid-19 patients in Italy who were admitted to 26 hospitals in Italy. From the observations, it was found that the average patient came with hypertension, diabetes, chronic obstructive pulmonary disease, coronary artery disease, heart failure, obesity, and chronic kidney disease [2]. Researchers also conducted a cross-sectional, multicenter observational national survey, which found that the prevalence of diabetes, coronary artery disease, and chronic kidney disease was lower in women than men. The survey was also used to understand gender differences in predictors of intensive care units admission among COVID-19 patients. Of the total study population, there were 395 patients in Italy who were admitted to the ICU. Those patients are more often male, with higher criteria for obesity, hypertension, diabetes, chronic kidney disease, and heart failure. So that the dominance of ICU recipients are 291 male patients and 104 female patients [2]. Heo et.al. [3] conducted a research to predict patients requiring ICU in South Korea by implementing multivariate logistic regression models. An integer-based scoring system was developed for predicting Covid-19 patients requiring intensive care, with high performance.

In this paper the authors identified what data need to be considered to determine whether a Covid-19 patient requires an ICU. Furthermore, a classification model of patients requiring ICU were built to prioritizing patients who need to receive treatment in the ICU when the availability ICU is limited. The model hopefully may aid decision to save the lives of Covid-19 patients.

To build the classification model, the authors used Support Vector Machine (SVM) algorithm. Support Vector Machine (SVM) is a classification algorithm for finding hyperplanes that can separate two sets of data from two different classes [4]. The concept of SVM classification starts from the problem of classifying two classes so that it requires positive and negative training sets.

II. RESEARCH METHODOLOGY

A. Data Description

The data used in this study is the Covid - 19 patient pre-condition dataset (Covid-19 data from the Mexican government) obtained from the Kaggle website [5]. This data was uploaded in Kaggle website on 19 July 2020 and updated on 22 July 2020 by Tanmoy Mukherjee. The researchers downloaded the data on November 24, 2020. For this research, the researchers use 58.769 rows and 23 columns (attributes). Several data attributes were selected in this research as described in preprocessing stage. The research was carried out by classifying ICU attributes using the SVM algorithm. The ICU attribute has two classification classes, namely 1 (yes) which represent patient who requires ICU and

2 (no) which represent patient who does not require ICU. Table 1 describes attributes of data to be processed.

TABLE I. DESCRIPTION OF ATTRIBUTES

No	Attribute	Description	Explanation
1	id	Patient ID	String type, example: 11ea2f
2	sex	Gender	1 = Female, 2 = Male
3	patient_type	Type of patient	1 = Outpatient, 2 = Inpatient
4	entry_date	Date of admission to hospital	Date type
5	date_symptoms	Date of first symptom	Date type
6	date_died	Date of death	Date type
7	intubed	Have a history of being intubated	1 = Yes, 2 = No 97,98,99 = NA
8	pneumonia	Have a history of pneumonia	1 = Yes, 2 = No, 97,98,99 = NA
9	age	Age	Numeric, Example: 45
10	pregnancy	Pregnant	1 = Yes, 2 = No, 97,98,99 = NA
11	diabetes	Have a history of diabetes	1 = Yes, 2 = No 97,98,99 = NA
12	copd	Have a history of chronic obstructive pulmonary disease	1 = Yes, 2 = No, 97,98,99 = NA
13	asthma	Have a history of asthma	1 = Yes, 2 = No, 97,98,99 = NA
14	inmsupr	Lack of immunity	1 = Yes, 2 = No, 97,98,99 = NA
15	hypertension	Have a history of hypertension	1 = Yes, 2 = No, 97,98,99 = NA
16	other_disease	Have a history of other diseases	1 = Yes, 2 = No, 97,98,99 = NA
17	cardiovascular	Have a history of heart disease	1 = Yes, 2 = No, 97,98,99 = NA
18	obesity	Experiencing obesity	1 = Yes, 2 = No, 97,98,99 = NA
19	renal_chronic	Have a history of chronic kidney	1 = Yes, 2 = No, 97,98,99 = NA
20	tobacco	Smoker	1 = Yes, 2 = No, 97,98,99 = NA
21	contact_other_covid	Contact with Covid patients	1 = Yes, 2 = No, 97,98,99 = NA
22	covid_res	Covid status	1 = Positive, 2 = Negative, 3 = Waiting for Result
23	icu	Enter the intensive care unit	1 = Yes, 2 = No, 97,98,99 = NA

B. *Design of Research*

To build the classification model, the research was divided into three stage: pre-processing stage, training, and testing the model.

Pre-processing stage

Four activities were performed in the pre-processing stage namely data cleaning, data selection, data balancing, and data transformation.

1. Data Cleaning

At this stage, empty data and missing values indicated by values of 97, 98, 99 or empty (NaN) were removed. A new attribute, namely "incubation_period" was added, containing the difference between "entry_date" and "date_symptoms" attributes. Several other attributes, namely "id, intubed, entry_date, date_symptoms, date_died, age, pregnancy" were also removed. This stage was performed by using PyCharm IDE.

2. Data Selection

This stage was performed to select data and determine the attributes that will be used. The cleaned data was processed using PyCharm. Based on information gain, the attributes were ranked and resulted on the following rank as described in Fig. 1.

The pneumonia attribute has the highest rank. Based on the ranking, 16 datasets were prepared by utilizing spreadsheet. Each dataset represents different combination of attributes that will be used in the experiment. For each dataset, ICU attribute was included as the class label.

```
pneumonia              0.013132
patient_type           0.012756
cardiovascular         0.012246
other_disease          0.011657
inmsupr                0.011387
tobacco                0.011063
asthma                 0.009479
renal_chronic          0.008782
copd                   0.008676
obesity                0.008525
diabetes               0.007912
contact_other_covid    0.007533
sex                    0.006977
hypertension           0.006820
covid_res              0.005151
incubation_period      0.002229
```

Fig. 1. Ranking of The Attributes

3. Data Balancing

As can be seen in Fig. 2, the original dataset was imbalance due to the number of patients who needed ICU is less than those who did not need ICU. Among 58.769 data, 6.364 data represent patients required ICU. Imbalance data might result on unsatisfying classification, therefore a balancing process was performed toward the dataset by making a replica of the minority data. The balancing process

was carried out by applying SMOTE technique using WEKA application. Fig. 3 shows the data distribution after balancing.

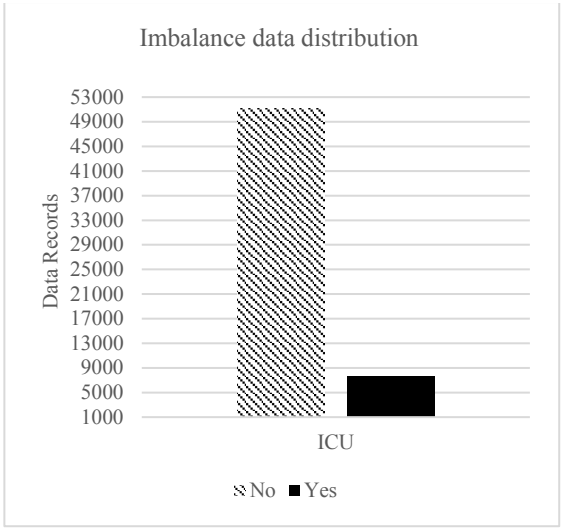


Fig. 2 Distribution of Original (Imbalance) Data

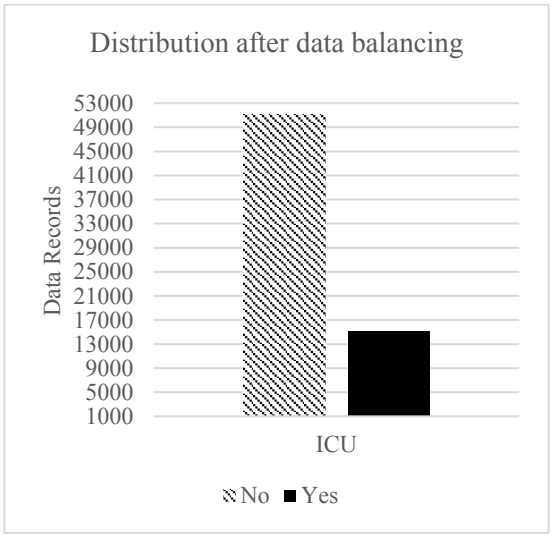


Fig. 3 Distribution Data after Balancing

4. Data Transformation

At this stage, data normalization was carried out so that each attribute in the dataset has the same weight. Normalization was done using min-max normalization.

Training & testing stage

After pre-processing, dataset was divided into training and testing set using k-fold cross validation technique. To get the optimum accuracy, in this research several experiments were performed by:

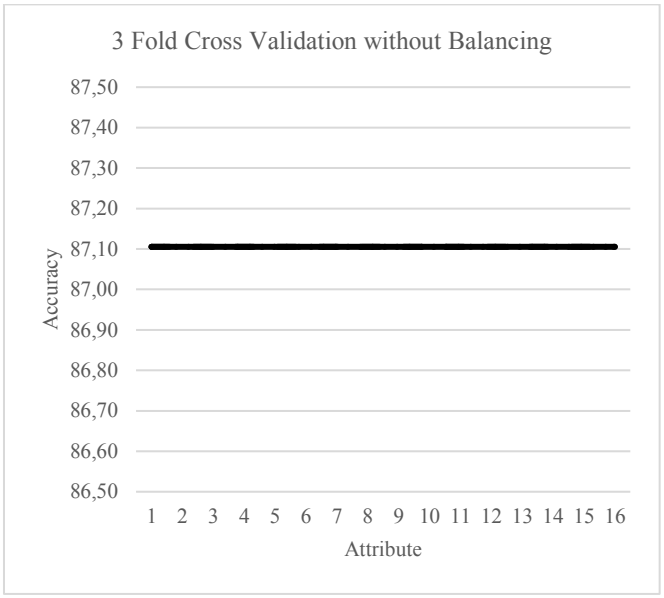
1. varying 3 types of SVM kernel, namely linear, Gaussian radial basic function (RBF), and polynomial kernel.
2. varying number of attributes in dataset
3. varying k values in k-fold cross validation

Those experiments were performed towards two groups of dataset: original dataset without balancing and dataset that has been balanced. All of the experiments were carried out by using Scikit-Learn containing SVM library.

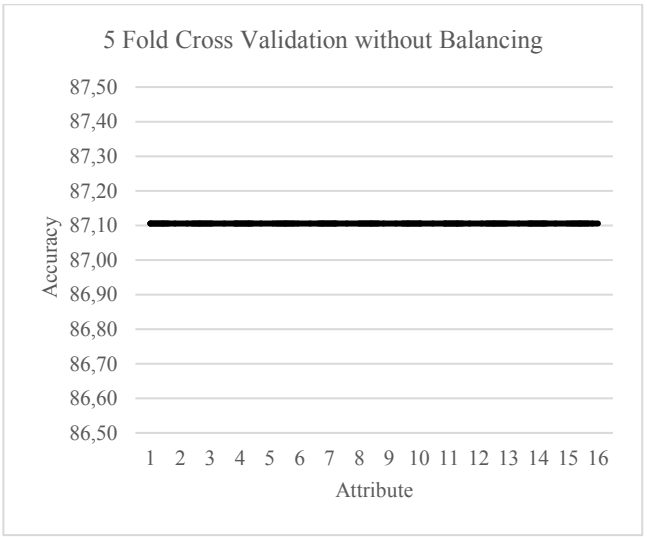
III. Results And Analysis

A. *Experiments Toward Dataset without Balancing*

Fig. 4 and Fig. 5 shows the result of experiments toward dataset without balancing using 3-fold cross validation and 5-fold cross-validation, accordingly. All variation of experiments with different number of attributes, different k-fold, and different type of kernel resulted on the same accuracy, which is 87.1055%. These two graphics show us that even using only by using one attribute (i.e. pneumonia), we can classify patient requiring ICU. Adding many more attributes in the model do not result on better accuracy.



. Fig. 4 Accuracy of Experiment using 3-fold Cross Validation towards Dataset without Balancing using Linear, RBF, Polynomial Kernel



. Fig. 5 Accuracy of Experiment using 5-fold Cross Validation towards Data without Balancing using Linear, RBF, Polynomial Kernel

B. *Experiment Towards Dataset After Balancing*

Fig. 6 and Fig. 7 shows the result of experiment towards dataset that has been balanced using 3-fold cross validation and 5-fold cross-validation, accordingly.

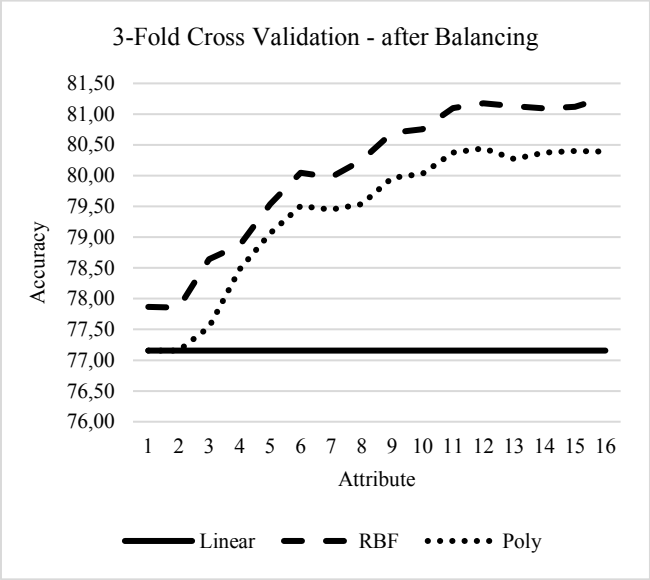


Fig. 6 Accuracy of Experiment using 3-fold Cross Validation Towards Dataset After Balancing

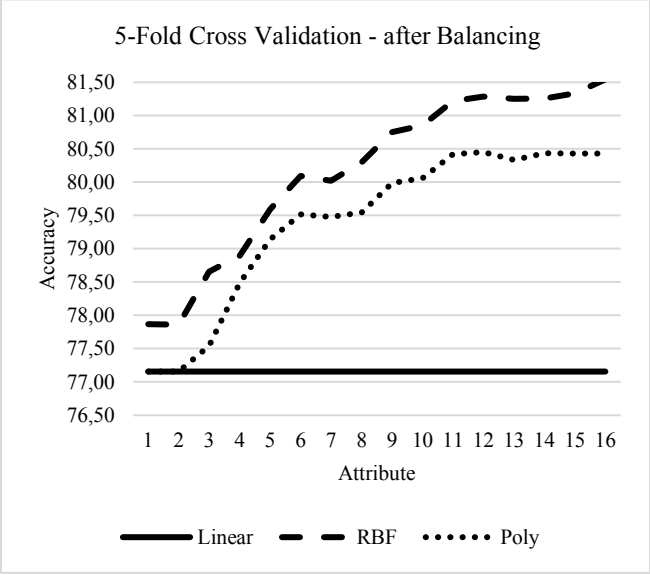


Fig. 7 Accuracy of Experiment using 5-fold Cross Validation Towards Dataset After Balancing

In Fig. 6, it is clearly shown that using linear kernel, the accuracy is stable at 77.1564% in any variation number of attributes. Using RBF kernel, the highest accuracy is 81.2712% that resulted from experiment with 16 attributes. Using polynomial kernel, the highest accuracy is 80.4437 % that resulted from experiment with 12 attributes.

Meanwhile, in the experiment with 5-fold cross validation, using linear kernel resulted on accuracy of 77.1564% in any variation of attributes. Using RBF kernel, the highest accuracy is 81.5439% that resulted from experiment with 16 attributes, while the highest of accuracy polynomial kernel is 80.4437 % that was obtained from experiment with 12 attributes.

IV. Conclusion

Based on the experiment, it can be concluded that SVM method can be used to classify Covid – 19 patients requiring ICU. The best accuracy that was obtained from experiment by varying type of kernel, number of attributes in dataset, and k values in k-fold cross validation, is 87.1055%. The accuracy was obtained using all types of kernel in 3-fold and 5-fold cross validation towards dataset without balancing. Variation number of attributes resulted on the same accuracy.

Further research might be performed by comparing different classification algorithm or different balancing techniques. In addition, since the research was performed on data taken from patients in Mexico, it might be biased towards race of patients. Validation of the model using data from different countries might be interesting and useful.

References

[1] H.K.N. Sumartiningtyas. "*Positif COVID – 19 Tinggi ICU Rumah Sakit Strategi dari Ahli"*. www.kompas.com/sains/read/2020/09/10/170100923/, Accessed December 3, 2020.

[2] G. Laccarino, et al., "Gender differences in predictors of intensive care units admission among COVID-19 patients: The results of the SARS-RAS study of the Italian Society of Hypertension", Plos One 16(9): e0257181, 2020, www.doi.org/10.1371/journal.pone.0237297, Accessed November 25, 2020.

[3] J.N. Heo, et.al., "Prediction of patients requiring intensive care for COVID-19: development and validation of an integer-based score using data from Centers for Disease Control and Prevention of South Korea", Journal of Intensive Care, 2021, https://jintensivecare.biomedcentral.com/articles/10.1186/s40560-021-00527-x, Accessed November 14, 2021.

[4] C. Cortes, and V. Vapnik, Support - Vector Network, Netherlands: Kluwer Academic Publisher, 1995.

[5] https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset