An XGBoost Model for Age Prediction from COVID-19 Blood Test

Nunung Nurul Qomariyah
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
nunung.qomariyah@binus.edu

Ardimas Andi Purwita
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
ardimas.purwita@binus.edu

Maria Seraphina Astriani
Computer Science Department,
Faculty of Computing and Media,
Bina Nusantara University,
Jakarta, Indonesia 11480
seraphina@binus.ac.id

Sri Dhuny Atas Asri
Head of Functional Medical Staff
Pulmonology Department,
Pasar Minggu Regional Hospital,
South Jakarta, Indonesia, 12550
dhunyatasasri@gmail.com

Dimitar Kazakov
Computer Science Department,
University of York,
York, UK, YO10 5GH
dimitar.kazakov@york.ac.uk

***Abstract*—COVID-19 was declared a pandemic by the World Health Organization (WHO) in January 2020. Many studies found that some specific age groups of people have a higher risk of contracting the disease. The gold standard test for the disease is a condition-specific test based on Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR). We have previously shown that the results of a standard suite of non-specific blood tests can be used to indicate the presence of a COVID-19 infection with a high likelihood. We continue our research in this area with a study of the connection between the patients' routine blood test results and their age. Predicting a person's age from blood chemistry is not new in health science. Most often, such results are used to detect the signs of diseases associated with aging and develop new medications. The experiment described here shows that the XGBoost algorithm can be used to predict the patients' age from their routine blood tests. The performance evaluation is very satisfactory, with $R^2 > 0.80$ and a normalized RMSE below 0.1.**

***Index Terms*—Artificial Intelligence, COVID-19, Coronavirus, Routine blood test, XGBoost, Decision Tree, Regression, Age Prediction**

I. Introduction

The 2019 Novel Coronavirus Disease (COVID-19) disease is caused by a virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2) which attacks the human respiratory system. The disease has spread all over the world and was declared a pandemic by the World Health Organization (WHO) in January 2020.

Many studies found that COVID-19 disease attacks differently to people of some age groups. The older age group above 50 [1] or 60 found to have higher risk when compared to the younger group in many countries [2]. Although, the age-based proportions of infected people differs from country to country.

Instead of examining broader view of countries scale, this study focuses on the age-based group analysis on their routine blood test. The dataset was collected retrospectively in 2020 from a public hospital in Jakarta, Indonesia. This study also performed some description analysis and conducted experiment on how the blood test can predict the age of the patients.

The importance of this study is to help the doctors and other health practitioners to analyse further about the disease in some specific age groups. The prediction can be performed not only by examining some specific patterns of the blood test in each age-based group patients, but also to conduct deeper analysis on the specific pattern found in the patients' blood test which can contribute to the age prediction. The doctors can then use our machine learning model to design a suitable clinical treatment for the patients. In other case, such as using telemedicine to treat COVID-19 patients, where people in some rural areas do not have hospital facility near them, the machine learning model can help the doctors to predict their real age by only examining from their blood test result. By using telemedicine, the blood result which can easily be obtained from any clinic which has laboratory can be uploaded to the system. The doctors will read and give advise from the distance.

Predicting age from blood chemistry is a common method in health science, nevertheless aging is a very complex process. In 2013, a research from King's College London has revealed that a panel of 22 metabolites - the small molecules - in the blood can be used to reveal age and possibly even reveal the signs of the disease associated with aging [3]. We can also predict age based on medical images. Using medical images to predict age has also been conducted by several studies for the clinical and forensic purposes. A study by Karargyris et al. [4] performed age prediction by using a large chest x-ray dataset.

Human organisms can change complexly as they age. The

transformation can affect any level, from organ system to cell organelles. Health scientists study this topic for the reason that the aging process in the human body is not a straight forward information [5].

The main contribution of this paper consists of: (1) reporting findings of age based study analysis from blood test data, and (2) make the pretrained machine learning model resulted from our experiment publicly available.

II. Related Work

Many studies found that people aged 50 years and older were having higher risk of mortality when exposed to COVID-19 disease as mentioned in [1]. Possible explanation is because older patients may have several common factors, including reduce immunity, low organ function, or coexisting comorbidities.

Based on a study conducted by Bauer et al. [6] which studied data of COVID-19 mortality cases collected from many different countries, they found that some age groups shows higher risk when compare to the other groups. The groups with less than one year old tend to have higher risk, then the risk decrease as they getting older, but start to increase again with age for young adults.

Another study which also conducted age analysis for COVID-19 case was performed by Jakhmola et al. [7]. The study shows that the reduced COVID-19 susceptibility for children may be caused by the reduced ACE-2 (Angiotensin-converting enzyme 2) expression, acts as the receptor for COVID-19 viral entry [8], in their nasal epithelium. Jakhmola et al. [7] performed analysis based on the data of confirmed case and deaths from many countries which is available publicly. They classified the data into three age-based group, i.e. under 19 years old, 20-49 years old, and above 50 years old. They found that the highest number of confirmed cases were in the middle age (20-49 years old) and old group (above 50 years old). For the middle age group, the increase number of confirmed case is possibly due to the higher exposure to the virus from outside activity, such as serving the society, working and other activities which require more contact with people. While the high number of confirmed case in the older age group can be caused by compromised immunity and prevalent health ailments, as also mentioned by similar findings from [9]. The other factor for the older age group, which related to their congregate living in some common facilities like nursing home, was also mentioned as possible caused.

A study conducted in Ontario, Canada, the children with COVID-19 case from January 2020 to June 2021, was found only 12.9% (70,187 cases) of the total of 545,398 cases. Within the children group, the rates of illness were highest on the group of older children, i.e. 14-17 years old [10]. Most cases of children were reported to have had previous close contact with confirmed adult cases, which may be responsible for transmission of the virus.

Another similar study, which focuses on age analysis of COVID-19, has also been performed by Monod et al. [11] in the United States. They studied the resurgent of the epidemics in 2020. They found that the adults, within the age of 20 - 49 years old, were the only age groups that have sustained COVID-19 transmission with reproduction number consistently above one. Reproduction number is a measurement of transmissibility of the infectious agents [12]. They also studied the correlation of the resurgence with the reopening of the schools. However, the results concluded that the evidence of the correlation was not very strong. Another study by Silveira [13] shows that the addition of patients' age as a feature with their blood count result in predicting COVID-19 can achieve a model's accuracy of 80.0% with XGBoost.

III. Research Methodology

A. Dataset

The study was conducted in accordance with current ethical and legal frameworks. Anonymized blood samples were obtained retrospectively from a COVID-19 referral hospital in Jakarta, Indonesia. From March 2020 to December 2020, we collected a sample of 1,000 inpatients data with confirmed COVID-19 case. Even though the title of this paper only mentions the COVID-19 blood test, we also collected a random sample of 1,000 inpatient with Pneumonia cases, and 1,000 inpatients with other disease, in the period before March 2020. This was done to find evidence whether the blood tests from COVID-19 patients have a robust pattern in predicting the patient's age.

Each patient blood test has been taken multiple times during their stay in the hospital. Each patient has been represented with several entries in the dataset. The data distribution in terms of age is shown in Figure 1. In this paper, the patient diagnosis in the dataset, i.e. 'COVID-19', 'pneumonia', and 'other' disease, will be discussed as the 'classes' (not to be confused with classification tasks). For each age group, the portion of classes is shown in Figure 2.

B. Data Preprocessing

We performed the same preprocessing technique as our previous study in [14]. The most important step is imputing the missing data. We performed the backward fill for imputation method and K-Nearest Neighbour (KNN) imputer method. The backward fill is used to impute the missing values from the same patient. The missing data was filled by using the previous data from the latest blood test. We assume that there is no change in the blood result, if no further test being administered. After this step has been performed, we then proceed with the KNN imputer to clean the whole dataset. The KNN works by finding the nearest k-neigbours, and fill the missing value by using the estimation found from the neighbours data. The neigbours is the data point which has the closest distance with the current data point based on the similarity metric.

We have class imbalance in the dataset, where for the COVID-19 class, the patients have more blood test entries than the other two classes. The data entries count in each class and the percentage of missing value before imputation

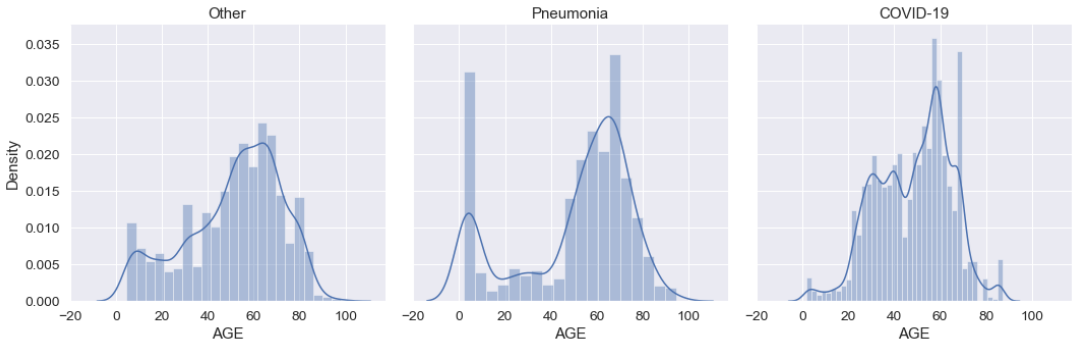
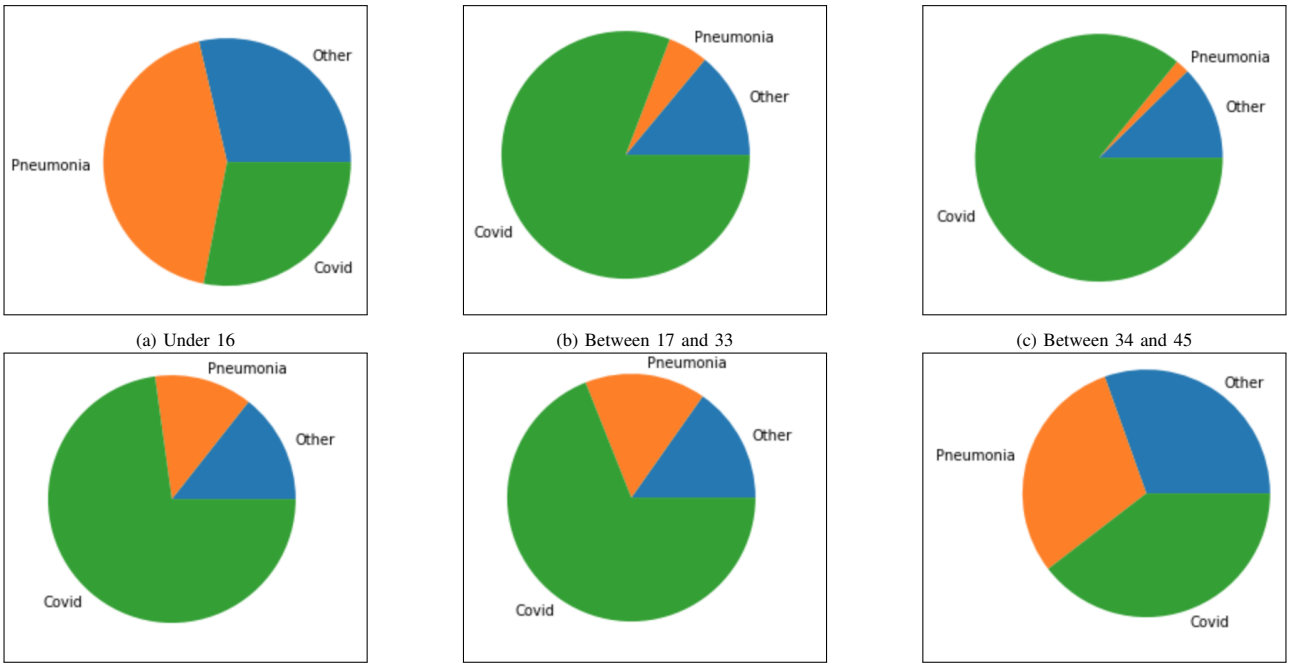


Fig. 1. Histogram of Age in Each Class



(a) Under 16 (b) Between 17 and 33 (c) Between 34 and 45

(d) Between 46 and 58 (e) Between 59 and 71 (f) Above 72

Fig. 2. Proportion of Class in Each Age Group

is shown in Table I. The reason why the two classes, Other and Pneumonia, have a large number of missing value in their data is because we select the feature based on the COVID-19 patient test. We used the COVID-19 patient data as the primary data, and the other two classes as additional information. The features that we used in the experiment is only 28. Detailed columns of the dataset used in the experiment can be seen in our previous paper [14, Table 1]. We also provide the summary in the Appendix (Table IV).

TABLE I. Dataset Count For Each Class

Class	Data Count	Missing Value Count
Other	2,634	48.12%
Pneumonia	3,136	44.36%
COVID-19	11,456	26.39%
Total Entries	17,226	32.99%

C. Machine Learning Technique

Our experiment focuses on predicting age based on the patients' routine blood test. We use five different regression algorithms, as explained in the following sections. Other than that, we also conducted the experiment with linear regression, however, the result was not included in this paper. This is due to the unacceptable performance and the nature of the data being non-linear.

1) Decision Tree: Both Decision Tree and XGBoost are included in our experiment. This is due to the satisfactory result we gained earlier in our previous study [14] which found that decision tree based algorithms (CART, XGBoost and Random Forest) performed well to explain the phenomena within the routine blood test for COVID-19 mortality prediction.

In this study, we use Decision Tree Regressor from Scikit-Learn Python package, which is an implementation of an optimized version of CART (Classification and Regression

Trees) algorithm. For regression task, we use the minimum Mean Squared Error (MSE or L2 error) as the criteria for determining future splits in the next node. In the experiment, we set the tree maximum depth as 20.

2) XGBoost: eXtreme Gradient Boosting is an highly efficient decision-tree-based ensemble machine learning library which implements the algorithm under Gradient Boosting framework. This algorithm can also be used for both classification and regression task. We run the experiment with the algorithm in several rounds to find the best setting and adjust the parameters. We found that the best parameters of the XGBoost model when trained with our dataset are as below:

- the objective function is mean squared error.
- maximum tree depth is 10, and
- learning rate is set to constant at 0.3.

3) Support Vector Machine: Support Vector Machine (SVM) is also used for both classification and regression problem. This algorithm was first proposed by Cortez and Vapnik from AT&T Bell Laboratories in 1995 [15]. SVM works by finding the optimal hyperplane that can maximize the distance between the classes. Later in 1997, they introduced SVM for regression [16] which works with the same principle but returns the continuous value. This algorithm is considered non-parametrics as its model relies on kernel functions. In our experiment, we use Radial Basis Function (RBF) as the kernel choice because our data is obviously cannot be described by using linear model.

4) Multi Layer Perceptron: MLPRegressor is an implementation of the Artifical Neural Network algorithm which used several nodes and hidden layers to estimate the outputs. These layers are inter-connected and pass the calculated values from the previous layers to the next ones. The model is built through iterative process by aiming to minimize the loss value when adjusting the parameters. In our experiment, we used Adam solver and initial learning rate 0.01. The size of hidden layer was 10.

5) Kernel Ridge Regression (KRR): This algorithm is a combination of the rigde regression (linear regression with L2-norm regularization) and the kernel trick, like in SVM. When compared to the SVM, it typically computes faster for medium size dataset. However, the loss function used is different. KRR uses squared error instead of epsilon-insensitive loss. The learned model is non-sparse, which make the learning time slower. For KRR, we used the polynomial kernel to evaluate the performance when compared to the SVR with RBF kernel.

D. Evaluation Technique

For the clustering task, we used k-means with elbow method to find the best number of clusters we can build from the data. Elbow method [17] is a well known heuristics calculation, which plots the value of average distortion at different values of k. Distortion means the distance from each cluster member to their respective centroids. The average distortion will decline as k increases, as each cluster will have fewer number of instance. The elbow is found when at specific k the distortion declined the most. At that point, there is no further cluster number will be better in explaining the data.

For regression task, we used the following metrics: R^2 (coefficient of determination), Root Mean Squared Error (RMSE) and Normalized RMSE. In regression, R^2 (R-squared) is a statistical measure which can be used to evaluate of how close are the prediction with the true value of the data. In our case, the prediction and the true value being compared is the age of the patients. Root Mean Squared Error (RMSE) is the measurement of how much is the error/residuals between the prediction and the true value. The lowest the RMSE the better the model fit. However, there is no bound in the RMSE value. In order to make better interpretation of the model's performance, we also use Normalized RMSE, which can facilitates the comparison of RMSE with different scales.

In order to evaluate the performance of our model, we use holdout technique, where we divide the data into 70% training and 30% testing. The holdout technique was chosen because the size of data used in the training was quite large, so there is no need to validate with cross-validation. Even though we also repeated the holdout technique several times to ensure the result are robust. Each model was trained on the same set of splitted data to avoid bias (by specifying the random state explicitly).

IV. Result and Discussion

A. Clustering Analysis

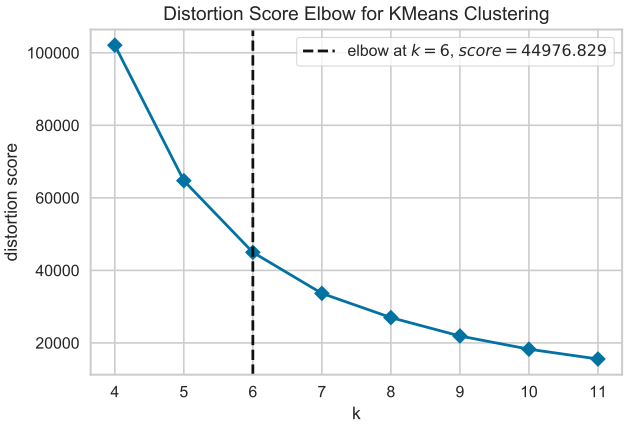


Fig. 3. Result of the Best Number of Clustering (k)

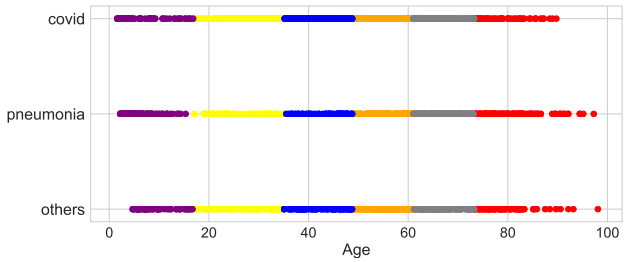


Fig. 4. K-Means Result (k=6)

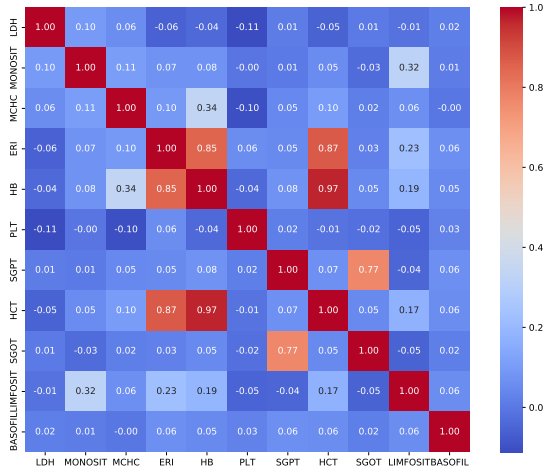


Fig. 5. 10 Most Important Features in Relation with Each Class

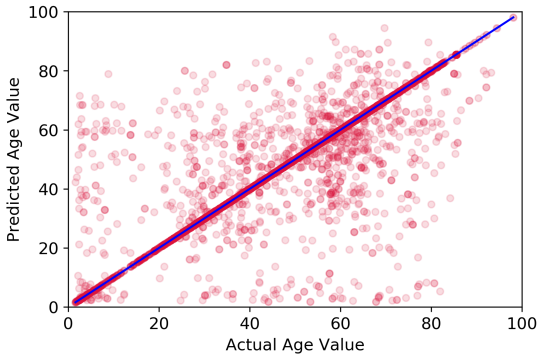
B. Regression Result

The task to predict age from routine blood test data can be performed differently by using five algorithms mentioned earlier. The result can be seen in Table II. While The result of trained model in each class is shown in Table III. An acceptable performance has been achieved by training the models with the data from all classes. The best performer was still hold by XGBoost algorithm, then followed by Decision Tree. When we observed in more detail for each class, the trained model in COVID-19 blood test shows a satisfactory result, followed by Pneumonia blood test as the second best. The R2 score shows the correlation of the *predicted* age value with the *true* age value. In the result, R2 score above 0.7 (for Decision Tree) and above 0.8 (for XGBoost) show that the strong correlation has been found. Both p-value also shows the significant correlation with the probability less than 0.05. Both models can predict the age very well. In addition, we also recorded the other measurement, which are RMSE and NRMSE. While the RMSE is difficult to be interpreted, we found that NRMSE result in both models was very low (close to zero). This result show that both models' error rate is very low. The high value of R^2 is also supported by the plot of actual versus predicted value shown in Figure 6. In Figure 6b, we can see that XGBoost prediction data is very close to the actual value. We also show the result of the Decision Tree prediction in Figure 6a. Decision Tree shows there are some couples correctly predicted value which lie next to the blue line (diagonal), while some predictions are spread out quite far from the diagonal line.

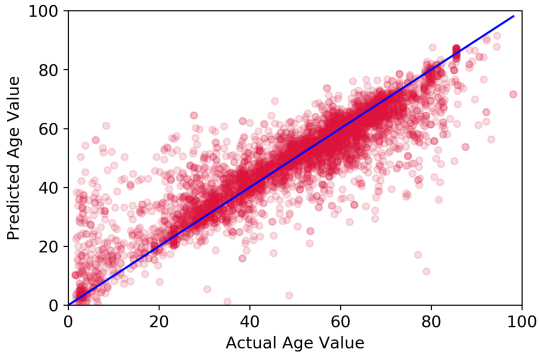
For the other three algorithms, SVR, MLP and Kernel Ridge Regression, the NRMSE score was also showing a low result. However, the R^2 score of these three models did not agree with the NRMSE result. The R^2 score shows a lower result, which means the correlation between the predicted age value and the true age value was very weak (less than 0.3). Therefore, we decided to not put the plot of actual versus predicted value, like the XGBoost and Decision Tree.

TABLE II. Regression Result from All Classes

	All Classes		
	R2 score	RMSE	NRMSE
XGBoost	0.82	8.23	0.08
Decision Tree	0.65	11.37	0.12
SVR	0.02	19.07	0.21
MLP	0.29	16.21	0.17
Kernel Ridge Regression	0.25	16.67	0.17



(a) Decision tree



(b) XGBoost

Fig. 6. Plot of actual versus predicted value from the model

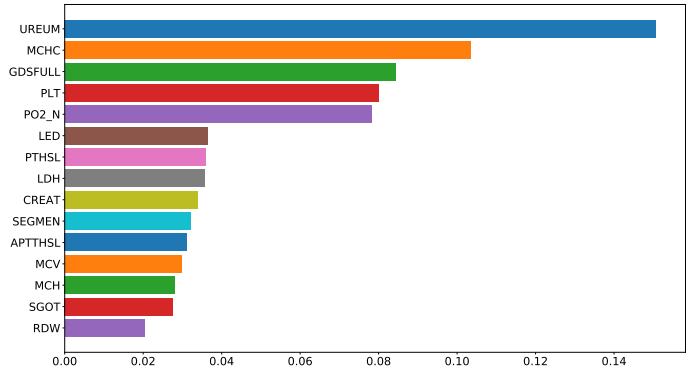
The top 15 important features found by the models when predicting age from the COVID-19 blood dataset are shown in Figure 7.

V. Conclusion and Future Work

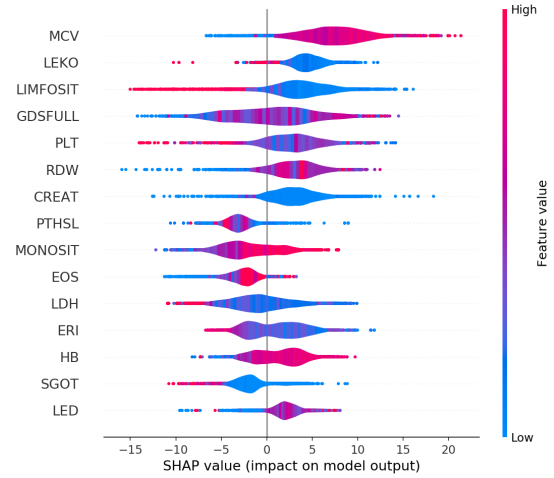
In this paper, we have explained our research on predicting age based on blood test from COVID-19 patients. As an additional information, we also compared the result with two other class, i.e. Pneumonia and other. The experiment shows that XGBoost and Decision Tree yielded a satisfactory result for age prediction. The prediction made from COVID-19 blood test data also shows the best performances when compare to the other two classes. In the future, we want to expand the research for COVID-19 mortality prediction from blood test data with correctly predicted patient's age. We want to examine whether these additional information can improve the prediction when compared to the blood test data alone.

TABLE III. Regression Result from Each Class

	R-squared (R^2) score			RMSE			NRMSE		
	COVID	Pneumonia	Other	COVID	Pneumonia	Other	COVID	Pneumonia	Other
XGBoost	0.87	0.86	0.80	6.15	8.91	9.32	0.07	0.09	0.10
Decision Tree	0.76	0.76	0.56	8.46	11.72	13.79	0.09	0.12	0.16
SVR	0.01	0.00	0.01	17.24	23.89	20.73	0.19	0.25	0.24
MLP	0.16	0.54	0.48	15.94	16.10	15.01	0.18	0.17	0.17
Kernel Ridge Regression	0.25	0.20	0.39	14.97	21.42	16.23	0.16	0.23	0.19



(a) Decision tree important features



(b) XGBoost SHAP value showing top important features

Fig. 7. Important features for predicting age from COVID-19 blood test dataset

Consent and Ethical Clearance

The patient medical records used in this study were collected by the data provider, including epidemiological, demographic, clinical, laboratory and mortality outcome information. This study has been approved by the Ethics Committee of the data provider, Pasar Minggu Regional Hospital in Jakarta. The requirement for patient's consent was waived as this was a secondary analysis of anonymized data.

Supplementary Files

The XGBOOST pretrained model is available at our Github repository: https://github.com/nnqomariyah/covid-age-prediction. Please cite this paper to use the model.

Acknowledgment

This work was partially supported by the Research and Technology Transfer Office of Bina Nusantara University as part of Bina Nusantara University's International Research Grant entitled "AI-Based Telemedicine for COVID-19 Patients" with the contract number of 017/VR.RTT/III/2021 and the contract date of 22 March 2021. The last author was supported by the British Council Newton Institutional Links Research grant "AI-Based Telemedicine for COVID-19 Patients".

Appendix

TABLE IV. Biomarkers Used in Dataset

Biomarker	**Feature code**
HEMATOLOGY	
Hemoglobin	HB
Hematocrit	HCT
Leukocytes	LEKO
Platelets	PLT
Erythrocytes	ERI
Red Cell Distribution Width	RDW
AVERAGE ERYTHROCYTE VALUE	
Mean Corpuscular Volume	MCV
Mean Corpuscular Hemoglobin	MCH
Mean Corpuscular Hemoglobin Concentration	MCHC
COUNT TYPE	
Basophils	BASOFIL
Eosinophils	EOS
Stem Neutrophils	NEUTB
Segmented Neutrophils	SEGMEN
Lymphocytes	LIMFOSIT
Monocytes	MONOSIT
Neutrophil-Lymphocyte Ratio	NLR1
Erythrocyte Sedimentation Rate	LED
HEMOSTASIS	
D-Dimer	DDIMER
prothrombin time	PTHSL
Activated Partial Thromboplastin Time	APTTHSL
BLOOD CHEMISTRY	
Arterial blood gas analysis	
Partial pressure of oxygen	PO2_N
Oxygen saturation	O2S_N
Liver function	
Serum Glutamic Oxaloacetic Transaminase	SGOT
Serum Glutamic Pyruvic Transaminase	
Diabetes	
Random Plasma Glucose Test	GDSFULL
Kidney Function	
Urea	UREUM
Creatinine	CREAT
Cardiac enzymes	
Lactate dehydrogenase	LDH

REFERENCES

[1] M. Biswas, S. Rahaman, T. K. Biswas, Z. Haque, and B. Ibrahim, "Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis," 2021.

[2] J. Q. Liu, J. W. Xu, C. Y. Sun, J. N. Wang, X. T. Wang, X. Chen, and S. L. Gao, "Age-stratified analysis of SARS-CoV-2 infection and case fatality rate in China, Italy, and South Korea," *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 23, 2020.

[3] C. Menni, G. Kastenmüller, A. K. Petersen, J. T. Bell, M. Psatha, P.-C. Tsai, C. Gieger, H. Schulz, I. Erte, S. John, M. J. Brosnan, S. G. Wilson, L. Tsaprouni, E. M. Lim, B. Stuckey, P. Deloukas, R. Mohney, K. Suhre, T. D. Spector, and A. M. Valdes, "Metabolomic markers reveal novel pathways of ageing and early development in human populations," *International Journal of Epidemiology*, vol. 42, no. 4, pp. 1111–1119, 8 2013. [Online]. Available: https://academic.oup.com/ije/article/42/4/1111/657994

[4] A. Karargyris, S. Kashyap, J. T. Wu, A. Sharma, M. Moradi, and T. Syeda-Mahmood, "Age prediction using a large chest x-ray dataset," in *Proc. SPIE 10950, Medical Imaging 2019: Computer-Aided Diagnosis, 109501U*, vol. 10950. SPIE, 3 2019, pp. 468–476.

[5] Z. Yu, G. Zhai, P. Singmann, Y. He, T. Xu, C. Prehn, W. Römisch-Margl, E. Lattka, C. Gieger, N. Soranzo, J. Heinrich, M. Standl, E. Thiering, K. Mittelstraß, H.-E. Wichmann, A. Peters, K. Suhre, Y. Li, J. Adamski, T. D. Spector, T. Illig, and R. Wang-Sattler, "Human serum metabolic profiles are age dependent," *Aging Cell*, vol. 11, no. 6, pp. 960–967, 12 2012.

[6] P. Bauer, J. Brugger, F. König, and M. Posch, "An international comparison of age and sex dependency of COVID-19 deaths in 2020: a descriptive analysis," *Scientific Reports*, vol. 11, no. 1, 2021.

[7] S. Jakhmola, B. Baral, and H. C. Jha, "A comparative analysis of COVID-19 outbreak on age groups and both the sexes of population from India and other countries," *Journal of Infection in Developing Countries*, vol. 15, no. 3, 2021.

[8] L. Samavati and B. D. Uhal, "ACE2, Much More Than Just a Receptor for SARS-COV-2," *Frontiers in Cellular and Infection Microbiology*, vol. 10, 2020.

[9] W.-j. Guan, Z.-y. Ni, Y. Hu, and others, "Clinical Characteristics of Coronavirus Disease 2019 in China," *New England Journal of Medicine*, vol. 382, no. 18, 2020.

[10] Public Health Ontario, "COVID-19 Infection in Children: January 15, 2020 to June 30, 2021," 2021. [Online]. Available: https://www.publichealthontario.ca/-/media/documents/ncov/epi/2020/05/covid-19-epi-infection-children.pdf?la=en

[11] M. Monod, A. Blenkinsop, X. Xi, and others, "Age groups that sustain resurging COVID-19 epidemics in the United States," *Science*, vol. 371, no. 6536, 2021.

[12] N. C. Achaiah, S. B. Subbarajasetty, and R. M. Shetty, "R0 and Re of COVID-19: Can we predict when the pandemic outbreak will be contained?" *Indian Journal of Critical Care Medicine*, vol. 24, no. 11, 2020.

[13] E. C. Silveira, "Prediction of COVID-19 From Hemogram Results and Age Using Machine Learning," *Frontiers in Health Informatics*, vol. 9, no. 1, 2020.

[14] N. N. Qomariyah, A. Andi Purwita, S. D. Atas Asri, and D. Kazakov, "A Tree-based Mortality Prediction Model of COVID-19 from Routine Blood Samples," in *International Conference on ICT For Smart Society (ICISS)*. Institute of Electrical and Electronics Engineers (IEEE), 9 2021, pp. 1–7.

[15] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, 1995.

[16] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapnik, and others, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[17] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.