The 2nd International Conference on Multidisciplinary Engineering
(2nd ICOMDEN 2024)

The 2nd International Conference on Multidisciplinary Engineering: 00015(2024) https://doi.org/10.29103/icomden.v2.709
Volume 2, 2024
eISSN: xxxx-xxxx
Research Original Article

Implementation Of Support Vector Regression In Prediction Air Quality Index In Banda Aceh city

Rizky Fasya Ramdhani✉[1] **Rozzi Kesuma Dinata**[2] **Ar Razi**[3]
[1]Department of Informatic Engineering, Malikussaleh University, Bukit Indah, Lhokseumawe, 24353, Indonesia, rizky.200170266@mhs.unimal.ac.id
[2]Department of Informatic Engineering, Malikussaleh University, Bukit Indah, Lhokseumawe, 24353, Indonesia, rozzi@unimal.ac.id
[3]Department of Informatic Engineering, Malikussaleh University, Bukit Indah, Lhokseumawe, 24353,Indonesia, ar.razi@unimal.ac.id
✉Corresponding Author: rizky.200170266@mhs.unimal.ac.id**author | Phone: +6282287713204**

Abstract

Air quality is one of the important aspects in maintaining environmental balance and public health. Increasing air quality in the environment is a matter of concern. Therefore, a method that can predict the Air Quality Index (AQI) effectively is needed to be able to monitor and support decision making on environmental impacts. This study aims to predict the Air Quality Index in Banda Aceh City using the Support Vector Regression algorithm, with five main parameters used in the study, namely particulate matter ($PM_{2.5}$), Sulfur dioxide ($S0_2$), Nitrogen dioxide ($N0_2$), Carbon monoxide (CO), and Ozone (0_3). In this research, the Support Vector Regression algorithm was chosen because of its ability to handle non-linear data and also because it can provide accurate predictions on data. The prediction system designed will be web-based using the flask framework and MySQL database, while the Support Vector Regression modeling will be done on google colab for the media used. In the process of modeling the data will be divided into 80% training data and 20% test data to ensure the model can capture long and short-term patterns. The results of the prediction will be compared using the Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) evaluation metrics. The results of the evaluation using both metrics yielded RMSE values of 1.9001 and MSE of 3.6015. These values indicate good performance of the model in predicting the data. This research is expected to provide insight for future similar research in terms of prediction using the Support Vector Regression algorithm.

Keywords: Support Vector Regression; Air Quality Index; Machine learning

INTRODUCTION

Air is a vital element of the environment that is essential for the survival of organisms. However, the development of the transportation sector and industrial expansion in modern times have the potential to threaten air quality. Human activities and industrial activities can produce pollutants that degrade air quality. This phenomenon, known as air pollution, occurs when polluting substances in both particulate and gaseous forms enter the atmosphere in a certain quantity and duration, causing negative effects on living things [1].

Air pollution has now become a worldwide environmental issue. According to World Health Organization (WHO) statistics, air pollution causes about 2 million premature deaths each year. Sources of air pollution can come from natural processes or anthropogenic activities. However, the majority of air pollution cases originate from fossil fuel utilization and industrial sector activities. In Indonesia, the air pollution problem has reached an alarming stage. The main factor is the large population accompanied by the high use of motorized vehicles, which produce significant amounts of exhaust emissions. In addition, the number of industrial zones that have not fully complied with emission regulation standards has contributed to the decline in air quality in various regions of Indonesia [2].

From these problems, this research wants to propose a system that can predict the air quality index, where the prediction results will be expected to help better air quality monitoring. The system will use machine learning algorithms. Machine learning (ML) is a branch of artificial intelligence (AI) that is widely applied to replicate or replace human functions in problem solving and automation [3]. Support Vector Regression is a machine learning algorithm that will be used. Support Vector Regression (SVR) was introduced by Vladimir N. Vapnik in 1999. SVR is an application of the Support Vector Machine (SVM) method in the context of regression[4]. The SVR method is used because this method has a good ability to handle data because the SVR method is designed to minimize overfitting this has an impact on stable prediction results.

The purpose of this research is to develop a system that implements the Support Vector Regression algorithm to be able to predict the air quality index. With this system, it is expected that both the community and the government are aware of air pollution in the community.

LITERATURE REVIEW

Air Quality Index

The Air Quality Index (AQI) is an indicator used to describe the status of air quality in an area based on the level of pollution in the area[5]. The AQI is calculated using several air pollutant parameters such as particulate matter ($PM_{2.5}$), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and nitrogen dioxide (NO_2). The AQI scale is a standard used to look at air quality conditions in a simple way for ordinary people to understand what the level of air pollution is based on a set index. The higher the AQI value, the worse the quality[6]. Based on the latest US EPA standards the AQI scale can be seen in the table below.

Table 1 Air Quality Index indicator

Level of concern	Values of index	Description of air quality
Good	0 – 50	Air quality is satisfactory, and air pollution poses little or no risk.
Moderate	101 – 150	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Unhealty fo sensitive groups	151 – 200	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Unhealty	201 - 250	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealty	251 – 300	Health alert: The risk of health effects is increased for everyone.
Hazardous	301-350	Health warning of emergency conditions: everyone is more likely to be affected.

Machine learning

Machine learning involves programming computers to improve their performance on a given task by learning from example data or past experiences[7]. Machine learning (ML) helps machines work with data more efficiently. Sometimes, even after looking at the data, we can't easily find useful information. That's when we turn to machine learning[8]. Through powerful algorithms, ML can sift through huge datasets and identify patterns and information otherwise imperceptible. This capacity allows us to predict more accurately, and to act more wisely, in every domain. From predicting consumer behaviour in advertising to disease diagnosis in medicine to optimising production on the factory line, machine learning is the tool to extract valuable knowledge from data sets that are too large to be explored by hand. The more data there is, the more machine learning will play in converting it to something useful.

Support Vector Regression

Support Vector Regression (SVR), the invention of Vladimir N.Vapnik in 1999. Support Vector Regression (SVR) is a generalisation of the Support Vector Machine (SVM) algorithm to the regression domain. SVM and SVR are identical except the application. The goal of SVM is to find the best separating hyperplane that can divide two types of objects by maximizing the distance between them, and SVR is finding some function whose error is no more than the deviation from the actual target value.

The formula for the support vector regression algorithm is as follows:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\mathrm{K}(x_i, x_j) + \mathrm{b} \qquad (1)$$

$\alpha_{i,}$ = Sets the contribution of data points that are above the margin of the prediction function
α_i^* = Sets the contribution of data points that are below the margin of the prediction function
K = Kernel function
x_i = training data
x_j = Data input for prediction
B = Bias

Hyperparameter Tuning

Hyperparameter tuning is the process of finding the optimal parameter values of the model used that cannot be obtained directly from the data. This parameter has a significant impact on the performance of a model, for this optimal parameter search is not done during the data training process but is done before the model training process. Hyperparameter tuning is often more important than choosing a machine learning algorithm[9].

Grid search optimization will be used to determine the hyperparameter that will be run with the *k-fold cross validation* technique. The*hyperparameter* that will be searched for is .In the value of k in *cross validation*, 3 values will be tested, namely 3, 5, and 10. The test results that have the smallest error will be used in model development.

Table 2 Result of Hyperparameter Tuning

k	ε	C	γ	error
3	0.01	100	0.1	2.11678
5	0.01	100	0.1	2.11578
10	0.1	100	0.1	2.10470

The results of *grid search* using *cross validation* show differences in the value of some parameters, specifically using k-10 the epsilon value obtained is 0.1 this epsilon value is different from *cross validation* using k 3 and 5 which is the same value of 0.01. From the test results, the error value obtained using k-10 is smaller than the others. The smallest error value obtained is 2.10470 with parameters ε =0.1, C = 100, γ = =0.*1* , then when the value of k = 3 with parameters ε =0.01, C = 100, γ =0. 1 the smallest error is 2.11678, then for k = 5 with parameters ε =0.01, C = 100, γ =0.*1* the error value obtained is 2.11578. So with this based on the smallest error, the parameter to be used is ε =0.1, C=100, γ =0.1.

MATERIALS & METHODS

Dataset

In this study the data used were collected from various sources to start making the model to be used. In this study the data to be used is sourced from Dinas Lingkungan Hidup, kebersihan kota Banda Aceh (DLHK3), the selection of appropriate data is needed so that this research can run smoothly, in this study the data variables to be used are particulate matter , *Sulfur dioxide* , *Nitrogen dioxide* , *Carbon monoxide*, and *Ozone* . This variable has been proven to affect the quality index based on previous research journals.

Table 3 Dataset of Air Quality Index

Date	$PM_{2.5}$	SO_2	CO	0_3	NO_2	IKU
28/2/2021	54	31	1	5	45	98
1/3/2021	65	0	2	33	41	110
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
22/12/2023	31	48	3	36	23	69
23/12/2023	30	50	3	37	23	68

The table is data on the air quality index and its air pollutants from January 2021 to December 2023. This dataset is an important material in the research to analyze the air quality index and to predict it. before being used to build this data model, it will be processed first according to a predetermined method.

Methods

In this research, the research process will use systematic steps that organize the process to get the desired thing. This systematic process will lead to getting the final output of the research. This process is important because it will facilitate the research in terms of its work. for the research scheme can be seen in the figure below

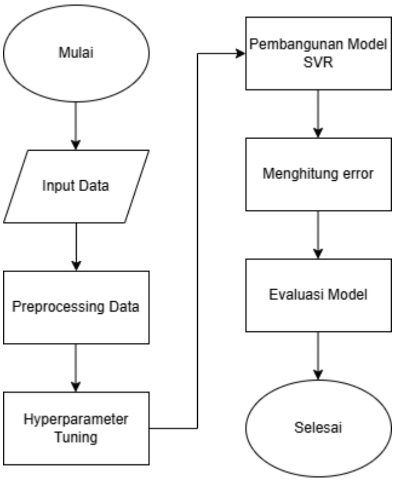

Figure 1 Research scheme

- Input data, At this stage data will be collected from various sources to start making the model to be used. In this study, the data to be used is sourced from the Banda Aceh Environmental and Hygiene Service (DLHK3), the selection of appropriate data is needed so that this research can run smoothly.
- Preprocessing Data the data will be subjected to several stages of processing starting from cleaning the data, normalizing and dividing the data into *training* and *testing* data. This aims to improve data quality and so that data can be easily processed when modeling.
- *Hyperparameter* Tuning, is the process of finding the best combination of parameters in the model. This is necessary due to the uncertainty of the quality of the parameters used randomly, which will have an impact on the performance of the model.
- Model Building, The model used is *Support Vector Regression,* SVR is a model that has advantages in generalization and is effective for non-linear data. Although this research will use the help of Google Colab, understanding how SVR works manually is still important to provide deeper insights for researchers.

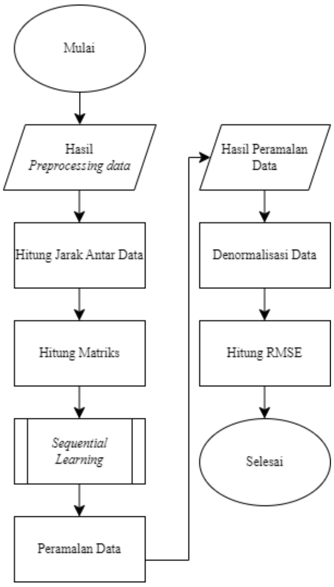

Figure 2 SVR Flowchart

In the SVR workflow there is a *predefine process* which is a procedure for *sequential learning,* the following *flowchart* at the *sequential learning* stage

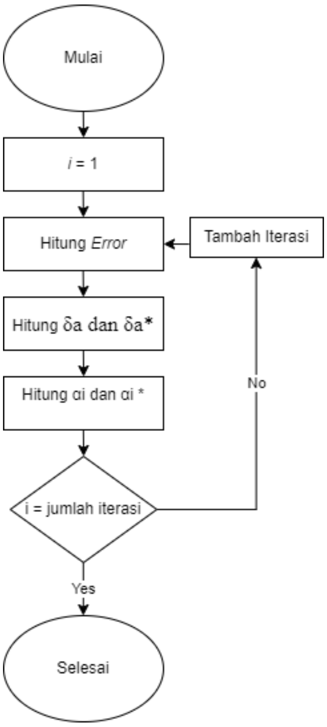

Figure 3 Sequential Learning Flowchart

- The model will be evaluated using the Root Mean Squared Error (RMSE) metric. RMSE will calculate the average error of the predicted value by squaring the difference and the original value, then taking the square root. The performance of the model will be seen from the RMSE value, the lower the RMSE value, the better the performance of the model in making predictions. This value provides an overview of the results of the model prediction with the original data.

RESULTS AND DISCUSSION

Analysis of data characteristics is an important stage before the model is built because each data has many different patterns and structures. By doing this analysis, the author can understand outliers, missing values, and the distribution of each variable. At this stage, the data used is historical ISPU data in Banda Aceh City, this data has 5 variables which are air pollution substances that affect the air quality index, the period to be used is from 1-January-2021 to 31-December-2023.

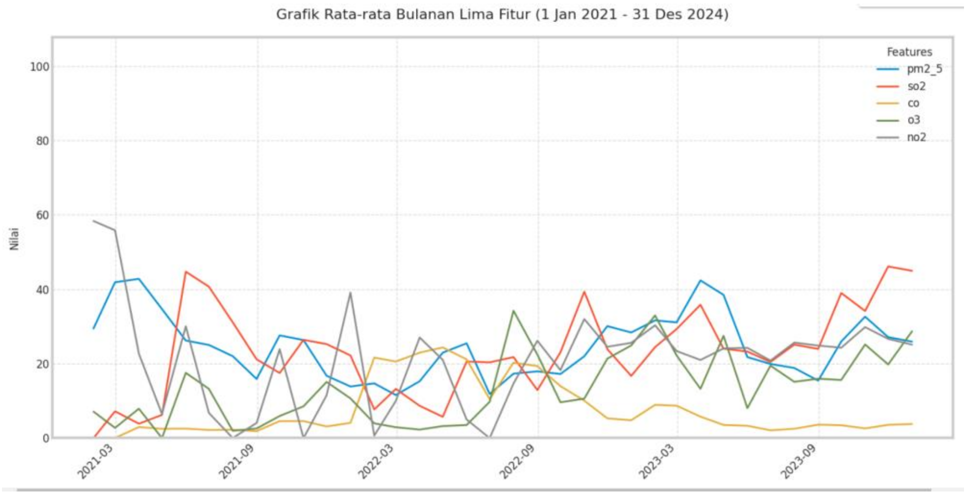

Figure 4 Feature data graph

The graph shows the changes in the monthly average values of five air quality parameters between January 2021 and December 2024. Each parameter experiences different fluctuations, with some, such as PM2.5 and NO2, showing sharp spikes in certain months that may be influenced by seasonal factors or human activity, such as increased pollution from vehicles or industry. Meanwhile, CO appears relatively stable without many major changes, suggesting the possibility of better regulation or consistent emission sources. On the other hand, some parameters, especially SO2 and PM2.5, show an upward trend at the end of the period, which could indicate an increase in emissions in recent years, possibly due to population changes or industrial activity. In addition, there are similar patterns in some parameters such as PM2.5 and SO2, suggesting that they may be influenced by the same pollution sources or similar weather conditions. These graphs provide important insights into seasonal and annual variations in air quality, which can form the basis for formulating future pollution control strategies.

Furthermore, to understand more about this AQI data, a *correlation map* visualization will be carried out on this data, this aims to see how close the relationship is between two or more variables in the dataset. This is certainly needed in this research because the author gets information on what variables are quite influential in this AQI.

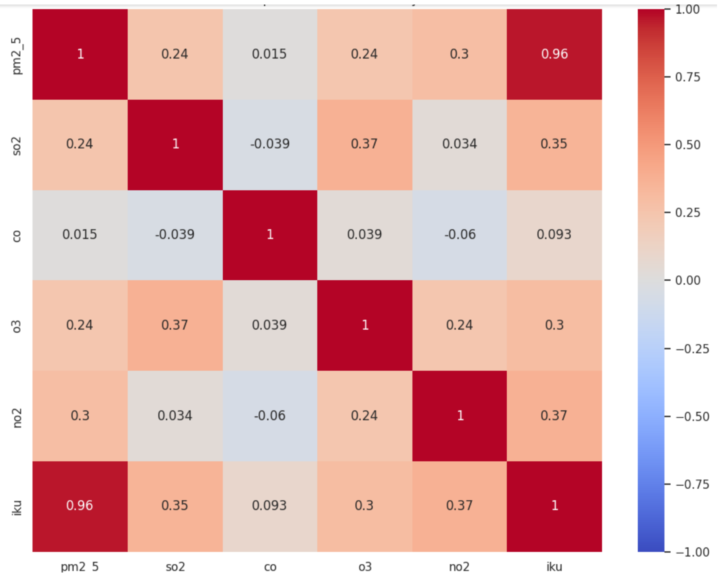

Figure 5 Correlation Heatmap feature

Based on the *Correlation map* in the figure above some variables appear to have a positive correlation, based on the figure PM2.5 and SO2 have a positive correlation indicating that both tend to increase or decrease together, then CO and NO2 and also have a strong positive correlation. Meanwhile, PM 2.5 and O3 have a weak negative correlation, indicating that higher levels of PM 2.5 can slightly reduce the formation of ozone.

Furthermore, the data that has been analyzed will be used to create machine learning in which this research is a Support Vector Regression algorithm. In making this model there are several things that need to be determined, the parameters used, the determination of the data division and also the kernel. then for the determination of these things will be taken from the previous analysis. the parameters used are the parameters of the results of hyperparameter tuning, then for the division will be divided into 80% training data and 20% testing data and finally for the kernel that will be used will use the Radial Basis Function kernel. then with that the model will be assisted by google colab, after that the model will be tested to see its ability to predict data by comparing the results of actual data prediction with the prediction data. The following is a graph of the data comparison

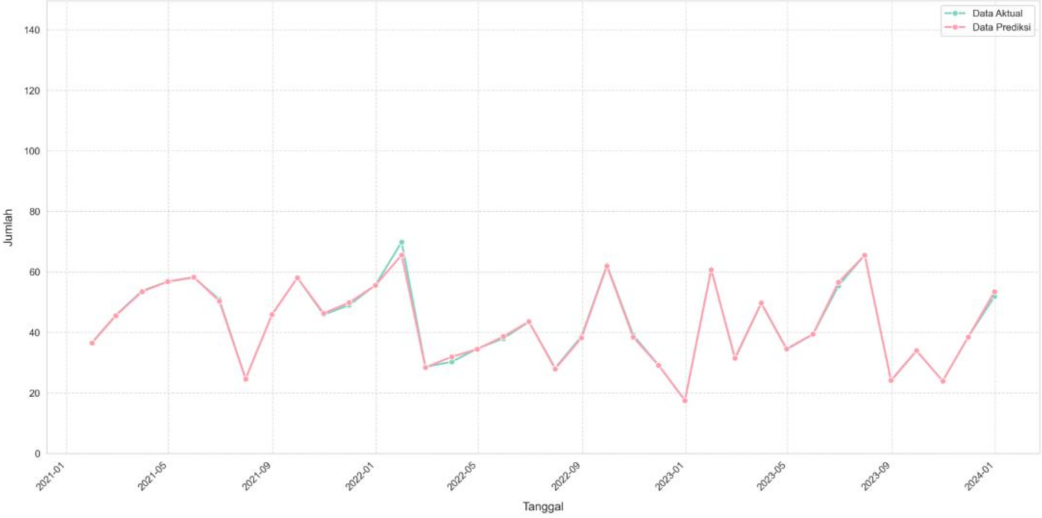

Figure 6 Actual data vs Prediction data

The results of the prediction show the predicted value and evaluation metrics of the mode that has been built. The parameters used in this model are ε =0.1, C=100, γ =0.1 . The table shows the average RMSE value obtained is 1.90014 while the MAPE value obtained is 3.61056. Then after the model has been successfully built, predictions will be made for the next 15 days. The following are the prediction results

Table 4 Future prediction data

Tanggal	Prediksi
01/01/2024	38.28195062
02/01/2024	17.40582056
03/01/2024	48.06461586
04/01/2024	62.69645518
05/01/2024	45.41382753
06/01/2024	70.17113237
07/01/2024	60.90941559
08/01/2024	68.12318152
09/01/2024	68.7825963
10/01/2024	66.67162214
11/01/2024	67.10215921
12/01/2024	67.03219927
13/01/2024	67.13690416
14/01/2024	67.19109285
15/01/2024	67.17294115

CONCLUSIONS

The results that have been carried out in this study to predict the air quality index using the Support Vector Regression algorithm are successful with the prediction value obtained against the actual value is very small. The error value obtained in the model for both evaluation metrics RMSE 1.9001 and MSE 3.6105 shows the effectiveness of the model in prediction is quite good. The results of the study also found that the PM 2.5 parameter has the strongest influence on the air quality index with a correlation value of 0.96.Implementation of the system into a website with a simple interface makes it easy for users to predict the Air Quality Index (AQI). The success of this implementation is expected to help monitor air quality effectively, reduce uncertainty in environmental decision making, and increase public awareness in maintaining air quality.

This research itself also still has shortcomings, so it is recommended that further research can improvise by using more variables in the data to predict the air quality index, this is because air pollution substances that affect the air quality index are quite varied and it is hoped that further research can use other machine learning methods to be able to compare the results of each method.

REFERENCES

[1] D. Kartikasari, “Analisis Faktor-Faktor Yang Mempengaruhi Level Polusi Udara Dengan Metode Regresi Logistik Biner,” *MATHunesa J. Ilm. Mat.*, vol. 8, no. 1, pp. 55–59, 2020, doi: 10.26740/mathunesa.v8n1.p55-59.

[2] A. Riyanto, A. Maheswara, R. Zulianty, V. M. Alegra, and ..., “Tanggung Jawab Pemerintah dalam Penyelesaian Masalah Polusi Udara di DKI Jakarta,” *J. Pendidik. Tambusai*, vol. 7, no. 3, pp. 27890–27896, 2023, [Online]. Available: https://www.jptam.org/index.php/jptam/article/view/11232%0Ahttps://www.jptam.org/index.php/jptam/article/download/11232/8850

[3] A. Nata and S. Suparmadi, “Analisis Sistem Pendukung Keputusan Dengan Model Klasifikasi Berbasis Machine Learning Dalam Penentuan Penerima Program Indonesia Pintar,” *J. Sci. Soc. Res.*, vol. 5, no. 3, p. 697, 2022, doi: 10.54314/jssr.v5i3.1041.

[4] R. E. Cahyono, J. P. Sugiono, and S. Tjandra, “Analisis Kinerja Metode Support Vector Regression (SVR) dalam Memprediksi Indeks Harga Konsumen,” *JTIM J. Teknol. Inf. dan Multimed.*, vol. 1, no. 2, pp. 106–116, 2019, doi: 10.35746/jtim.v1i2.22.

[5] H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” *Appl. Sci.*, vol. 9, no. 19, 2019, doi: 10.3390/app9194069.

[6] J. Wang, X. Li, L. Jin, J. Li, Q. Sun, and H. Wang, “An air quality index prediction model based on CNN-ILSTM,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-12355-6.

[7] E. Alpaydin, *Alpaydin, E., 2020. Introduction to machine learning. MIT press.* 2020.

[8] Mahesh, “Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), pp.381-386.,” *Int. J. Sci. Res.*, vol. 9(1), no. pp.381-386., 2020.

[9] N. Lavesson and P. Davidsson, “Quantifying the impact of learning algorithm parameter tuning,” *Proc. Natl. Conf. Artif. Intell.*, vol. 1, no. 1, pp. 395–400, 2006.

[10] S. Pakpahan, A. Faâ, and others, “Sistem Informasi Pengelolaan Dana Desa Pada Desa Hilizoliga Berbasis Web,” *J. Tek. Inform. UNIKA St. Thomas*, pp. 109–117, 2020.