A Geospatial Big Data Approaches to Estimate Granular Level Poverty Distribution in East Java, Indonesia using Machine Learning and Deep Learning Regressions

R Ramadhan[1], **A W Wijayanto**[1,2,*], **S Pramana** [1,2]

[1] Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia
[2] BPS-Statistics Indonesia, Jakarta, Indonesia

*Corresponding author's e-mail: ariewahyu@stis.ac.id

Abstract. One of the economic development the focus of the Indonesian government's efforts is for reducing poverty. In Indonesia, collecting poverty data uses the conventional method, the name is National Socio-Economic Survey (SUSENAS) which takes a large cost, time, and effort. To overcome these limitations, there is a need for additional data to provide more detailed poverty data. Recent studies show that the use of geospatial big data could identify poverty at a granular level, with a lower cost and faster update because of their unique and unbiased capacity to identify physical and socioeconomic phenomena. The integrated multi-source satellite imagery data such as the normalized difference vegetation index (NDVI) for detecting rural areas based on vegetation, built-up index (BUI) for identifying urban areas through building distribution, normalized difference water index (NDWI) for land cover detection, day time land surface temperature (LST) for identifying urban regions based on surface temperature, and pollutants such as carbon monoxide (CO), nitrogen dioxide (NO2), and sulfur dioxide (SO2) to evaluate economic activities based on pollution. Additionally, point of interest (POI) density and minimum POI distance are used to measure area accessibility. Therefore, the contribution of this research is to implement the utilization of geospatial big data to estimate the numbers of poverties at a granular level to the 666 sub-districts in East Java Province using machine learning and deep learning regression models. The evaluation results to estimate sub-district level poverty shows that the best model development using Support Vector Regression (SVR) in machine learning was the best root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) values of 0.365, 0.293, and 0.032 with R-squared of 0.59 and MLP in deep learning algorithm with 0.444, 0.345, and 0.039 values of RMSE, MAE, and MAPE with R^2 0.52. In addition, the results of visual identification revealed that high estimates of lower poverty are typically found in urban areas with high accessibility, and these areas are not spatially deprived areas with limited accessibility.

1. Introduction

In September 2015, 193 countries approved the United Nations Sustainable Development Goals (SDGs), which were first published with 17 objectives and 169 targets for "Transforming our World" [1], and poverty is one of the issues that this comprehensive framework is aiming to solve on a worldwide socioeconomic level. In Indonesia, the official poverty data is gathered by the conventional way of the National Socio-Economic Survey (SUSENAS) which involves performing on-the-ground household surveys every six months which has drawbacks in terms of scope, expense, labor, and time commitment

[2]. Indonesia's official poverty statistics are only permitted to be made public once a year at the district level, despite the need for more comprehensive data for poverty monitoring, and the ineffective delivery of social assistance is thought to be significantly impacted by this data shortage [3].

The most recent advances in Earth observation techniques and satellite image analytics have created numerous opportunities for precise and effective monitoring of Earth's surface geospatial features [28,29,32]. There is now a need for alternative poverty estimation data to supplement the official poverty statistics in order to hasten Indonesia's SDGs poverty target attainment. These alternative data should offer better coverage granularity, cheaper costs, and faster updates. Contrarily, geospatial big data and remote sensing satellite imagery, including point of interest (POI), show to be valuable resources because of their unique and unbiased capacity to identify physical and socioeconomic phenomena across various scales with effectiveness, cost savings, regular updates, and precise representation of the coverage are [4][5]. The advent of using satellite imagery from remote sensing data for poverty monitoring allows for the quick update of the current situation at a low cost or for free, but the quality of the data is dependent on cloud cover. Then, the other geospatial big data, such as Points of Interest (POI) from the OpenStreetMap (OSM) platform, which includes a sizable number of notable sites, can show the accessibility and economic density of a region [6][7].

The recent studies have shown that geospatial big data such as nighttime light (NTL) can show the population density [8], gross domestic product (GDP) [9][10], and electric power consumption [6] which could identify the socioeconomic conditions and poverty [30,31]. Besides that, using the normalized difference vegetation index (NDVI) was significantly high positive correlated with poverty [11], and land surface temperature (LST) could show the high and low income with urban thermal which could identify the regional poverty [12]. The normalized difference built-up index (NDBI) could have the potential to identify the urban areas [33-35], with the normalized difference water index (NDWI) which provide the accurate urban land to show the poverty area [13][14]. Therefore, the difference of geographical characteristic could show the regional poverty. Moreover, the air pollution such as the carbon monoxide (CO) and the nitrogen dioxide (NO_2) is related with the regional economic growth and GDP [15][16], and the sulfur dioxide (SO_2) could be used to identify the economic growth and energy consumption [17]. The other geospatial big data such as POI density and POI cost distance could show the regional economic development to identify the poverty also [6]. In addition, several studies have tried to implement of using the satellite imagery and POI for estimating poverty mapping with machine learning and deep learning. In Thailand, the using of the integration of geospatial big data such as nighttime light (NTL), some POI data, land cover with vegetation and urban index, and also the land surface temperature (LST) was achieving R^2 value greater than 0.8 with Random Forest algorithm [18]. In the Southwest China to predict the contiguous extremely poor area, the best algorithm with the Extreme Gradient Boosting (XGBoost) model had the highest R^2 of 0.61 using multisource spatial data such as NTL, POI, land cover, and the digital elevation model (DEM) [19]. The using of deep learning by transfer learning have shown the strong predictive of both average household consumption expenditure in 37% to 55% of variation and asset wealth in 55% to 75% either in African countries [20].

In Indonesia, the utilizing of geospatial big data such as population density data with economic spatial distribution from NTL, the geographical characteristic with land cover (NDVI, NDWI, and NDBI), the quality of environment by the air pollution (NO, CO_2, and SO_2), and geo-accessibility with POI for poverty mapping is still limited in granular level. The accessibility of this data can attest to the official poverty statistics' inadequacies. In this study, machine learning and deep learning techniques are used to estimate poverty up to the sub-district level utilizing geographic big data, such as satellite images from multi-source remote sensing and POI using OSM platform. This studied is focused in East Java which was the province who have the highest poverty rate in 4.23 million of poor individuals in Indonesia in 2022 [21]. However, this research is proposed the map of poverty that can be updated more quickly and affordably, supplementing the official poverty data now in use. In the end, the goal is to make policy decisions that are more effective and efficient and align with the first Sustainable Development Goal with zero poverty.

2. Method

2.1. Study Area

The East Java is the province in Indonesia which has 38 regencies/municipalities with 666 sub-districts and has been selected in this research due to the highest number of poverty in Indonesia in 2022 in 4.23 million individuals [21]. Based on BPS-Statistics Indonesia in 2022, the number of urban poverties in East Java in September 2022 increased by 24.18 thousand people (from 1.721 million in March 2022 to 1.752 million in September 2022). Meanwhile, in the same period, the number of rural poverties increased by 24.2 thousand people (from 2.459 million in March 2022 to 2.484 million in September 2022) [22].

Poverty is concentrated in the northeastern region of East Java, most notably on the island of Madura especially in Sampang Regency with 217,970 people and Sumenep Regency with the poverty 206,200 people in 2022 [23]. Besides that, the poverty in rural area such as Kota Surabaya (the second-largest metropolitan in Indonesia) has 138,210 people and Kota Malang has 38,560 people [23]. Thus, the researcher took the locus of study to East Java Province.

2.2. Data Used

In this research, point-of-interest (POI and multi-source satellite images were used to construct a model for estimating poverty. Detailed information about datasets, can be found in Table 1.

Table 1. Data summary

Data Source	Variable	Band	Unit	Spatial Resolution
Visible Infrared Imaging Radiometer Suite (VIIRS)	Nighttime Light Intensity (NTL)	avg_radian	nanoWatts/ cm2/sr	750 m
Sentinel Multispectral Level 2A	Normalized Difference Vegetation Index (NDVI)	B4 (Red) dan B8 (NIR)	index	10 m
	Normalized Difference Water Index (NDWI)	B3 (Green) dan B8 (NIR)		
	Normalized Difference Built-Up Index (NDBI)	B8 (NIR) dan B11 (SWIR 1)		
Moderate-resolution Imaging Spectroradiometer (MODIS)	Day Time Land Surface Temperature (LST)	LST Day 1 km	Kelvin	1000 m
Sentinel-5P	Carbon Monoxide (CO)	CO Column Number Density	mol/m2	1113.2 m
	Nitrogen Dioxide (NO2)	NO2 Column Number Density		
	Sulfur Dioxide (SO2)	SO2 Column Number Density		
OpenStreetMap	POI Density	-	point	dynamic

Data Source	Variable	Band	Unit	Spatial Resolution
(OSM)	POI Distance	-	meter	dynamic
BPS-Statistics Indonesia	Total Population in sub-district	-	population	-
BPS-Statistics Indonesia	The numbers of poverties in regency/municipality	-	population	-

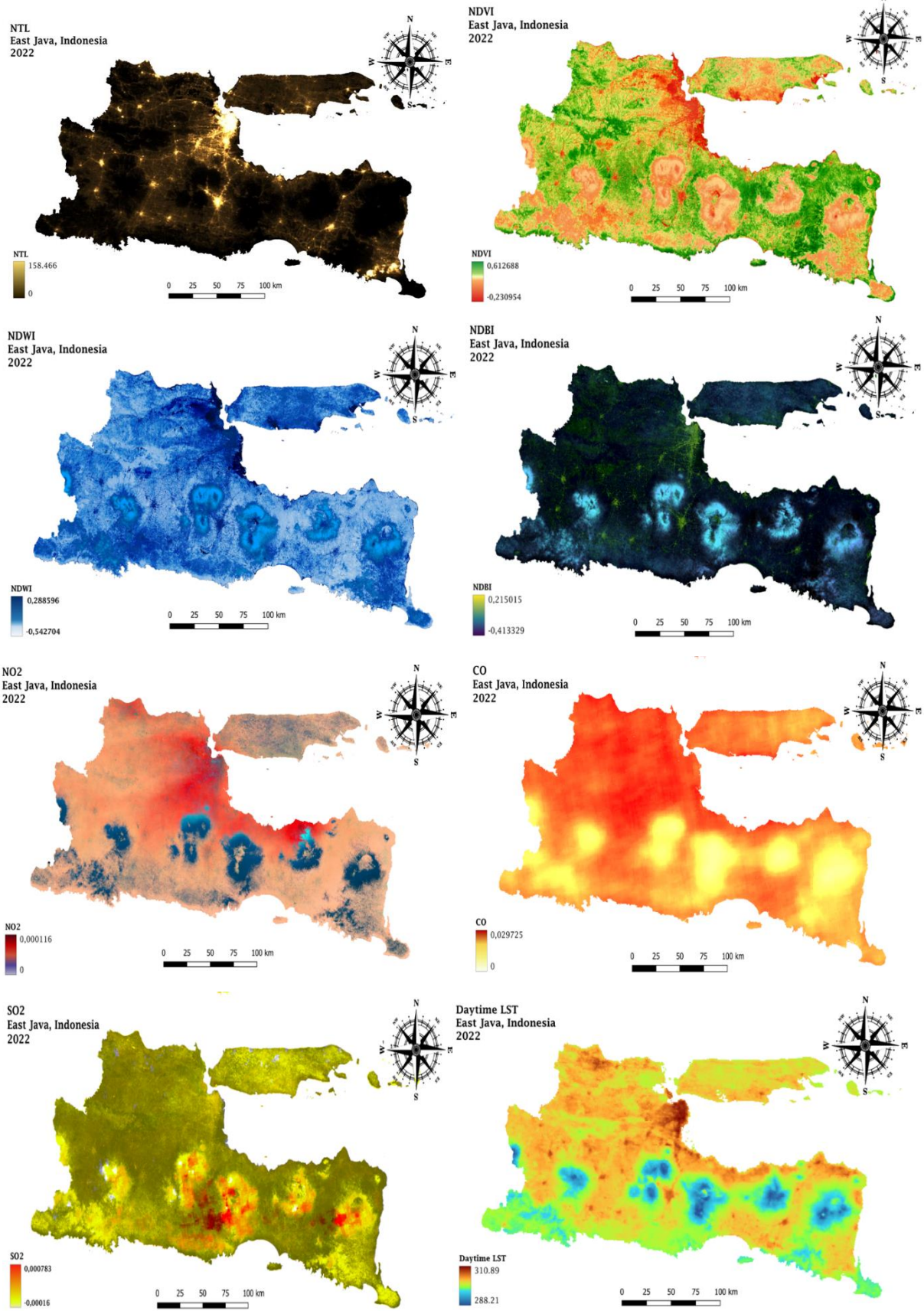
This study uses data sourced from remote sensing, namely multi-source satellite imagery and other geospatial big data, namely Point of Interest (POI) data originating from OpenStreetMap (OSM). The multi-source satellite images used in this study are Nighttime Light (NTL) intensity from NOAA-VIIRS, Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Built-Up Index (NDBI) from Sentinel Multispectral Level 2A. NDVI is a vegetation index that is analyzed through reflection brightness and absorption of Near-Infrared (NIR) and red band and positively correlated with poverty [11]. Then, NDBI could have the potential to identify the urban areas [13], and NDWI which provide the accurate urban land and water area [14], that both can identify the poverty areas. These are the compositing index that used in this study, they are NDVI, NDBI, and NDWI is s calculated based on the following formula:

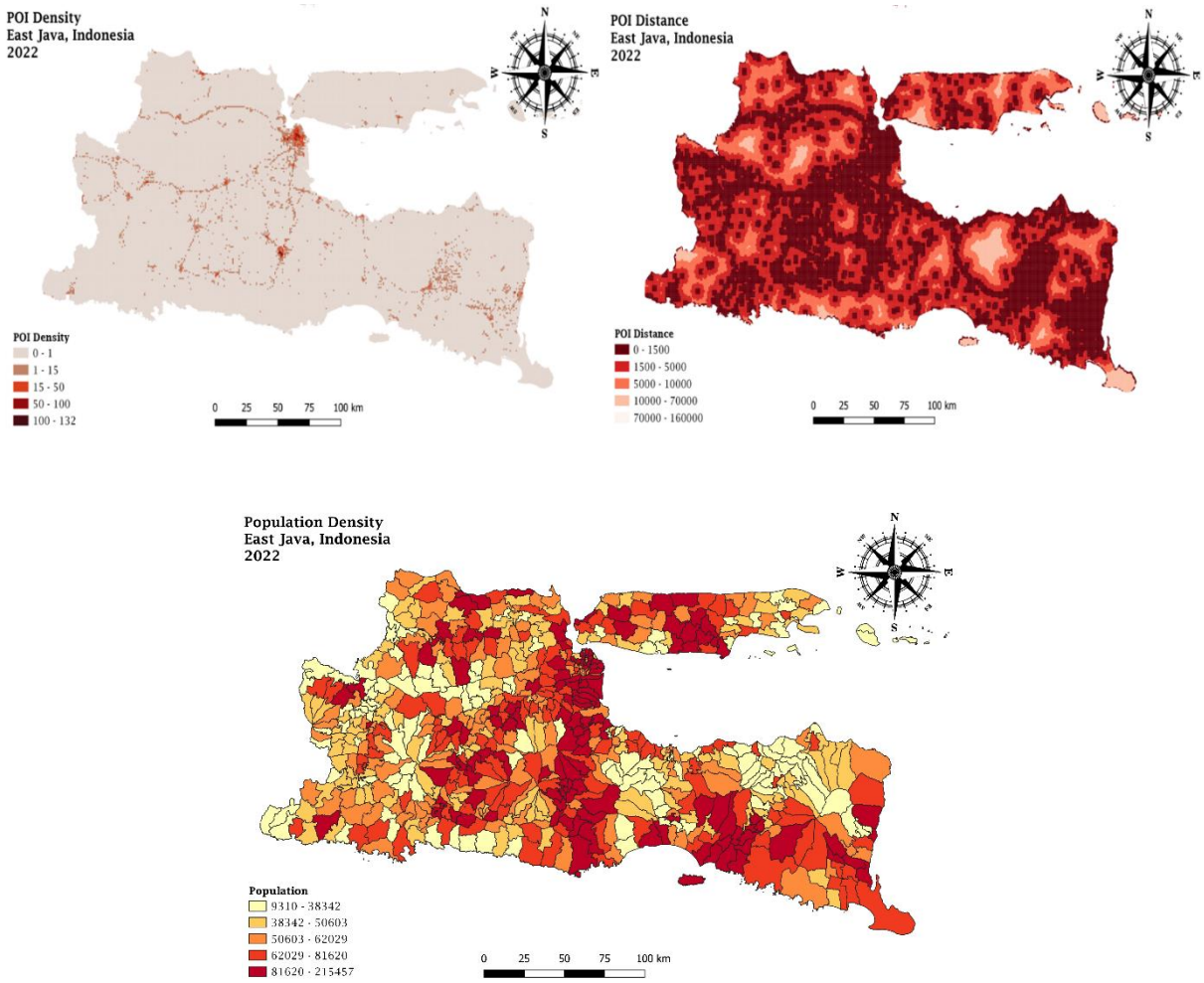
$$NDVI = \frac{NIR_{band\ 8} - RED_{band\ 4}}{NIR_{band\ 8} + RED_{band\ 4}} \quad (1)$$

$$NDWI = \frac{Green_{band\ 3} - NIR_{band\ 8}}{Green_{band\ 3} + NIR_{band\ 8}} \quad (2)$$

$$NDBI = \frac{SWIR_{band\ 1} - NIR_{band\ 8}}{SWIR_{band\ 1} + NIR_{band\ 8}} \quad (3)$$

Besides that, this study used the other variables such as Daytime Land Surface Temperature (LST) from MODIS, Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), and Sulfur Dioxide (SO_2) from Sentinel-5P. Remote sensing satellite imagery data is collected and processed through a cloud-based platform designed to store and process Earth's geographic data, namely Google Earth Engine (GEE). Meanwhile, POI data is another geospatial big data set used in this study and collected through OSM. POI data describes the accessibility of an area; in this study, there were 17,542 points, which are places of public access in East Java Province, and they were filtered through several categories such as education, health, economy, tourist attractions, and so on. The official statistics data used in this study are population data for each regency/municipality, and sub-district in East Java Province in 2022, which is from the BPS-Statistics Indonesia. The data visualization in this study has shown in Figures 1.

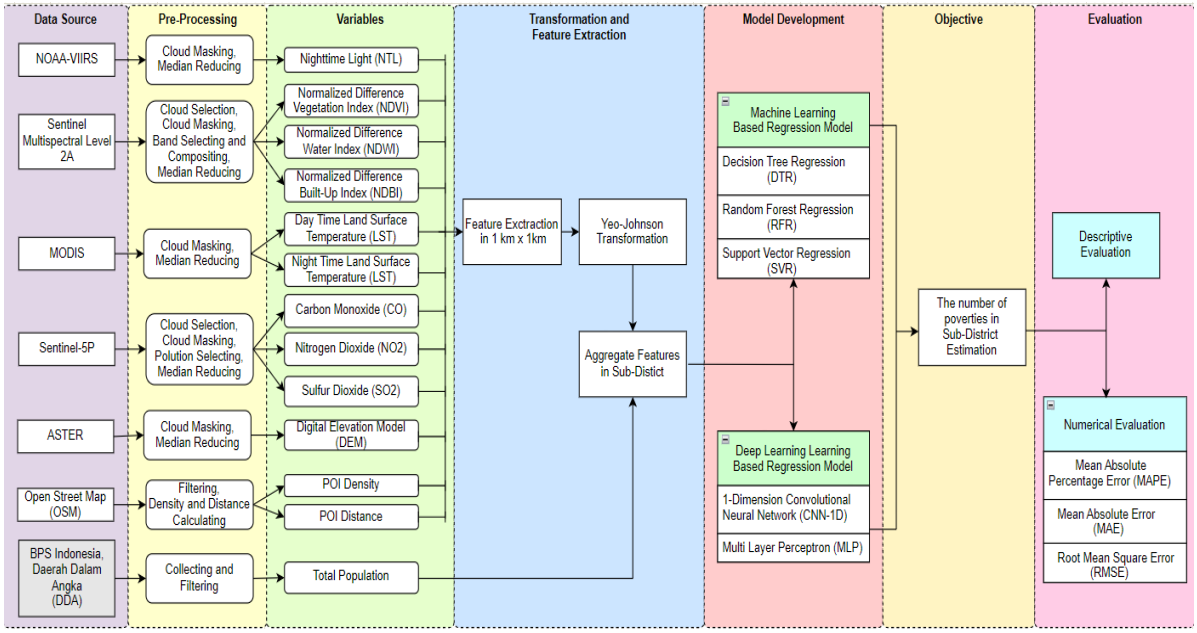

NTL
East Java, Indonesia
2022
NTL
158.466
0
0 25 50 75 100 km
NDVI
East Java, Indonesia
2022
NDVI
0,612688
-0,230954
0 25 50 75 100 km
NDWI
East Java, Indonesia
2022
NDWI
0,288596
-0,542704
0 25 50 75 100 km
NDBI
East Java, Indonesia
2022
NDBI
0,215015
-0,413329
0 25 50 75 100 km
NO2
East Java, Indonesia
2022
NO2
0,000116
0
0 25 50 75 100 km
CO
East Java, Indonesia
2022
CO
0,029725
0
0 25 50 75 100 km
SO2
East Java, Indonesia
2022
SO2
0,000783
-0,00016
0 25 50 75 100 km
Daytime LST
East Java, Indonesia
2022
Daytime LST
310.89
288.21
0 25 50 75 100 km



Figures 1. Visualization of NTL (nanoWatts/cm^2/sr), NDVI (index), NDWI (index), NDBI (index), Daytime LST dan Nighttime LST (Kelvin), NO_2 (mol/m^2), CO (mol/m^2), SO_2 (mol/m^2), POI density (points), POI distance (meter), dan Population Density (people) in East Java, Indonesia 2022.

2.3. Research Framework

This research focuses on the utilization of multi-source satellite imagery big data and POI data originating from OpenStreetMap (OSM) to build estimates of poverty mapping at the sub-district levels using machine learning and deep learning algorithm. The research solves the problem that underlies the reason for conducting the research, namely the weakness of collecting data by conventional survey method to calculate poverty, which then proposes solutions to achieve the goals until the evaluation of target achievement. Figure 2 shows the proposed research framework.



Figures 2. The research framework

In this study, the feature extraction is carried out by applying statistics based on a zone of 1 km x 1 km and transformed using the Yeo-Johnson method. This transformation is used because the variables used have different units and different values, positive and negative and this transformation method is effective for this problem and transformed it to more normal distribution [24]. The extraction data that had been transformed are also used to build models using machine learning and deep learning to estimate the poverty down to the sub-district in East Java. The machine learning methods used are Decision Tree Regression (DTR), Random Forest Regresion (RFR), and Support Vector Regressions (SVR) that have their respective modelling. Then the deep learning algorithms used are 1-Dimension Convolutional Neural Network (CNN-1D) and Multi Layer Perceptron (MLP) which have their respectively modelling as well. The results obtained are then evaluated descriptively and numerically through the calculation of the correlation coefficient, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE). The measurement is used because the modeling used is regression-based and requires numerical evaluation measurements [25]. The formula is calculated in (6), (7), and (8) forms. To facilitate interpretation and stakeholders' participation in decision-making, a visualization mapping the results of the poverty estimation to the sub-district levels in East Java was carried out.

2.4. Broad Area Ratio Estimation (BARE) for Poverty Preprocessing

More granular poverty information, namely at the sub-district level, is needed for the purpose of estimating poverty because the poverty data published by BPS is only limited to the regency/municipality level. Therefore, one of the basic methods of small area estimation (SAE), namely the Broad Area Ratio Estimation (BARE), is used so that a more representative target variable value is obtained for the sub-district level. Supporting data used to apply this method are the total population of regency/municipality and sub-districts from the BPS for each district or city, as well as poverty data for each regency/municipality in East Java. The BARE method uses direct estimates of the variables of interest for a given area, for which data can be obtained from small-area population surveys or population censuses and other demographic estimation data [26]. The small area of interest is the sub-district, and the available data at the regency/municipality (broad area) level are direct estimates from surveys and population data for the small area. The amount of poor people in a sub-district can be calculated by applying the regency/municipality poverty incidence to the sub-district level population [26]. The major assumption is that small areas have the same characteristics as large areas so that unbiased estimates can be made. Additionally, making reliable small area estimates requires calculation

of the direct estimate for the large area from a survey with a sufficiently large sample size. Additionally, BAREs can serve as references to confirm the results of small area estimates from more complex methods [26].They are applied based on the following formula (4).

$$\hat{Y}_i = \bar{Y}_p \times N_i \tag{4}$$

BARE is a basic small area estimation (SAE) technique that assumes the $\bar{Y}_p$ is a direct estimator that is calculated through $\frac{\hat{Y}_p}{N_p}$, namely the ratio of the poor population of a regency/municipality p to its population and N_i is the population of a sub-district in regency/municipality p. The main assumption is that the small area has the same characteristics as the large area above it which will produce an unbiased estimation. In addition, producing reliable small area estimation requires direct estimator calculations with a large enough sample size so that the BARE method can be used as a reference in confirming small area estimation results from more complicated methods.

2.5. Data Transformation

The Yeo-Johnson power transformation, a variation of the Box-Cox transformation, is used in this study to handle data values that can be both positive and negative [27]. This transformation method is effective for addressing variables with dissimilar units across all ranges by reshaping them to conform to a Gaussian distribution or a more normal distribution [24]. The specific data transformation approach used in this study is as follows:

$$y_\lambda(x) \begin{cases} \frac{(1+x)^\lambda - 1}{\lambda}, \lambda \neq 0 \; dan \; x \geq 0 \\ log(1+x), \lambda = 0 \; dan \; x \geq 0 \\ -\frac{(1-x)^{2-\lambda}}{2-\lambda}, \lambda \neq 2 \; dan \; x < 0 \\ -log(1-x), \lambda = 2 \; dan \; x < 0 \end{cases} \tag{5}$$

For each individual variable or input data value, denoted as x, the Yeo-Johnson transformation is applied using a parameter λ, which is estimated through the Maximum Likelihood technique, if the variables adhere to a normal distribution. When the family of transformations employs a parameter value of λ=1, a linear relationship is achieved. The transformation then adjusts the distribution by either thickening or condensing the right tail when λ<1, thereby making the distribution right-skewed and closer to symmetry. Conversely, when λ>1, it makes the left tail more symmetrical, particularly in cases of left-skewed distributions.

2.6 Model Development and Evaluation

The results of feature extraction that has been carried out at each level of the 1 km x 1 km grid are ten variables of satellite imagery data and other geospatial data, namely POI, which are transformed using the Yeo-Johnson method. Then, aggregation is carried out to the sub-district level, and then modelling is carried out with an additional variable, namely the population at the sub-district level. The results of this aggregation will be the independent variable in each model used in predicting poverty at the sub-district level as the dependent variable.

The development of the estimation model used is based on machine learning and deep learning. In the machine learning approach, the algorithms used are Decision Tree Regression (DTR), Random Forest Regression (RFR), and Support Vector Regression (SVR).

The selection of the best parameters and hyperparameters for each model is carried out by conducting a grid search with 10-fold cross-validation. Then, in the deep learning approach, algorithms are used, namely Multi-Layer Perceptron (MLP) and Convolutional Neural Networks (CNN-1D). Parameter selection was carried out in a random experiment with a 10-fold cross-validation evaluation so that two models of poverty estimation results were obtained. In this study, a performance comparison was made between machine learning and deep learning algorithms in predicting the dependent variable, and

numerical evaluation measures were carried out, namely RMSE, MAE, and MAPE, which had the smallest average at the sub-district level with 10-fold cross-validation test data. RMSE, MAE, and MAPE values are calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (7)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| x\ 100\% \quad (8)$$

Where y_i denotes the true value (in this research is the numbers of poverty using BARE), $\hat{y}_i$ denotes the predicted value using machine learning and deep learning algorithm, and n denotes the number of observations. Numerical evaluations were performed at the sub-district level using SAE poverty data using official poverty statistics because the limited of official data in sub-district level.

3. Result

The performance of the models is evaluated using a 10-fold cross-validation approach. This means that the dataset is divided into 10 subsets, and each model is trained and tested on different combinations of these subsets. The evaluation results of the 10-fold cross-validation on the test data for each fold are presented in Table 2. To determine which poverty mapping result is better, the numerical evaluation should preferably be done at the sub-district level. However, due to the limitations of poverty data that is only available at the district level and the results of the SAE (small area estimation) estimation carried out only up to the sub-district level, numerical evaluations will only be carried out at the sub-district. Table 2 shows the evaluation results of 10-fold cross validation model development with machine learning (DTR, RFR, and SVR) and deep learning (MLP and CNN-1D) at the sub-district levels. Table 3 shows that the aggregated poverty map estimated evaluation in sub-district level from 10-fold cross validation, it shows the MLP model (based on multi-source satellite imagery and POI) has the lowest RMSE, MAE, and MAPE for deep learning algorithm and SVR is for machine learning. This shows that this poverty estimation is closer to poverty at the sub-district level. As a result, the MLP and SVR poverty map based on multi-source satellite imagery and POI was chosen as the best poverty map in this study.

Table 2. Evaluation results of 10-fold cross validation model development

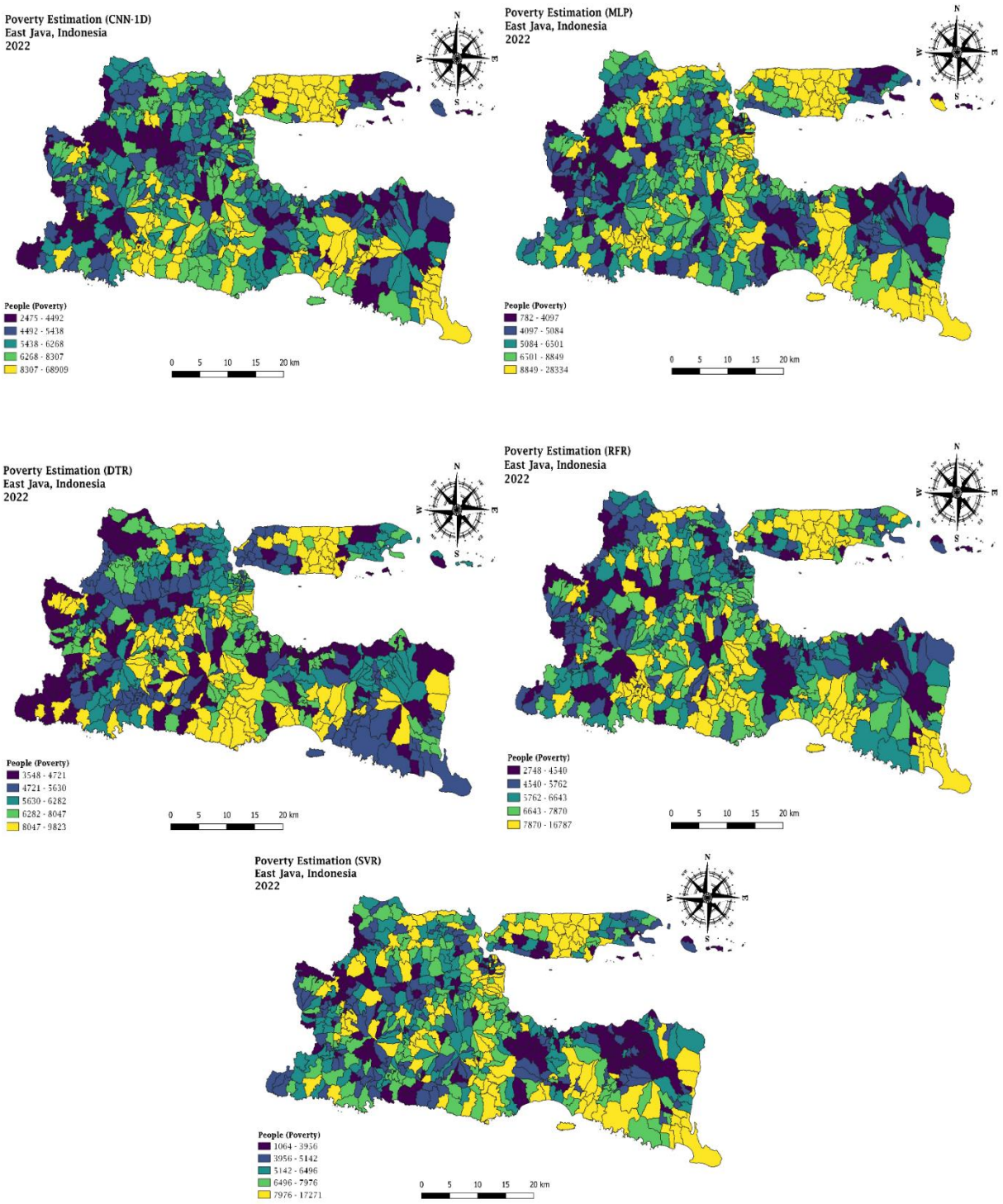
Fold	Algorithm	RMSE	MAE	MAPE
1	DTR	1.1335	0.8771	8.9029%
	RFR	**1.1099**	**0.8769**	**8.9554%**
	SVR	1.1519	0.9033	9.2404%
	MLP	**1.1702**	**0.9017**	**9.1981%**
	CNN-1D	1.1436	0.9447	9.9651%
2	DTR	0.5705	0.4292	5.1423%
	RFR	0.3593	0.2491	2.4908%
	SVR	**0.3279**	**0.2124**	**2.4285%**
	MLP	**0.2952**	**0.1896**	**2.1812%**
	CNN-1D	0.4439	0.3393	4.0083%
3	DTR	0.3521	0.2851	3.2616%
	RFR	0.2911	0.2369	2.7608%
	SVR	**0.2115**	**0.1664**	**1.9269%**
	MLP	**0.2870**	**0.2285**	**2.6317%**
	CNN-1D	0.6977	0.5393	6.1478%
4	DTR	0.3883	0.3002	3.6389%

Fold	Algorithm	RMSE	MAE	MAPE
	RFR	0.3708	0.2642	3.2342%
	SVR	**0.1971**	**0.1292**	**1.5631%**
	MLP	**0.3536**	**0.2638**	**3.1981%**
	CNN-1D	0.4172	0.3129	3.7516%
5	DTR	0.3645	0.3162	3.7951%
	RFR	0.3649	0.3117	3.7885%
	SVR	**0.2484**	**0.2177**	**2.6315%**
	MLP	**0.1744**	**0.1457**	**1.7231%**
	CNN-1D	0.4189	0.3270	3.9771%
6	DTR	0.2932	0.2282	2.6524%
	RFR	0.2675	0.2019	2.3441%
	SVR	**0.1571**	**0.1257**	**1.4598%**
	MLP	**0.2062**	**0.1638**	**1.8923%**
	CNN-1D	0.2641	0.1964	2.3025%
7	DTR	0.2856	0.2363	2.7303%
	RFR	0.2501	0.1964	2.2635%
	SVR	**0.2105**	**0.1643**	**1.8653%**
	MLP	0.2783	0.2343	2.6653%
	CNN-1D	**0.3021**	**0.2292**	**2.5886%**
8	DTR	0.6114	0.5068	5.7569%
	RFR	**0.4786**	**0.4275**	**4.8533%**
	SVR	0.5278	0.4999	5.6361%
	MLP	**0.3961**	**0.3177**	**3.6408%**
	CNN-1D	0.4379	0.3389	3.8817%
9	DTR	0.5391	0.4309	5.1478%
	RFR	0.3584	0.3025	3.5677%
	SVR	**0.3122**	**0.2557**	**2.9573%**
	MLP	**0.5789**	**0.4544**	**5.2173%**
	CNN-1D	0.5823	0.4679	5.3658%
10	DTR	0.4966	0.4015	4.7794%
	RFR	0.6062	0.4767	5.6661%
	SVR	**0.3061**	**0.2528**	**2.9998%**
	MLP	0.7047	0.5547	6.6203%
	CNN-1D	**0.6828**	**0.5047**	**6.0216%**

Table 3. The aggregated evaluation results of 10-fold cross validation model development

Model	The Average Test of 10-*fold Cross Validation*		
	RMSE	MAE	MAPE
DTR	0.50348	0.40115	4.5808%
RFR	0.44568	0.35438	3.9924%
SVR	**0.36505**	**0.29274**	**3.2709%**
MLP	**0.44446**	**0.34542**	**3.8968%**
CNN-1D	0.53905	0.42003	4.8010%

Based on table 3, it can be seen that the machine learning algorithm which has the best average RMSE, MAE, and MAPE is SVR in estimating poverty (natural logarithm transformation) at the sub-district level with a RMSE value of 0.36505, MAE of 0.29274, and MAPE worth 3.2709%. Then, in the deep learning algorithm, the best algorithm is MLP with a RMSE value of 0.44446, MAE of 0.34542, and MAPE of 3.8968%. The poverty estimation results of each deep learning model and machine learning are mapped to make it easier to visualize and interpret as shown in figure 3.



Figures 3. Poverty Estimation Mapping with Machine Learning and Deep Learning Modeling

The figure 3 hows the resulting the number of poverties poverty map visual identification of modelling estimation. The figure 3 shows that sparsely populated areas surrounded by agricultural areas, or rural areas, have higher estimated poverty values, such as in one of the regions in Sampang, Tuban, Jember, and Magetan. Conversely, densely populated areas with better access, or urban areas, have lower poverty estimates, such as one of the region in Surabaya and Kota Malang. This is aligned with Statistics Indonesia's report, which claims that poverty in rural regions is higher than in urban areas

(Statistics Indonesia (BPS), 2020). Figure 4 shows a scatter plot comparing the predicted value with the actual value of poverty at the district and sub-district level. It is shown that the result of the constructed poverty mapping is strongly correlated (Sugiyono, 2010) with poverty at the sub-district with correlation value is 0.768 Pearson correlation coefficient for SVR poverty estimation and 0.7237 Pearson correlation coefficient for MLP poverty estimation. The adjusted R^2 value represents that the independent variables in the model can explain 59% of the variance in poverty at the sub-district level with SVR poverty estimation and 52.38% of the of the variance in poverty at the sub-district level with MLP poverty estimation

Figures 4. The Scatter Plot of SVR and MLP Estimation with SAE

4. Conclusion

This research proposes a novel method intended to produce a grid-level poverty map with a geographical resolution of sub-districts. To fulfil the first Sustainable Development Goal (SDG), which is to eradicate poverty, the major goal is to improve poverty monitoring. The advantage of this method for updating the poverty map is that it saves money and time. Support Vector Regression (SVR) model was found to be the best machine learning model for the first scenario through the assessment of model development, while Multilayer Perceptron (MLP) model was chosen as the best deep learning model for the second scenario. The first scenario's poverty map, which was produced using the SVR model and official

poverty data, had the highest accuracy, with an RMSE value of 0.36505, a MAE value of 0.29274, and a MAPE value of 3.2709%. The Pearson correlation and adjusted R^2 were both 0.768 at the sub-district level. The northern and northeastern parts of the province of East Java, as well as Madura Island, all showed high levels of poverty. According to the visual study, locations with high poverty estimates were also generally sparsely populated and bordered by uninhabited terrain, which frequently represented agricultural areas. On the other hand, areas with low estimates of poverty tended to be populous and accessible. These results support a study from Statistics Indonesia that claims rural regions have greater rates of poverty than metropolitan areas.

Without the conventional data, big geospatial data can be used as a proxy or alternative data because of its advantage that can estimate on more granular areas for the provision of data poverty so as not as a substitute for existing data. Then, the use of data and methods in this study can be an early indicator to look at areas with good or bad economic sides so that it can give an indication of such poverty. The data used can not provide data namely by address in detail but can provide supporting data related to social protection data by identifying areas with high poverty estimates or deprived areas. Later, this modeling can also be used as a monitoring for poverty estimates by using the best existing data modeling using the same new data source or by training the data again and adding or combining it with some other official statistics data.

References

[1] M. A. B. Omer and T. Noguchi, "A conceptual framework for understanding the contribution of building materials in the achievement of Sustainable Development Goals (SDGs)," *Sustain. Cities Soc.*, vol. 52, p. 101869, 2020, doi: 10.1016/j.scs.2019.101869.

[2] M. Jerven, "Benefits and costs of the data for development targets for the Post-2015 Development Agenda," *Data Dev. Assess. Pap. Work. Pap.*, vol. Copenhagen, no. September 2014, p. 41, 2014, [Online]. Available: http://www.copenhagenconsensus.com/sites/default/files/data_assessment_-_jerven.pdf.

[3] U. N. Afifah and R. Faradis, "Optimalisasi Data Survei Sosial Dan Ekonomi Nasional (Susenas) Dengan Small Area Estimation (Sae)," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, pp. 132–139, 2020, doi: 10.34123/semnasoffstat.v2019i1.147.

[4] A. W. Wijayanto, D. W. Triscowati, and A. H. Marsuhandi, "Maize field area detection in East Java, Indonesia: An integrated multispectral remote sensing and machine learning approach," *ICITEE 2020 - Proc. 12th Int. Conf. Inf. Technol. Electr. Eng.*, pp. 168–173, 2020, doi: 10.1109/ICITEE49829.2020.9271683.

[5] N. Pokhriyal, O. Zambrano, J. Linares, and H. Hernández, "Estimating and Forecasting Income Poverty and Inequality in Haiti Using Satellite Imagery and Mobile Phone Data," *Estim. Forecast. Income Poverty Inequal. Haiti Using Satell. Imag. Mob. Phone Data*, 2020, doi: 10.18235/0002466.

[6] K. Shi, Z. Chang, Z. Chen, J. Wu, and B. Yu, "Identifying and evaluating poverty using multisource remote sensing and point of interest (POI) data: A case study of Chongqing, China," *J. Clean. Prod.*, vol. 255, p. 120245, 2020, doi: 10.1016/j.jclepro.2020.120245.

[7] I. Tingzon *et al.*, "Mapping Poverty in the Philippines Using Machine Learning, Satellite Imagery, and Crowd-Sourced Geospatial Information," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 42, no. 4/W19, pp. 425–431, 2019, doi: 10.5194/isprs-archives-XLII-4-W19-425-2019.

[8] K. Shi *et al.*, "Modeling and mapping total freight traffic in China using NPP-VIIRS nighttime light composite data," *GIScience Remote Sens.*, vol. 52, no. 3, pp. 274–289, 2015, doi: 10.1080/15481603.2015.1022420.

[9] Y. Gu, Z. Shao, X. Huang, and B. Cai, "GDP Forecasting Model for China's Provinces Using Nighttime Light Remote Sensing Data," *Remote Sens.*, vol. 14, no. 15, 2022, doi: 10.3390/rs14153671.

[10] Z. Zhao *et al.*, "Analysis of the Spatial and Temporal Evolution of the GDP in Henan Province

Based on Nighttime Light Data," *Remote Sens.*, vol. 15, no. 3, 2023, doi: 10.3390/rs15030716.
[11] T. Dawson, J. S. Onésimo Sandoval, V. Sagan, and T. Crawford, "A spatial analysis of the relationship between vegetation and poverty," *ISPRS Int. J. Geo-Information*, vol. 7, no. 3, 2018, doi: 10.3390/ijgi7030083.
[12] G. Huang, W. Zhou, and M. L. Cadenasso, "Is everyone hot in the city? Spatial pattern of land surface temperatures, land cover and neighborhood socioeconomic characteristics in Baltimore, MD," *J. Environ. Manage.*, vol. 92, no. 7, pp. 1753–1759, 2011, doi: 10.1016/j.jenvman.2011.02.006.
[13] S. Ahmed, "Assessment of urban heat islands and impact of climate change on socioeconomic over Suez Governorate using remote sensing and GIS techniques," *Egypt. J. Remote Sens. Sp. Sci.*, vol. 21, no. 1, pp. 15–25, 2018, doi: 10.1016/j.ejrs.2017.08.001.
[14] Y. Zheng, Q. Zhou, Y. He, C. Wang, X. Wang, and H. Wang, "An optimized approach for extracting urban land based on log-transformed dmsp-ols nighttime light, ndvi, and ndwi," *Remote Sens.*, vol. 13, no. 4, pp. 1–22, 2021, doi: 10.3390/rs13040766.
[15] Y. Wang *et al.*, "The impact of carbon monoxide on years of life lost and modified effect by individual- and city-level characteristics: Evidence from a nationwide time-series study in China," *Ecotoxicol. Environ. Saf.*, vol. 210, p. 111884, 2021, doi: 10.1016/j.ecoenv.2020.111884.
[16] C. Han, Z. Gu, and H. Yang, "Ekc test of the relationship between nitrogen dioxide pollution and economic growth—a spatial econometric analysis based on chinese city data," *Int. J. Environ. Res. Public Health*, vol. 18, no. 18, 2021, doi: 10.3390/ijerph18189697.
[17] K. Bakhsh, T. Akmal, T. Ahmad, and Q. Abbas, "Investigating the nexus among sulfur dioxide emission, energy consumption, and economic growth: empirical evidence from Pakistan," *Environ. Sci. Pollut. Res.*, vol. 29, no. 5, pp. 7214–7224, 2022, doi: 10.1007/s11356-021-15898-9.
[18] N. Puttanapong, A. Martinez, J. A. N. Bulan, M. Addawe, R. L. Durante, and M. Martillan, "Predicting Poverty Using Geospatial Data in Thailand," *ISPRS Int. J. Geo-Information*, vol. 11, no. 5, 2022, doi: 10.3390/ijgi11050293.
[19] Y. Xu, Y. Mo, and S. Zhu, "Poverty mapping in the dian-gui-qian contiguous extremely poor area of southwest china based on multi-source geospatial data," *Sustain.*, vol. 13, no. 16, 2021, doi: 10.3390/su13168717.
[20] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science (80-.).*, vol. 353, no. 6301, pp. 790–794, 2016, doi: 10.1126/science.aaf7894.
[21] Badan Pusat Statistik Indonesia, "Profil kemiskinan di indonesia september 2023," *Ber. Resmi Stat.*, vol. 01, no. 05, pp. 1–16, 2023.
[22] BPS Jawa Timur, "BRS Profil Kemiskinan jawa Timur," no. 06, 2022.
[23] Badan Pusat Statistik RI, *Data dan Informasi Kemiskinan Kabupaten/Kota di Indonesia*. Jakarta, Indonesia.
[24] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," *Mach. Learn.*, no. May 2020, 2021, doi: 10.1007/s10994-021-05960-5.
[25] A. Botchkarev, "Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3177507.
[26] Asian Development Bank, "Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices," no. May, 2020, p. 99.
[27] Y. I.-K. and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 156, no. I, pp. 87–90, 1989.
[28] Afira N, Wijayanto AW 2022 Mono-temporal and multi-temporal approaches for burnt area detection using Sentinel-2 satellite imagery (a case study of Rokan Hilir Regency, Indonesia), Ecological Informatics, 69, 101677, Elsevier

[29] Wijayanto AW, Afira N, Nurkarim W 2022 Machine Learning Approaches using Satellite Data for Oil Palm Area Detection in Pekanbaru City, Riau, Proceedings of the 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom).
[30] Putri, SR, Wijayanto, AW, & Permana, S. (2023). Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches. Remote Sensing Applications: Society and Environment, 29, 100889.
[31] Putri, SR, Wijayanto, AW, & Sakti, AD (2022). Developing relative spatial poverty index using integrated remote sensing and geospatial big data approach: a case study of East Java, Indonesia. ISPRS International Journal of Geo-Information, 11(5), 275.
[32] Putri SR, Wijayanto AW 2022 Learning Bayesian Network for Rainfall Prediction Modeling in Urban Area using Remote Sensing Satellite Data (Case Study: Jakarta, Indonesia), Proceedings of The International Conference on Data Science and Official Statistics, 2021, 1, 77-90
[33] Saadi T D T and Wijayanto A W 2021 Machine learning applied to Sentinel-2 and Landsat-8 multispectral and medium-resolution satellite imagery for the detection of rice production area in Nganjuk, East Java, Indonesia International Journal of Remote Sensing and Earth Sciences 18 19-32
[34] Wijayanto AW, Triscowati DW, Marsuhandi AH 2020 Maize Field Area Detection in East Java, Indonesia: An Integrated Multispectral Remote Sensing and Machine Learning Approach. 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE).
[35] Nurmasari Y, Wijayanto AW 2021 Oil Palm Plantation Detection in Indonesia using Sentinel-2 and Landsat-8 Optical Satellite Imagery (Case Study: Rokan Hulu Regency, Riau Province), International Journal of Remote Sensing and Earth Sciences (IJReSES), 18, 1, 1-18, LAPAN