Classification of Urban and Rural Villages with Machine Learning on Satellite Image Data and Points of Interest

A S R Zega[1], B P Josaphat[1*]

[1] Department of Statistical Computing, Politeknik Statistik STIS, Jakarta, Indonesia

*Corresponding author's email: bonyp@stis.ac.id

Abstract. An evaluation of the Sustainable Development Goals with data disaggregated by residential area, namely urban and rural areas, is essential. This study proposes the use of satellite imagery and point of interest (POI) data with machine learning methods to classify urban and rural villages, specifically in North Sumatra Province. The data used includes satellite imagery from various sources, such as NOAA-20, Sentinel-2, Sentinel-5P, and Terra, as well as Google Maps, covering various variables including NTL, NDVI, NDBI, NDWI, NO_2, CO, and LST, along with POIs categorized under education, economy, health, and entertainment. The machine learning methods used were Decision Tree and Support Vector Machine, with data imbalance addressed through resampling techniques such as Random Under sampling (RUS). The results of the study show that the Support Vector Machine model with RUS produced the best weighted average F1-score of 87.74% for the classification of urban and rural villages, with NTL being the most important feature in the model formation. This study is expected to be an alternative for BPS in the classification of urban and rural villages.

Keyword: Machine Learning, POI, Remote Sensing, Rural, Urban.

1. Introduction

One important part of SDG metadata is data disaggregation. Data disaggregation/separation/grouping aims to reveal differences that are hidden in aggregate data by breaking down the data into appropriate groups[1]. One form of data disaggregation in Indonesia's SDG achievement indicators is disaggregation by place of residence, specifically disaggregation based on urban and rural areas. Disaggregation by place of residence is necessary to support more in-depth analysis of the data, particularly in identifying various aspects such as access to basic services[2][3][4], economicdisparities[5], environmental quality[6], legal issues[7], and other aspects. With this separation of data, more precise and relevant policy formulation can be carried out to ensure appropriate interventions in urban and rural areas down to the village level.

The BPS has established criteria for classifying urban and rural villages based on *Peraturan Kepala BPS Nomor 120 Tahun 2020 Tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia Tahun 2020*[8]. This classification of villages into urban and rural villages is used for statistical purposes as a basis for sampling for the allocation of census or survey officers and for determining strategic estimates[8]. This classification of urban and rural villages is based on three main indicators: population density per km^2, the percentage of families working in the agricultural sector, and access to urban facilities. A village will be categorized as an urban area if the total value or score obtained from these indicators reaches or exceeds nine. Meanwhile, areas with a score below nine are classified as rural villages.

The data used to calculate the classification of urban and rural villages comes from the *Potensi Desa* (Podes) data[8]. The frequency of Podes implementation, which is not carried out every year, means that the classification of urban and rural villages is not always updated in line with changes in regional conditions. Changes in the conditions of villages in Indonesia are influenced, among other things, by the Village Fund program, which has significantly improved basic infrastructure, access to facilities, and economic activities in underdeveloped villages, especially in remote areas[9]. This program has accelerated village development, enabling changes in village classification through improvements in facilities and the economy, from rural to urban. Additionally, rapid urbanization is occurring in Indonesian regions due to population growth[10], public infrastructure investment[11], and changes in lifestyle, consumption patterns, mindset, and social interaction patterns[12]. Therefore, alternative, renewable data is needed to support the classification of urban and rural villages due to changes in the conditions of rural areas in Indonesia. In addition, in conducting the Podes Data Collection, BPS collects data from various respondents, including village heads or lurah, village or kelurahan officials, community leaders, and relevant agencies to obtain comprehensive information on the characteristics of villages or kelurahan[13]. This requires significant costs and manpower. Although Podes data is not only used for the classification of urban and rural villages, there is a need for alternative data sources that are more efficient in terms of cost and manpower. This has prompted improvements in the efficiency of the BPS work system, particularly in the classification of urban and rural villages.

The 2020 BPS classification method for urban and rural villages is an improvement on the 2000 criteria, which used three main indicators: population density per km^2, percentage of families working in the agricultural sector, and access to urban facilities[8]. This classification method emphasizes physical urban facilities rather than urban functions, so that areas close to urban areas but without their own urban facilities may be rated low and categorized as villages, even though their functions are actually urban. Additionally, the scores assigned are discrete (1–8 for population density/number of people working in agriculture and 0–1 for facilities). This standardizes the detection of urban conditions across all regions, which may have differing or unbalanced conditions between regions[14]. This means that the reliance of this method on variable selection, scoring interval determination, and the determination of urban facility distances for classifying urban and rural villages may not fully reflect the actual conditions of the area. Therefore, an alternative method is needed that can more broadly capture the real conditions on the ground.

Several machine learning algorithms such as Random Forest and Support Vector Machine have been used previously for urban area detection and urban–rural classification [15][16]. The type of classification used in this study is binary classification because there are only two categories in the classification, namely urban and rural. Based on various studies related to binary classification, the Decision Tree and Support Vector Machine algorithms are suitable for use because they have a strong ability to clearly separate the two classes and consistently show high classification performance [17][18][19]. Therefore, all two algorithms will be used and compared in this study.

Artificial Neural Network (ANN) and other deep learning methods are not used in this study because they generally require a large volume of training data, extensive computational resources, and longer training times. In contrast, the dataset used in this research has a relatively limited number of samples, and the main objective is to evaluate model interpretability and variable importance rather than achieving the highest possible prediction accuracy. Tree-based and kernel-based methods such as

Decision Tree and SVM are more interpretable and efficient for structured tabular data with limited sample sizes, making them more appropriate for this research context.

The use of satellite imagery data obtained through remote sensing is a potential solution in supporting the classification of urban and rural villages. Remote sensing is defined as the acquisition of information about an object without physical contact, where information is obtained by detecting and measuring changes caused by the object in its surrounding environment, whether electromagnetic, acoustic, or potential fields[20]. Remote sensing technology has played a role in supporting global observation since 1972 to the present and has driven the sustainability of satellite innovation[21]. Satellite imagery from remote sensing is also available for free access as open-access products and serves as a tool for comparing and explaining socioeconomic data[22]. This makes satellite image big data an efficient and renewable alternative data source. In addition to satellite image big data, points of interest (POIs) have the potential to be another supporting source in classifying urban and rural villages. POIs are specific locations on a map that are considered important or interesting because they have a specific function or use. POIs have been widely used to extract information related to urban functional areas[23]. POI data is widely available and can be provided by many web map services, such as Google Maps, OpenStreetMap, Bing Maps, Baidu Maps, and Gaode Maps[24].

Early studies that defined the boundary between urban and non-urban areas from satellite imagery were based on vegetation and non-vegetation boundaries[25] using the NDVI. Subsequently, studies on urban areas evolved to include the Urban Index (UI), which focuses on spectral information of built-up areas[26]. After 2000, urban indices that initially focused on general urban characteristics shifted their focus to urban infrastructure and eventually emphasized impervious surface area (ISA) characteristics as an indicator for distinguishing urban and non-urban areas[25]. With the development of urban mapping techniques, several complex indices have been developed by combining optical data, NTL (Nighttime Light), and thermal data. Additionally, other studies have incorporated POI data into urban detection. Urban boundaries can be effectively identified through the distribution of POI density, demonstrating the potential for identifying POI data within urban-scale boundaries[15]. This further supports the detection of urban and non-urban areas.

Machine learning has been used in previous studies to classify urban and non-urban (rural) villages or, more broadly, to classify urban and non-urban areas. Machine learning algorithms utilize large amounts of input and output data to identify patterns, enabling them to “learn” independently in training machines to provide recommendations or make decisions automatically[27]. Machine learning has the capacity to handle high-dimensional data and map classes with complex characteristics[28]. The use of features from satellite imagery big data and POI can be classified effectively using machine learning. This demonstrates that the machine learning approach enables better classification results in classifying urban and rural villages.

2. Research Method

This study focuses on classifying urban and rural villages in North Sumatra Province using satellite imagery and Point of Interest (POI) data, with 2020 as the reference year due to the availability of official classification data from BPS. Predictor variables were derived from satellite indices (e.g., NDVI, NDBI, NDWI), air pollutants, and POI features representing socioeconomic infrastructure. Satellite data was processed through Google Earth Engine and POI data collected via the Google Maps API. Feature extraction employed zonal statistics for satellite data and POI density and distance calculations for POI data. The features were transformed using the Yeo-Johnson method and standardized using z-score normalization, with the target variable encoded as binary. Correlation analysis and feature selection were based on point biserial correlation. Machine learning models—Decision Tree and SVM—were developed with an 80:20 train-test split using stratified sampling, optimized through Bayesian search, and evaluated with accuracy, precision, recall, F1-score, AUC, MCC, and log loss.

2.1. Study Area

North Sumatra Province was selected as the case study area with 2020 as the reference year. North Sumatra Province was chosen because it has 6,113 villages, making it the province with the fourth-highest number of villages in Indonesia in 2020. The selection of 2020 as the reference year was due to the fact that the latest official statistics on the classification of urban and rural villages by BPS were only available for 2020. This study focuses on the village level.

2.2. Data and Data Sources

This study uses target variables in the form of urban and rural village classifications based on official data from BPS. This classification refers to *Peraturan Kepala BPS Nomor 120 Tahun 2020 Tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia Tahun 2020*[8]. In this regulation, villages are categorized into two main classes, namely urban and rural. This classification is based on criteria specified in the regulation.

The predictor variables used in this study were obtained from various indicators representing urban and rural villages sourced from satellite imagery and POIs. The satellite imagery variables used include Night-Time Light (NTL) from NOAA-20, Normalized Difference Vegetation Index (NDVI), Normalized Difference Built-up Index (NDBI), and Normalized Difference Water Index (NDWI) from Sentinel-2, NO_2 concentration, SO_2, and CO from Sentinel-5P, and Land Surface Temperature (LST) from Terra [25][29][30][31]. These images were collected between January 1, 2020, and December 31, 2020, making them relevant to the official BPS classification. The use of these images is based on previous literature studies and the researchers' identification of indicators capable of representing the characteristics of urban and rural villages. All satellite images were collected and processed through Google Earth Engine.

In the POI data, the variables used are POI Density and POI Distance for the categories of education, economy, health, and entertainment. The use of these categories is matched based on the *Peraturan Kepala BPS Nomor 120 Tahun 2020 Tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia Tahun 2020*[8]. For the education category, the POIs utilized include locations for kindergartens/preschools, junior high schools/MTs, and senior high schools/vocational schools/MA. Meanwhile, for the economic category, POIs for markets and shops are used. In the health category, POIs include hospitals, including maternity hospitals. The entertainment category includes hotels, billiards, pubs, discos, karaoke venues, and salons. Table 1 show relationship between research variables representing BPS Criteria

Table 1. Relationship between Research Variables and BPS Criteria

Research Variables	BPS Classification Variables Represented	Relationship and Rationale	Source of Approach
Night-Time Light (NTL)	Population density per km^2 and access to urban facilities	Higher NTL reflects greater human activity, including economic activity, population, and infrastructure.	[32]
Normalized Difference Vegetation Indeks (NDVI)	Percentage of families working in the agricultural sector	High NDVI indicates the presence of extensive agricultural land and dense vegetation, characteristic of rural areas.	[33][29]

Normalized Difference Built-up Indeks (NDBI)	Population density per km^2	A high NDBI indicates the dominance of built-up areas such as buildings and infrastructure, which is directly related to high population density, as areas with many buildings are usually more densely populated.	[34]
Normalized Difference Water Indeks (NDWI)	Population density per km^2 and percentage of families working in the agricultural sector	NDWI values above 0 generally represent the presence of water bodies, while NDWI values below 0 indicate non-water areas such as soil or vegetation. Combining NDWI with other indices can improve the mapping of built-up areas for urban and rural classification.	[35][29]
NO_2, SO_2, dan CO concentration	Population density per km^2	High concentrations of pollutant gases reflect high levels of transportation, industrial activity, and population density.	[36][31][37][38][39][40]
Land Surface Temperature (LST)	Population density per km^2	High land surface temperatures indicate the urban heat island effect that often occurs in densely populated urban areas.	[41][42]
POI Density	Access to urban facilities	POI density indicates the number of facilities available in a given area. Areas with high POI density—such as many schools, hospitals, markets, and entertainment venues—indicate good accessibility and urban characteristics.	[43][44]
POI Distance	Access to urban facilities	Distance to POI indicates the level of ease of access to these facilities. Villages with a long distance to POI have limited access to public services, reflecting rural conditions.	[43][44]

2.3. Data Collection and Preprocessing

The official classification of urban and rural villages was obtained from *Peraturan Kepala BPS Nomor 120 Tahun 2020 Tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia Tahun 2020*[8]. Meanwhile, satellite imagery data was collected using Google Earth Engine, and POI data was collected from Google Maps. During the satellite image data collection process, data preprocessing was also carried out simultaneously.

In the four satellite data sources used (NOAA-20, Sentinel-2, Sentinel-5p, Terra), several preprocessing steps were applied to prepare the data for further analysis. The first step was time filtering by selecting the reference time range from January 1, 2020, to December 31, 2020. Next, the area was cropped to limit the data to only include the relevant geographical area, namely North Sumatra. Finally, to reduce fluctuations or noise that could arise due to daily variations or the influence of weather conditions, the median value of the images from each satellite collection was calculated during the relevant period, namely 2020. To obtain the NDVI, NDBI, and NDWI values from the Sentinel-2 images, band compositing was performed by applying the formulas described in equations (1), (2), and (3).

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

$$NDBI = \frac{SWIR_{band\ 11} - NIR_{band\ 8}}{SWIR_{band\ 11} + NIR_{band\ 8}} \quad (2)$$

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

POI data collection is carried out automatically using the Google Maps API service (via the gomaps.pro proxy endpoint). This process begins by defining search queries that represent POI categories in accordance with BPS Regulation No. 120 of 2020[8], namely “tk”, “ra”, “ba”, “smp”, “mts”, “sma”, “smk”, “ma” for the education category, “pasar” and “toko” for the economic category, “rs” for the health category, and “hotel,” “bilyar,” “pub,” “diskotek,” “tempat karaoke,” and “salon” for the entertainment category. All searches were conducted within a 20-kilometer radius of the coordinates in the selected subdistrict in a particular district (example query: “tk di Kecamatan Idano Gawo Kabupaten Nias”). The search results included the name of the place, address, type of place, and geographical coordinates (latitude and longitude).

2.4. Zonal Statistics, POI Density, and POI Distance

Feature extraction is the process of converting raw data into a more meaningful and informative representation by extracting the most relevant characteristics from the data. In this study, feature extraction for satellite image data involves converting raster data into vector table data. This conversion of raster data into vector table data uses zonal statistics techniques. Zonal statistics calculate the statistical values of raster cells (grids) that fall within a zone (e.g., administrative boundaries, land areas, or ecological zones)[45], which are typically defined by vector polygons such as shapefiles. The statistics used are the median, and the zones used are the administrative boundaries at the village level in North Sumatra Province. The result of this feature extraction is a shapefile containing the median values of raster cells within each administrative village zone in North Sumatra Province for all features (NTL, NDVI, NDBI, NDWI, NO2, SO2, CO, LST).

In addition to satellite imagery, feature extraction was also performed on POI data to represent the socio-economic and infrastructure aspects of the region. Two types of features extracted from POI data are POI Density and POI Distance. POI Density is calculated based on the number of POIs in each village administrative zone according to the categories of education, economy, health, and entertainment. Meanwhile, POI Distance measures the Euclidean distance from the center of the village administrative zone to the nearest POI in each category of education, economy, health, and entertainment.

2.5. Data Transformation, Standardization, and Encoding

The extracted features were then transformed using the Yeo-Johnson transformation[46]. The Yeo-Johnson transformation is an extension of the Box-Cox transformation that can be used for both positive and negative data. The purpose of the Yeo-Johnson transformation is to make the data distribution closer to normal (Gaussian) and the features to be on a uniform scale. In addition, some machine learning algorithms are more sensitive to data scale, so it is important to normalize or standardize numerical features so that they are within the same range[47]. In this study, z-score standardization was performed, which involves changing the data so that it has a mean of zero and a variance of one. In addition to standardizing numerical features, coding was performed for target variable, namely the classification of urban and rural villages. Urban villages were coded as 1 and rural villages were coded as 0.

2.6. Correlation Analysis and Feature Selection

Point biserial correlation (r_{pb}) is the Pearson product moment correlation value when one variable is dichotomous, taking only two possible values coded as 0 and 1 (binary data), and the other variable is numerical (interval or ratio)[48]. The following is the point biserial correlation equation.

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{s}_Y} \sqrt{\frac{N_1 N_0}{N(N-1)}} \quad (4)$$

Point biserial correlation can be interpreted as a measure of effect for the difference in means between two groups[48]. Based on the correlation analysis results, the features used in the modeling are those with a correlation coefficient value ≥ 0.4 (medium) with the official classification of urban and rural villages.

2.7. Development of Classification Models

A classification model for urban and rural villages was developed using satellite imagery data obtained from remote sensing and machine learning. The machine learning algorithms used were Decision Tree (DT) and Support Vector Machine (SVM). In implementing these classification algorithms, the data was divided into 80% training data and 20% test data to train and test the performance of the developed model. This 80:20 division is a common practice in machine learning because it provides sufficient data for model training while leaving an adequate proportion for model evaluation. Data splitting was performed using the stratified split technique, ensuring that the proportions of each class (urban and rural villages) in the training and testing data remained balanced as in the original data distribution[49]. This aims to avoid bias in model training due to class distribution imbalance in the data subsets used.

A decision tree (DT) is a tree structure similar to a flowchart, where each internal node (a node that is not a leaf) represents a test on an attribute, each branch represents the result of that test, and each leaf node (or terminal node) contains a class label, with the top node in a tree being the root node[50]. The DT induction process is performed recursively using a divide-and-conquer approach.

Support Vector Machines were introduced by Vladimir Vapnik and his colleagues Bernhard Boser and Isabelle Guyon[51]. In its implementation, SVM seeks to maximize the margin so that it can create the widest possible distance between the separating hyperplane and the data from the two existing classes.

2.8. Model Optimization and Imbalanced Data Handling

To optimize the modeling results, the best hyperparameters for each classification algorithm were searched for using Bayesian Optimization through the bayes search technique. This process used Stratified 5-Fold Cross-Validation to maintain the class proportions in each fold. The training data was evaluated using several metrics, namely accuracy, precision, recall, and f1-score, with the best model selected based on the f1-score value as the main metric.

In overcoming imbalanced data, data resampling techniques are applied, specifically oversampling and undersampling, with the Random Undersampling (RUS) technique being selected. The data imbalance occurs because the study area — North Sumatra Province — is predominantly composed of rural villages rather than urban ones. Consequently, the number of samples in the rural class is much larger than in the urban class, resulting in an imbalanced dataset that can bias the model toward predicting the majority class. The RUS technique is used to address this imbalance by randomly reducing the number of samples in the majority (rural) class to match the minority (urban) class. This approach is considered appropriate in this study because of the relative homogeneity among rural villages in North Sumatra, where the characteristics of rural areas tend to be similar in terms of vegetation cover, accessibility, and socioeconomic structure. Therefore, random removal of some rural samples is unlikely to eliminate meaningful variation or cause significant information loss. The output produced from the development of this urban–rural village classification model is the classification results from all algorithms using resampled data and their respective combinations.

2.9. Classification Model Evaluation

An evaluation of urban and rural village classification models resulting from machine learning algorithms was conducted. There are nine models whose performance will be tested, consisting of combinations of algorithms and resampling techniques. Model evaluation uses a confusion matrix with model evaluation metrics of precision, recall, and F1-Score[52] for each class, namely the urban village class and the rural village class. For the combined urban village and rural village classes, the metrics used were accuracy, AUC score, MCC, and log loss[52]. The best model was determined based on the highest frequency of achieving the highest values across all evaluation metrics. The following are the formulas for all classification evaluations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$MCC = \frac{TN \cdot TP - FN \,.\, FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

2.10. SHAP

To obtain a more in-depth interpretation of the model, the SHAP approach was used. SHAP is a game theory-based model interpretability method[53] for interpreting the output of any black-box model, such as most machine learning models. Each SHAP value reflects how much a feature contributes to increasing or decreasing the probability of a prediction for a class[54] using Shapley values[55]. By using SHAP, analysis is not limited to the overall importance of features, but can also show how each feature influences the model's decision on each observation, resulting in a more transparent and detailed interpretation.

3. Result and Discussion

3.1. Visual Feature Identification

To gain a visual understanding of satellite imagery data and POIs, the results are visualized on a map of North Sumatra Province with village/subdistrict administrative boundaries, using satellite imagery data and POIs extracted, transformed, and standardized.

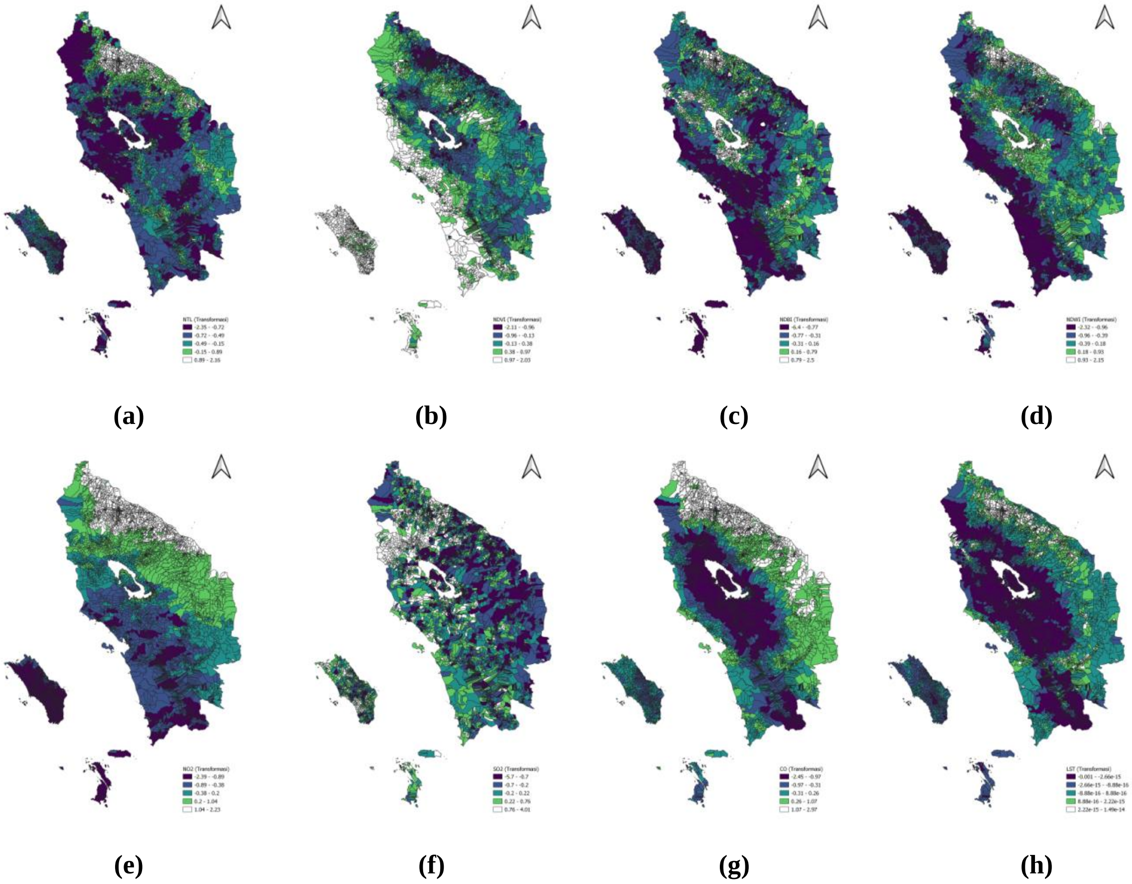
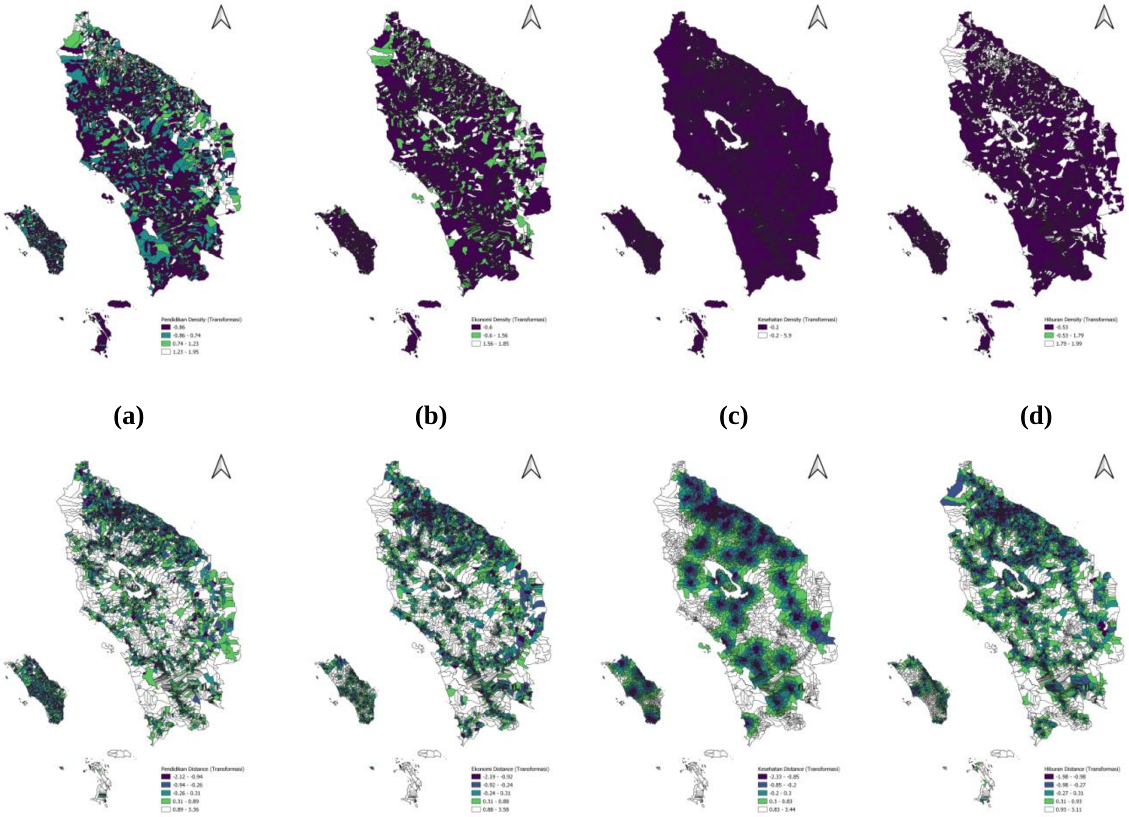


Figure 1. Visual Identification of Satellite Images Resulting from Zonal Statistics and Transformation of (a) NTL, (b) NDVI, (c) NDBI, (d) NDWI, (e) NO2, (f) SO2, (g) CO, and (h) LST

Based on Figure 1 (a, c, d, e, g, h), it can be seen that there are similarities in the patterns of the NTL, NDBI, NDWI, NO2, CO, and LST features. The similarity in the patterns of these features is the high value of these features in the city of Medan and its surroundings. Based on the official classification of urban and rural villages, these areas are the main contributors to urban villages. This pattern indicates a potential positive relationship between the high values of these features and the presence of urban villages. Meanwhile, based on Figure 1 (b), it can be seen that the NDVI feature shows the opposite pattern. The NDVI feature in areas dominated by urban villages such as Medan City and its surroundings actually has a low NDVI value. This indicates a negative relationship between the NDVI feature and urban villages. Additionally, the SO_2 feature in Figure 1 (f) shows a random pattern concentrated in the Karo Regency area. This irregularity is likely caused by volcanic activity from Mount Sinabung in the region, resulting in high SO_2 values. This indicates no relationship between the SO_2 feature and either urban or rural villages.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2. Visual Identification of Points of Interest Based on Density and Distance Calculations and Transformations (a) Education Density, (b) Economic Density, (c) Health Density, (d) Entertainment Density, (e) Education Distance, (f) Economic Distance, (g) Health Distance, and (h) Entertainment Distance

Based on Figure 2 (a, b, c, d), it can be seen that there is no pattern between the POI Education Density, POI Economy Density, POI Health Density, and POI Entertainment Density features. This irregular pattern is likely caused by the uneven and localized distribution of POIs. The existence of these facilities does not always reflect the characteristics of urban areas as a whole, thus showing a pattern. Additionally, some POIs may be scattered in rural villages as local service centers without showing a direct relationship with the overall village classification. Conversely, in Figure 2 (e, f, g, h), a clearer pattern is observed in the POI Education Distance, POI Economy Distance, POI Health Distance, and POI Entertainment Distance features. The values of these features tend to be lower in the Medan City area and its surroundings, which are contributors to urban villages. This reflects the proximity of residents in urban areas to various facilities. This indicates a potential negative relationship between POI Distance values and the presence of urban villages. In other words, the closer the distance to POIs for all categories, the greater the likelihood that the area is classified as an urban village.

3.2. Correlation Analysis Results and Feature Selection

Based on the correlation analysis, the features used in modeling are those with a correlation coefficient ≥ 0.4 (medium) with the official classification of urban and rural villages. Therefore, the features NTL, NDVI, NDBI, NDWI, NO_2, CO, LST, Entertainment POI Density, Education POI Distance, Economic POI Distance, Health POI Distance, and Entertainment POI Distance were selected as representative features in the classification of urban and rural villages. Figure 3 presents the visualization of feature selection results based on the point biserial correlation coefficient. The figure shows that the NTL, NDBI, and NDWI features have the strongest positive association with urban classification, indicating that areas with higher nighttime brightness and built-up intensity are more likely to be categorized as urban. Conversely, NDVI exhibit negative correlations, reflecting the dominance of vegetation typically found in rural areas. Air pollution indicators (NO_2 and CO) and LST also display moderate positive relationships with urban classification, reinforcing the linkage between anthropogenic activities and urban environments. Meanwhile, POI-related features show a clear spatial distinction between urban and rural areas. Urban villages generally exhibit higher POI density, indicating a greater concentration of facilities such as economic centers, schools, health services, and entertainment venues within a smaller area. In contrast, rural villages tend to have higher POI distances across all categories—education, economy, health, and entertainment—reflecting lower accessibility due to more dispersed facility locations and less developed infrastructure. This pattern aligns with the typical spatial organization where urban areas function as service hubs with dense facility networks, while rural areas are characterized by wider spatial distribution and limited access to essential public services. Overall, the visualization in Figure 3 supports the selection of these features as they effectively capture the multidimensional contrasts between urban and rural characteristics, both from environmental and accessibility perspectives.

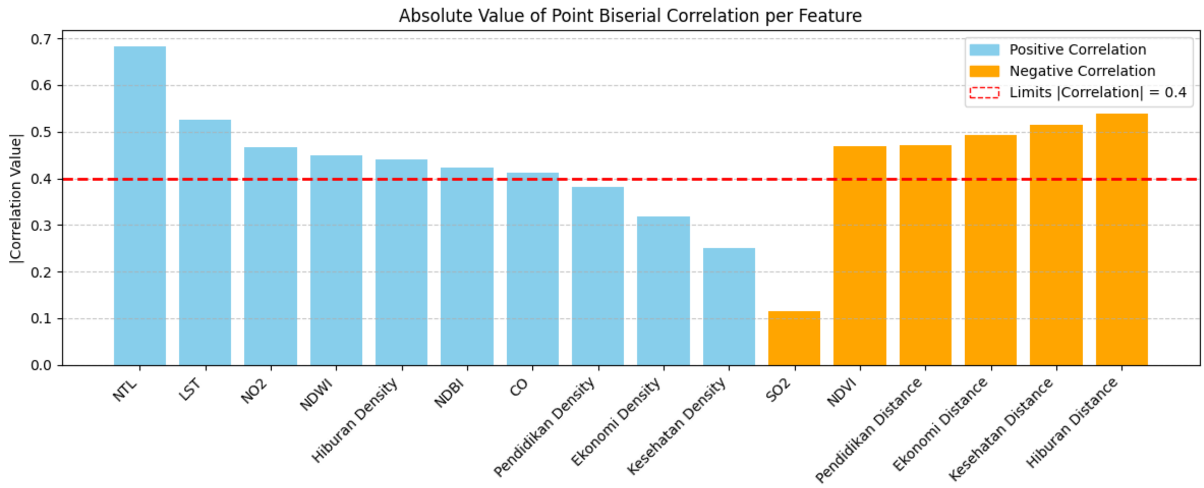


Figure 3. Feature Selection Visualization Based on Point Biserial Correlation Coefficient

3.3. Machine Learning Model Development Results

Based on the best hyperparameters selected, the F1-score evaluation results from the 5-Fold Cross Validation were obtained. These results can be seen in Table 2. From the table, it can be seen that for each fold, the Support Vector Machine model with RUS training data tends to have the highest F1-score value. This indicates that the Support Vector Machine model with RUS training data is able to improve its ability to recognize urban and rural village classes, thereby achieving a better balance between precision and recall.

Table 2. F1-score evaluation results from 5-fold cross-validation Model development based on the best hyperparameters

Model	Data	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
	Original	0.7780	0.7451	0.7541	0.7152	0.7556	0.7496

Decision Tree (DT)	RUS	0.7532	0.7665	0.7617	0.7701	0.7059	0.7515
Support Vector Machine (SVM)	Original	0.7656	0.7822	0.7556	0.7655	0.7637	0.7665
	RUS	0.7759	0.7510	0.7702	0.7756	0.7705	0.7686

3.4. *Results of the Best Machine Learning Model Selection*

Evaluate the model using precision, recall, and F1-score evaluation metrics. These evaluation metrics are calculated from the confusion matrix. In modeling, the positive class is defined as urban villages. Table 3 shows the evaluation of the machine learning model.

Table 3. Classification Performance per Class

Class	Model	Data	Precision	Recall	F1-Score	Support
Urban Class	Decision Tree (DT)	Original	0.8008	0.6614	0.7244	316
		RUS	0.6742	0.7595	0.7143	316
	Support Vector Machine (SVM)	Original	0.8031	0.6582	0.7235	316
		RUS	0.8074	0.6899	0.7440	316
Rural Class	Decision Tree (DT)	Original	0.8888	0.9427	0.9149	907
		RUS	0.9123	0.8721	0.8918	907
	Support Vector Machine (SVM)	Original	0.8880	0.9438	0.9150	907
		RUS	0.8972	0.9427	0.9194	907

Based on Table 3, the model performance shows that Support Vector Machine, especially when combined with data balancing techniques RUS, provides the best results in classifying urban and rural villages. For the urban class, although the initial precision was high but the recall was low in the original data, the application of RUS successfully improved the recall significantly. The highest F1-Score for this class was achieved by Support Vector Machine with RUS (74.40%). Meanwhile, for the rural class, all models showed very high performance with precision and recall approaching perfection, where Support Vector Machine with the RUS achieved the highest F1-Score (91.94%).

Table 4. Overall Model Performance (Macro and Weighted Averages)

Class	Model	Data	Precision	Recall	F1-Score	Support
Macro Average	Decision Tree (DT)	Original	0.8448	0.8020	0.8197	1223
		RUS	0.7932	0.8158	0.8030	1223
	Support Vector Machine (SVM)	Original	0.8455	0.8010	0.8192	1223
		RUS	0.8523	0.8163	0.8317	1223
Weighted Average	Decision Tree (DT)	Original	0.8660	0.8700	0.8657	1223
		RUS	0.8508	0.8430	0.8459	1223
	Support Vector Machine (SVM)	Original	0.8660	0.8700	0.8655	1223
		RUS	0.8740	0.8774	0.8741	1223

Furthermore, as shown in Table 4, the overall model performance based on macro and weighted averages reinforces these findings. The Support Vector Machine with RUS yielded the highest macro average F1-Score (83.17%) and weighted average F1-Score (87.41%), indicating that the model not only performs well across both classes but also maintains balanced predictive capability when accounting for class distribution. This suggests that the RUS technique effectively mitigates class imbalance, enabling the Support Vector Machine to generalize better and achieve more stable classification performance compared to the Decision Tree model.

The macro and weighted average F1-scores are below 90% primarily because of the differences in model performance between the two classes and the characteristics of the dataset, rather than being caused by the algorithm itself. Although the rural class achieves a very high F1-score (91.94%), the urban class shows a noticeably lower F1-score (74.40%), indicating that a relatively larger number of urban villages were misclassified as rural. This misclassification occurs because the urban class has more heterogeneous and overlapping feature distributions, making it harder for the model to draw clear boundaries between urban and rural areas. Several features derived from remote sensing indices and POI data may not fully capture subtle urban characteristics, such as moderate building density or mixed land use, which often resemble semi-urban or rural patterns in certain regions. As a result, even robust models like Support Vector Machines cannot perfectly separate the two classes. Although the application of RUS improved the balance between classes and enhanced recall for the urban class, it also reduced some informative samples from the rural class, slightly decreasing the model’s overall generalization ability. Consequently, these factors collectively contribute to the macro and weighted average F1-scores remaining below 90%.

Table 5. Evaluation of Accuracy, AUC Score, MCC, and Log-loss of Machine Learning Models

Model	Data	Accuracy	AUC Score	MCC	Log-Loss
Decision Tree (DT)	Original	0.8700	0.8946	0.6454	1.1749
	RUS	0.8430	0.8885	0.6086	1.2595
Support Vector Machine (SVM)	Original	0.8700	0.8786	0.6450	0.3324
	RUS	0.8774	0.9154	0.6676	0.3298

Based on Table 5, the application of RUS in the Support Vector Machine model reinforces the evaluation results with the previous metrics, namely providing optimal results. An accuracy value of 87.74% in Support Vector Machine with RUS indicates that the model is able to correctly predict classes in almost 88% of all observations. Additionally, the AUC score of 0.9154 indicates that the model performs very well in distinguishing between urban and rural villages. The MCC value of 0.6676 indicates a strong correlation between the actual classification and the model's predictions. This suggests that the Support Vector Machine model with RUS successfully improves the balance of predictions between the two classes. Meanwhile, the log loss value of 0.3298 indicates that the Support Vector Machine model not only produces accurate predictions but also provides high probabilities for those predictions. This signifies that the application of RUS successfully enhances the model's confidence in the predictions made.

3.5. Results of the Best Feature Importance Model in Machine Learning

Feature importance analysis was conducted to determine the contribution of features in the formation of the best model, namely Random Forest with SMOTE, to the classification of urban and rural villages. Figure 4 shows the order of the contribution of features in the formation of the best model, namely Random Forest with SMOTE.

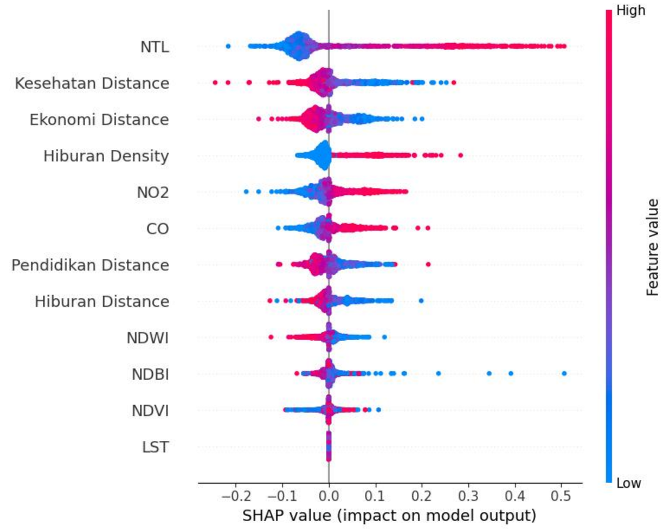


Figure 4. Features Importance based SHAP from Support Vector Machine Model with RUS

Based on Figure 4, an in-depth picture of the direction and magnitude of each feature's contribution to the prediction of urban and rural village classification models can be seen. The NTL feature shows that high values tend to drive predictions towards urban villages, while low values lead to rural villages. On the other hand, the distance to public facilities features such as POI Health Distance, POI Economy Distance, POI Education Distance, and POI Entertainment Distance with low values contribute positively to the prediction of urban villages, while high values lead to rural villages. NO_2, and CO features show that high values drive the model to predict urban villages. The NDWI, NDBI, NDVI, and LST features in the SHAP summary plot graph appear to cluster around zero. SHAP values close to zero indicate that changes in the values of these features do not significantly affect the model output in most observations. In other words, the Support Vector Machine model with RUS does not rely heavily on the NDWI, NDBI, NDVI, and LST features to make classification decisions.

3.6. Analysis of Misclassification Causes and Recommendations for Statistics Indonesia (BPS)

The misclassification observed in the urban–rural classification model primarily stems from the characteristics and representativeness of the variables used. Several features derived from remote sensing and POI data do not fully capture the subtle distinctions between urban and rural areas, especially in transition zones or semi-urban regions. For instance, some villages with moderate NDBI and NDVI values exhibit a mix of built-up and vegetated areas, making them difficult to categorize precisely. Similarly, NDWI values, which can indicate both the presence of water bodies and dry built-up surfaces, may overlap between classes, leading to ambiguity.

Moreover, Night-Time Light (NTL) intensity varies not only by urbanization level but also due to differences in electrification or infrastructure development. Rural villages with limited lighting infrastructure but high human activity (e.g., near plantations or industrial sites) may be incorrectly classified as rural instead of urban. The POI Distance and POI Density features, although informative, also introduce complexity—small urban centers may have relatively few POIs compared to larger cities, while some rural areas with nearby facilities (such as schools or health centers) may appear more urban in spatial terms. These overlaps in variable representation contribute to the model’s difficulty in drawing clear decision boundaries, resulting in misclassification of certain villages.

To improve classification accuracy, future research should consider integrating additional or refined variables that better capture socioeconomic and spatial characteristics, such as road network density, building footprint data, daytime population mobility, or temporal patterns of satellite imagery. The use of higher spatial resolution imagery (e.g., Sentinel-1 or PlanetScope) may also enhance the model's sensitivity to small-scale variations within villages.

For Statistics Indonesia (BPS), these findings highlight the potential of combining machine learning techniques with remote sensing and POI data as a complementary approach to the official urban–rural classification framework. Machine learning offers a scalable and data-driven method that can assist in updating village classifications more frequently and objectively. However, to ensure methodological consistency, BPS should align such implementations with the existing classification dimensions—population density, agricultural livelihood, access to facilities, and social conditions—and promote the use of interpretable models that can provide insights into the contribution of each variable to classification outcomes.

4. Conclusion

Overall, this research demonstrates that the integration of remote sensing data and Point of Interest (POI) information can effectively be used to classify urban and rural villages using machine learning approaches. The application of the Support Vector Machine algorithm combined with Random Undersampling (RUS) provides the best classification performance, successfully addressing the issue of class imbalance while maintaining robust accuracy and generalization capability. The model's strong performance metrics—particularly the weighted average F1-score of 87.88% and AUC of 0.9154—indicate that the developed approach is reliable for regional classification at the village level. The findings also emphasize the crucial role of the Nighttime Light (NTL) feature in distinguishing urban and rural areas, as it effectively captures variations in human activity and infrastructure intensity. Despite these promising results, the model's slightly lower performance in classifying urban villages suggests that certain features still overlap between classes, highlighting the need for more discriminative variables or higher-resolution data in future research. Therefore, subsequent studies are encouraged to explore the integration of additional spatial, socioeconomic, and temporal features, as well as the use of advanced ensemble or deep learning methods, to further enhance the precision and interpretability of village-level urban–rural classification models.

Acknowledgement

If there is any, the author may express gratitude to people, organizations, or institutions that helped or supported the work in some way.

References

[1] M. Patole, "Localization of SDGs through disaggregation of KPIs," *Economies*, vol. 6, no. 1, p. 15, 2018, doi: 10.3390/economies6010015.

[2] R. D. Wulandari, A. D. Laksono, Z. K. Nantabah, N. Rohmah, and Z. Zuardin, "Hospital utilization in Indonesia in 2018: do urban–rural disparities exist?," *BMC Health Serv. Res.*, vol. 22, no. 491, pp. 1–11, 2022, doi: 10.1186/s12913-022-07896-5.

[3] M. Hayashi, "Roles of Education in Expenditure Inequality between Urban and Rural Areas: Indonesia, the Philippines, and India," in *Industrial Location and Vitalization of Regional Economy*, Singapore: Springer, 2023, pp. 165–184. doi: 10.1007/978-981-19-8128-9_9.

[4] R. N. T. Wahyuni, M. Ikhsan, A. Damayanti, and K. Khoirunurrofik, "Inter-District Road Infrastructure and Spatial Inequality in Rural Indonesia," *Economies*, vol. 10, no. 9, p. 229, 2022, doi: 10.3390/economies10090229.

[5] B. van Leeuwen and P. Földvári, "The Development of Inequality and Poverty in Indonesia, 1932–2008," *Bull. Indones. Econ. Stud.*, vol. 52, no. 3, pp. 379–402, 2016, doi: 10.1080/00074918.2016.1184226.

[6] A. Setyadharma, S. Oktavilia, S. U. Audina, and R. Noormalitasari, "Impacts of Education and Environmental Sustainability on Rural Income Inequality in Indonesia," in *E3S Web of Conferences, 2021*. doi: 10.1051/e3sconf/202123204003.

[7] S. Cook and T. Winfield, "The urban–rural divide, regional disaggregation and the convergence of crime," *Appl. Econ., vol.* 47, no. 47, pp. 5072–5087, 2015, doi: 10.1080/00036846.2015.1042143.

[8] BPS, *Peraturan Kepala Badan Pusat Statistik Nomor 120 Tahun 2020 Tentang Klasifikasi Desa Perkotaan dan Perdesaan di Indonesia 2020 Buku 1 Sumatera*. Jakarta: Badan Pusat Statistik, 2020.

[9] N. Hartojo, M. Ikhsan, T. Dartanto, and S. Sumarto, "A Growing Light in the Lagging Region in Indonesia: The Impact of Village Fund on Rural Economic Growth," *Economies*, vol. 10, no. 9, 2022, doi: 10.3390/economies10090217.

[10] F. H. Mardiansjah, P. Rahayu, and D. Rukmana, "New Patterns of Urbanization in Indonesia: Emergence of Non-statutory Towns and New Extended Urban Regions," *Environ. Urban. ASIA*, vol. 12, no. 1, pp. 11–26, 2021, doi: 10.1177/0975425321990384.

[11] C. Silver, "Rapid Urbanization: The Challenges and Opportunities for Planning in Indonesian Cities," in *The Indonesian Economy and the Surrounding Regions in the 21st Century*, Singapore: Springer, 2024. doi: 10.1007/978-981-97-0122-3_3.

[12] J. Junaedi, D. Dikrurohman, and A. Abdullah, "Analysis of Social Change in Rural Communities Due to the Influence of Urbanization and Globalization in Indonesia," *Edunity Soc. Educ.* Stud., vol. 2, no. 3, pp. 431–441, 2023, doi: 10.57096/edunity.v2i3.76.

[13] BPS, *Pedoman Pencacah Pemutakhiran Data Perkembangan Desa*. Jakarta: Badan Pusat Statistik, 2020.

[14] S. Wilonoyudho, R. Rijanta, Y. T. Keban, and B. Setiawan, "Urbanization and Regional Imbalances in Indonesia," *Indones. J. Geogr.*, vol. 49, no. 2, pp. 125–132, 2017, doi: 10.22146/ijg.13039.

[15] Z. Wang, H. Wang, F. Qin, Z. Han, and C. Miao, "Mapping an Urban Boundary Based on Multi-Temporal Sentinel-2 and POI Data: A Case Study of Zhengzhou City," *Remote Sens*., vol. 12, no. 24, p. 4103, 2020, doi: 10.3390/rs12244103.

[16] M. S. Sari, D. Safitri, and S. Sugito, "Klasifikasi Wilayah Desa-Perdesaan Dan Desa-Perkotaan Wilayah Kabupaten Semarang Dengan Support Vector Machine (SVM)," *J. Gaussian*, vol. 3, no. 4, pp. 751–760, 2014, doi: 10.14710/j.gauss.3.4.751-760.

[17] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

[18] R. Kumari and S. K. Srivastava, "Machine Learning: A Review on Binary Classification," *Int. J. Comput. Appl*., vol. 160, pp. 11–15, 2017.

[19] J. Wainer, "Comparison of 14 different families of classification algorithms on 115 binary datasets," *arXiv Prepr.* arXiv1606.00930, pp. 1–36, 2016, doi: 10.48550/arXiv.1606.00930.

[20] C. Elachi and J. van Zyl, *Introduction to the Physics and Techniques of Remote Sensing:* Second Edition. 2006. doi: 10.1002/0471783390.

[21] P. K. Hargreaves and G. R. Watmough, "Satellite Earth Observation to Support Sustainable Rural Development," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103. 2021. doi: 10.1016/j.jag.2021.102466.

[22] T. A. Kugler et al., "People and Pixels 20 Years Later: The Current Data Landscape and Research Trends Blending Population and Environmental Data," *Popul. Environ.*, vol. 41, no. 2, 2019, doi: 10.1007/s11111-019-00326-5.

[23] X. Liu and Y. Long, "Automated identification and characterization of parcels with OpenStreetMap and points of interest," *Environ. Plan. B Plan. Des.*, vol. 43, no. 2, pp. 341–360, 2015, doi: 10.1177/0265813515604767.

[24] Z. Deng, Y. Chen, X. Pan, Z. Peng, and J. Yang, "Integrating GIS-Based Point of Interest and Community Boundary Datasets for Urban Building Energy Modeling," *Energies*, vol. 14, no. 4, p. 1049, 2021, doi: 10.3390/en14041049.

[25] A. Javed et al., "Review of Spectral Indices for Urban Remote Sensing," *Photogramm. Eng. Remote Sensing*, vol. 87, no. 7, 2021, doi: 10.14358/PERS.87.7.513.

[26] M. Kawamura, S. Jayamanna, and Y. Tsujiko, "Quantitative Evaluation of Urbanization In Developing Countries Using Satellite Data," *Doboku Gakkai Ronbunshu*, vol. 1997, no. 580, 1997, doi: 10.2208/jscej.1997.580_45.

[27] J. M. Helm et al., "Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions," *Current Reviews in Musculoskeletal Medicine*, vol. 13, no. 1. 2020. doi: 10.1007/s12178-020-09600-8.

[28] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *International Journal of Remote Sensing*, vol. 39, no. 9. 2018. doi: 10.1080/01431161.2018.1433343.

[29] Y. Zheng, L. Tang, and H. Wang, "An improved approach for monitoring urban built-up areas by combining NPP-VIIRS nighttime light, NDVI, NDWI, and NDBI," *J. Clean. Prod*., vol. 328, 2021, doi: 10.1016/j.jclepro.2021.129488.

[30] N. Kadhim, N. T. Ismael, and N. M. Kadhim, "Urban Landscape Fragmentation as an Indicator of Urban Expansion Using Sentinel-2 Imageries," *Civ. Eng. J.*, vol. 8, no. 9, 2022, doi: 10.28991/CEJ-2022-08-09-04.

[31] M. Iqbal, B. Susilo, and D. R. Hizbaron, "Spatial Variation of Pollutant (NO2, SO2 & CO) and Its Impact Factors in Jakarta: An Application of Sentinel-5P Products," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1406, no. 1, pp. 1–10, 2024, doi: 10.1088/1755-1315/1406/1/012007.

[32] P. Xiao, X. Wang, X. Feng, X. Zhang, and Y. Yang, "Detecting China's urban expansion over the past three decades using nighttime light data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 10, 2014, doi: 10.1109/JSTARS.2014.2302855.
[33] N. Pettorelli, J. O. Vik, A. Mysterud, J. M. Gaillard, C. J. Tucker, and N. C. Stenseth, "Using the satellite-derived NDVI to assess ecological responses to environmental change," *Trends in Ecology and Evolution*, vol. 20, no. 9. 2005. doi: 10.1016/j.tree.2005.05.011.
[34] Y. Zha, J. Gao, and S. Ni, "Use of normalized difference built-up index in automatically mapping urban areas from TM imagery," *Int. J. Remote Sens.*, vol. 24, no. 3, 2003, doi: 10.1080/01431160304987.
[35] S. K. McFeeters, "The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, 1996, doi: 10.1080/01431169608948714.
[36] WHO, *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. WHO, 2021.
[37] M. Vîrghileanu, I. Săvulescu, B. A. Mihai, C. Nistor, and R. Dobre, "Nitrogen dioxide (No2) pollution monitoring with sentinel-5p satellite imagery over europe during the coronavirus pandemic outbreak," *Remote Sens.*, vol. 12, no. 21, 2020, doi: 10.3390/rs12213575.
[38] L. Perez et al., "Chronic burden of near-roadway traffic pollution in 10 European cities (APHEKOM network)," *Eur. Respir. J.*, vol. 42, no. 3, 2013, doi: 10.1183/09031936.00031112.
[39] B. Muslimbekov, N. Teshaev, S. Abdurakhmonov, and O. Gaybulloev, "Monitoring Trends of SO level Using Time-Series Sentinel-5 Images Based on Google Earth Engine," *E3S Web Conf.*, vol. 563, p. 03068, Aug. 2024, doi: 10.1051/e3sconf/202456303068.
[40] L. Mamić, "CO level over the Republic of Croatia using SENTINEL-5P," *GIS Odyssey J.*, vol. 1, no. 1, 2021.
[41] S. Singh, S. K. P. Kushwaha, and K. Jain, "A correlation analysis of land surface temperature and evapotranspiration in urban setting," in *International Archives of the Photogrammetry*, Remote Sensing and Spatial Information Sciences - ISPRS Archives, 2023. doi: 10.5194/isprs-archives-XLVIII-1-W2-2023-1569-2023.
[42] L. Yang, F. Qian, D.-X. Song, and K.-J. Zheng, "Research on Urban Heat-island Effect," *Procedia Eng.*, vol. 169, no. 1, pp. 11–18, 2016, doi: 10.1016/j.proeng.2016.10.002.
[43] Z. Weifeng, L. Qingquan, and L. Bijun, "Extracting hierarchical landmarks from urban POI data," *J. Remote Sens.*, vol. 15, pp. 973–988, 2011.
[44] T. Pateman, "Rural and urban areas: comparing lives using rural/urban classifications," *Reg. Trends*, vol. 43, pp. 11–86, 2011, doi: 10.1057/rt.2011.2.
[45] J. A. Cardille, M. A. Crowley, D. Saah, and N. E. Clinton, *Cloud-Based Remote Sensing with Google Earth Engine*. Mountain View: Springer, 2024. doi: 10.1007/978-3-031-26588-4.
[46] I.-K. Yeo and R. A. Johnson, "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.
[47] P. Bintoro, Ratnasari, E. Wihardjo, I. P. Putri, and A. Asari, *Pengantar Machine Learnin*g. 2024. [Online]. Available: https://repository.um.ac.id/5619/1/fullteks.pdf
[48] D. Kornbrot, "Point Biserial Correlation," *Wiley StatsRef Stat. Ref. Online*, 2014.
[49] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Massachusetts*: The MIT Press, 2015.
[50] M. K. and J. P. Jiawei Han, *Data Mining: Concepts and Techniques, Third Edition*, Morgan Kaufmann Publishers, 2011.
[51] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," *Proc. Fifth Annu. ACM Work. Comput. Learn. Theory*, pp. 144–152, 1992, doi: 10.1145/130385.130401.
[52] O. Rainio, J. Teuho, and R. Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-66611-y.
[53] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014, doi: 10.1007/s10115-013-0679-x.
[54] S. Wang, H. Peng, Q. Hu, and M. Jiang, "Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method," *J. Hydrol. Reg. Stud.*, vol. 42, 2022, doi: 10.1016/j.ejrh.2022.101139.
[55] L. S. Shapley, "A value for n-person games," in *In contribution to the theory of games volume II*, New Jersey: Princeton University Press, 1953, pp. 307–318.