IJIES-Application Of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets

by X Y

Submission date: 12-Jun-2022 11:20PM (UTC+0700)
Submission ID: 1855335085
File name: IJIES_Format_Anik_Vega_Vitianingsih_RV-3.pdf (541.47K)
Word count: 9020
Character count: 46452

Received: October 18, 2020. Revised: December 20, 2020.

International Journal of
Intelligent Engineering & Systems

INASS

http://www.inass.org/

Application Of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets

Anik Vega Vitianingsih[1]* Zahriah Othman[2] Safiza Suhana Kamal Baharin[3] Aji Suraji[4]
Anastasia Lidya Maukar[5]

[1]*Informatics Departments, Universitas Dr. Soetomo, Surabaya, Indonesia*
[1,2,3]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia*
[4]*Department of Civil Engineering, University of Widyagama Malang, Malang, Indonesia*
[5]*Industrial Engineering Department, President University, Bekasi, Indonesia*
* Corresponding author's Email: vega@unitomo.ac.id

Abstract: The difficulty of acquiring data from numerous intergovernmental agencies/institutions for prone road traffic accidents (PRTA) spatial datasets produces a small-scale dataset that causes dataset imbalance. Class imbalance in small-scale datasets causes uncertainty in the results of the modeling PRTA classification. The proposed research is a scenario-based case representation model on the pre-processing data stage to increase the sensitivity of algorithm classification in a small-scale dataset that causes dataset imbalance using machine learning (ML), the synthetic over-sampling method. The retrieval of attributes from the spatial dataset is transformed into the raw dataset, the normalized dataset, the synthetic minority oversampling technique (SMOTE) raw dataset, and SMOTE normalized dataset scenarios. Balancing datasets using four variants of SMOTE, namely ADASYN, Borderline-SMOTE, K-Means SMOTE, and SVM-SMOTE resampled. To evaluate how well the PRTA classification model performed, we utilized the hyper-parameters optimization technique and the genetic algorithm (GA) search cross-validation. Experiments were run with the ML classifier method, including the k-nearest neighbor (KNN), support vector machines (SVM), multilayer perceptron (MLP), naive bayes (NB), logistic regression (LR), and random forest (RF). The Area Under Curve (AUC) was used to evaluate the results of the experiments. The results of the dataset test in a predetermined scenario conclude that a single algorithm that is computationally light to produce an optimal classifier tends to use a raw dataset that is balanced using SMOTE. The KNN method as a single algorithm for classification based on the distance between samples is superior to using datasets on K-Means SMOTE from all algorithms to handle small data sets in unbalanced classes with an average AUC value of 83%, including the good classification category.

Keywords: Spatial analysis, Unbalanced spatial dataset, Over-sampling method, Hyper-parameter optimization, Spatial cross-validation, Machine learning.

1. Introduction

The prone road traffic accidents (PRTA) classification is a critical research topic to contribute to intelligent transportation systems (ITS) [1]–[4]. Researchs with good performance for PRTA classification have offered many ITS methods. However, the method robustness has not been satisfactory [5][6]. Different studies in the requirement gathering for spatial dataset parameters from expert judgments will affect the PRTA classification with the resulting model accuracy value.

Spatial data modeling in geographic information systems (GIS) is related to behavior and the behavior of heterogeneous spatial-temporal datasets (HSTA) [7] because 96% of the data uses private spatial datasets [8][9]. The HSTA data types occur because of GIS characteristics used to solve specific region problems [10]–[16]. The term specific region can be interpreted as a linear network in spatial statistics

[17]. The heterogeneous geospatial data is a result of the acquisition of data from several different government agencies, which causes the data to have many different formats with various structures; the public ontology is used to resolve government-owned public data types in order to show inconsistencies in the acquisition of data sources [18]. These characteristics will affect uncertainty in the spatial-temporal data obtained from many agencies or institutions interested in the spatial data modeling process PRTA classification. The existence of conflicting conflicts provides a small-scale dataset that causes dataset imbalance. Class imbalance in small-scale datasets produces uncertainty in the findings of the modeling PRTA classification. The numerical data set is considered unbalanced if the minority class reaches 40%, which comes from one of the two classes in the classification. As a result, the classification algorithm to be tested will be biased in classifying instances of the minority class [19]. An unbalanced dataset will affect the classification sensitivity value in the minority class if it is not represented evenly [20].

The process of constructing an Artificial Intelligence (AI) model and combining it with experiments on spatial datasets is known as spatial analysis modeling [21], collecting spatial knowledge through the use of spatial datasets and supplying knowledge of models used in the framework through the use of artificial intelligence approaches based on machine learning models from a variety of sources. In GIS, spatial datasets take on the role of the fundamental framework for the development of spatial analysis algorithms, the investigation of algorithmic principles, or the modification of pre-existing algorithms. [22].

The objective of the spatial analysis model is to provide a description of the GIS software that will be produced, as well as to carry out simulations with the goal of putting models based on the AI in ML approach that will be utilized in the proposed framework that has previously been outlined. Spatial datasets in GIS refer to how primary and secondary data are gathered through the collection process, as well as how the data are processed through spatial analysis in order to become information that can be used in a decision support system [23]. Cloud-terminal Integration GIS makes it possible to visualize spatial data and provides a convenient means of doing spatial analysis on a variety of spatial datasets [24] as well as an information retrieval system that is based on an aggregation of spatial datasets [25].

In the field of spatial data mining (SDM), spatial datasets as the key to the value of big data refer to a description of attribute data requirements, how the data is gathered, and what AI approach is utilized to execute spatial analysis of the data [26][24]. In the discipline of machine learning, the categorization model is widely used [27] to be used to research in the field of geographic information system (GIS) spatial analysis. However, due to the fact that the accuracy tests in each study employ different types of sample data, there is no definitive judgment that can be made regarding which classification algorithm is the most effective to apply. In addition to this, it is dependent on the field of study, which is never the same as the subject of the research that is carried out.

Many machines learning (ML) techniques have been proposed over the last three decades to improve the accuracy of the GIS-Spatial for PRTA classification. The unfortunate truth is that spatial data modeling does not now have any available techniques that have a high-performance accuracy value that can be used on the behavior of various spatial datasets (temporal dependency, spatial dependency, spatial-temporal dependency, and exogenous dependency), there is no guarantee that the performance will be satisfactory when one method is applied to different spatial datasets [28]. The ML integrates various algorithms with combined machine learning models to complete tasks in the data mining field, including their classification, clustering, prediction, etc. [29] to improve the robustness [30] [4]. The methods for the single ML classifiers including Decision Tree (DT), NB, kNN, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), LR, SGD Classifier, SVMs, SVM-linear (SVM-L), SVM-RBF (SVM-R), SVM-polynomial (SVM-P), and MLP [31].

Researchers widely use the ML method to achieve the best performance from the model's accuracy value, which is very dependent on the hyper-parameter optimization technique used. [32]. The choice of tuning parameter technique in the ML model used has an essential role in determining the resulting sensitivity analysis [33] and validating the increased best performance model [34][35]. A hyper-parameter optimization technique is a process of tuning parameters that is suitable for training on the ML model [36]. The classification technique in the ML model has a complexity assessment level for hyper-parameter optimization [32]. The hyper-parameter process is determined before the training data is carried out, where the weight and bias of the model used in ML are the parameters that will be learned from the data during training [36].

This paper aims to determine the appropriate approach through scenario procedures at the pre-processing spatial datasets in the GIS spatial data modeling field with a sample of private datasets in the PRTA classification field. The pre-processing stage involves scenarios for prediction modeling PRTA using the raw dataset, the normalized dataset, the synthetic minority oversampling technique (SMOTE) raw dataset, and the SMOTE normalized dataset. This stage handles behavior on small spatial datasets that causes unbalanced classes. The new dataset from the scenario procedure with the best area under the curve (AUC) of receiver operating characteristic (ROC) will be used in the data balancing process with the SMOTE variant, namely ADASYN, Borderline-SMOTE, K-Means SMOTE, and SVM-SMOTE resampled. In order to evaluate the performance of various classification algorithms, including KNN, SVM, MLP, NB, and RF. The model performance derived from the over-sampling method variant used a hyperparameter optimization technique that was performed with a genetic algorithm (GA) search cross-validation.

The results of this study stated that the KNN method is superior with an average AUC value of 83% to all single competition algorithms to handle small datasets in unbalanced classes by pre-processing data using a raw, balanced dataset SMOTE. The results of this study can be recommendation steps that must be carried out in the process of pre-processing spatial analysis for the type of private spatial datasets. This recommendation function can improve the performance of the proposed model. PRTA classification is a very important research topic to contribute to intelligent transportation systems (ITS) [1]–[4]. Researchers with good performance for PRTA classification have offered many ITS methods. However, the method's robustness has not been satisfactory [5][6]. Different studies in the requirement gathering for spatial dataset parameters from expert judgments will affect the PRTA classification with the resulting model accuracy value.

The next discussion in this paper will be explained in sections 2 to 5. Section 2 discusses the related work. Section 3 discusses research methodology related to spatial data collection and imbalanced data techniques. Section 4 discusses the results and discussion for the effectiveness of scenarios on ML classifier and the effects of synthetic data on ML classifier performance using hyper-parameter optimization. Section 5 discusses the conclusions of the entire process in the discussion of this paper.

2. Related work

The discussion in this section will review several previous studies related to the imbalanced data techniques to create new synthetic data, the classification method approach used to test the best-imbalanced data techniques, and the hyper-parameter tuning method used to improve the performance of the classification method.

The research [37] uses Generative adversarial networks (GANs) to create new synthetic data on unbalanced datasets with classification techniques using Convolutional Neural Network (CNN), ResNet-50, Transfer ResNet-50, and Transfer ResNet-50+, where the research results state that The ResNet-50 model has superior classification performance to be applied with the data augmentation in overcoming the problem of small data sets in unbalanced classes for image-based datasets using GAN method. Meanwhile, the research [] proposed a hybrid CSO-FL model, which is the Chicken Swarm Optimization (CSO) and Fuzzy Logic (FL) method used to handle unbalanced datasets using KNN, DT, SVM, and NB classifiers which claim that it can increase accuracy greater than 90% [38].

Many previous researchers have carried out Studies related to the hyper-parameter method on the ML model. The hyper-parameter methods, i.e., manual tuning, grid search, randomized search, GA, PSO, and Bayesian optimization (BO) methods [36][34]. The researchers [39] used the Hyper-parameter technique to determine the sensitivity of the model by testing the accuracy of model validation in a case study of predicting injury severity of traffic accidents using the Recurrent Neural Network (RNN) method; the results of the RNN method were 71.77% superior to the MLP model which only reached 65.48%, and the Bayesian Logistic Regression (BLR) only reached 58.30%. The Bayesian inference method uses a random parameter approach to model the hyper-parameter effects of road-level factors on crash frequency. However, this model is limited to data with small sample size and is only suitable for hierarchical model structures [40][41]. The most widely used hyper-parameter methods are random search, grid search, and manual search. However, this method is computationally impractical [32]. The GA [42] and PSO methods are popular methods used for hyper-parameter techniques [34][43][44].

Based on the results of the literature study in previous studies, this study will propose a scenario model to overcome unbalanced data in creating new synthetic data. The function of this scenario will be used to assess the best dataset that can be used to improve the performance of the selected classification method, namely KNN SVM, MLP, NB, LR, and RF methods. 1st scenario using the raw dataset, 2nd scenario using the normalized raw dataset, 3rd scenario using a raw dataset that is processed with the SMOTE algorithm, and 4th uses a Min-Max normalized dataset that is processed using the SMOTE algorithm. All scenario models will be tested on the selected classification method by looking at the resulting performance value. The best scenario data model based on the results of the classification chosen method will be utilized to tune hyperparameters through GA search cross-validation.

3. Research methodology

The experimental procedure based on the flow in Figure 1 is used at the pre-processing data stage in the proposed machine learning approach for spatial analysis on PRTA classification. This is done to handle small datasets, which cause unbalanced class classifications in spatial datasets for attribute data categories.

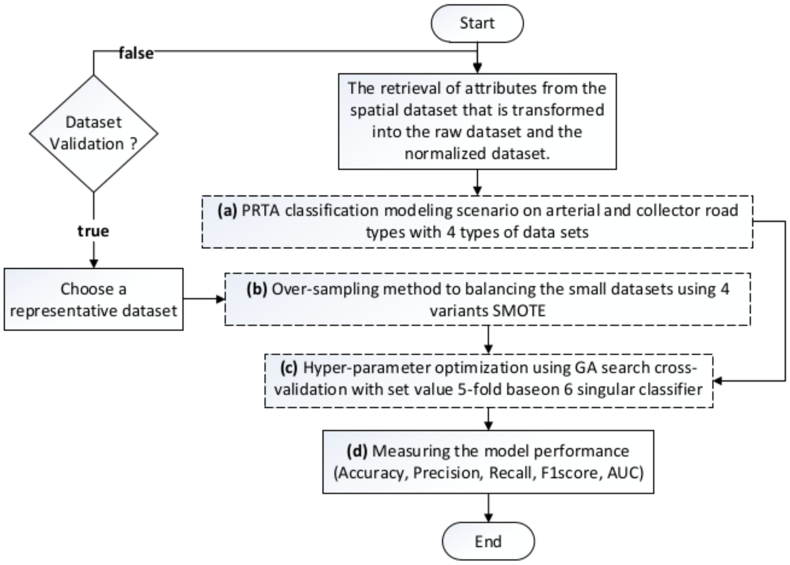

Start
false
Dataset Validation ?
true
Choose a representative dataset
The retrieval of attributes from the spatial dataset that is transformed into the raw dataset and the normalized dataset.
(a) PRTA classification modeling scenario on arterial and collector road types with 4 types of data sets
(b) Over-sampling method to balancing the small datasets using 4 variants SMOTE
(c) Hyper-parameter optimization using GA search cross-validation with set value 5-fold baseon 6 singular classifier
(d) Measuring the model performance (Accuracy, Precision, Recall, F1score, AUC)
End

Figure 1. The experiment procedure

The experiment procedure step:

(a) The dataset validation by taking attributes from private spatial datasets to be transformed into types of raw datasets and normalized raw datasets. The two types of datasets will be used for scenarios for prediction modeling for PRTA classification on arterial and collector road types, including the raw dataset, the normalized raw dataset, the raw dataset which is processed by the SMOTE oversampling method (SMOTE raw dataset), and the normalized dataset which is then processed to the SMOTE oversampling method (SMOTE normalized dataset).

(b) Validation of the performance of the selected dataset scenario based on the highest AUC value resulting from the performance of the classification method in machine learning, namely KNN, SVM, MLP, Naïve Bayes, Logistic Regression, and Random Forest.

(c) This research proposes a hyper-parameter optimization technique model using genetic algorithm (GA) search cross-validation to improve the optimization of the P-RTA classification parameters. This technique aims to improve the performance accuracy value in PRTA classification. Choose a representative dataset to balance classes on small dataset types using four variants SMOTE, such as ADASYN, Borderline-SMOTE, K-Means SMOTE, and SVM-SMOTE resampled to form each new synthetic dataset. GA is a global optimization algorithm based on natural selection theory. To solve the optimization problem, GA represents the intelligent exploits of random searches used to solve optimization problems [45][46]. Although randomly assigned, GA is not at all random, but they exploit historical information to direct the search to better performing areas in the search space. The basic process of genetic algorithm is as follows, although a number of variations are possible [47].

(d) Measuring the Model Performance on the new synthetic dataset is done by comparing the performance results of each classification method in machine learning (including KNN, SVM, MLP, NB, LR, and RF) through the acquisition of accuracy, precision, recall, F1score, and AUC. Classifications with scores between 91-100 percent (very good), 81-90 percent (good), 71-80 percent (fair), and 61-70 percent (poor), and values below 60 percent are considered to be false classifications [48].

3.1 Spatial Datasets Collection

The spatial datasets used in the discussion of this paper use private spatial datasets type for the classification of PRTA based on multi-criteria parameters. The primer data is a map of the arterial and collector road network from a specific region in the National Road Implementation Center for East Java, Bali, Indonesia. The multi-criteria parameters used for spatial data modeling include volume-to-capacity ratio, international roughness index, vehicle type, horizontal alignment, vertical alignment, design speed, and shoulder [49][50].

3.2 Imbalanced data techniques

The class imbalance for small datasets will be overcome using state-of-the-art oversampling algorithms, including ADASYN, Borderline-SMOTE, K-Means SMOTE, and SVM-SMOTE. Data processing uses a new dataset generated from 4 procedural scenarios, including raw dataset, normalized raw dataset, SMOTE raw dataset, and SMOTE normalized dataset.

3.2.1 Synthetic Minority Oversampling Technique (SMOTE)

The SMOTE method works at random observations to increase the number of minority class examples to be equivalent to the majority class through data synthesis based on a k-nearest neighbor. The synthetic sample quality can be done using the first five KNN [20] using Eq. (1) by obtaining a Value Difference Metric (VDM) to make the distance between the two observation vectors through the value of weight (w) and distance (δ) [51]. The data set points from the SMOTE method are placed at any point on the extrapolation line[52].

$$\Delta(X,Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r \qquad (1)$$

where the $\Delta(X,Y)$ the variable is the observation distance between vector X and Y, the $w_x w_y$ variable represents the weight of VDM, the N variable is the number of predictors, the r variable is the synthetic data generator measured by its proximity, the value of *r=1* if using Manhattan distance for categorical data, and *r=2* if using Euclidean distance for numerical data.

3.2.2 ADASYN (Adaptive Synthetic Sampling) resampled

ADASYN is a method for balancing data by approaching it through sampling from unbalanced datasets [53]. The purpose of this method is to reduce the bias caused by class imbalance by learning adaptively about the classification decision. ADASYN generates more synthetic data [53] using Eq. (2) for minority class examples that are harder to learn than minority class examples that are easier to learn. The number of synthetic samples using the ADASYN method will be calculated automatically by determining the weight size for each minority class sample [54].

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \qquad (2)$$

where the x_i variable is the minority class examples for each sample data, the x_{zi} is minority data selected randomly from k- nearest neighbors on data x_i, the $x_{zi} - x_i$ variable is a vector value that states the difference between raw and synthetic data, and the λ variable is a random value of $\lambda \in [0, 1]$.

3.2.3 Borderline-SMOTE resampled

The classification results on the algorithm will achieve better predictive results if learn each class in the training datasets on the borderline instance. Borderline-SMOTE is an over-sampling method that only processes borderline instances from the over-sampled minority class [55]. The synthetic data generator is only carried out in the example borderline [55] to generate a new instance using Eq. (3) by measuring between borderline instances and minority instants using k-nearest neighbors [19].

$$New\ instace = P_i' + gap\ \times (distance\ (P_i', P_j) \qquad (3)$$

Where the P_i' variable is the borderline minority instance class, gap is a random value between 0 and 1, and the P_j variable is the data set chosen at random on the minority instance.

3.2.4 K-Means SMOTE resampled

Received: October 18, 2020. Revised: December 20, 2020.

K-Means SMOTE is an oversampling technique that handles unbalanced classes, consisting of clustering, filtering, and oversampling [56]. Identifying locations in the input space generates synthetic data [57] based on Eq. (4). The sample class clustered by K-means and the original sample class are calculated to select safe samples whose sample classes have not been modified. The new sample synthesis data was obtained from linear interpolation on the safe sample class [58].

$$sample\_weight[k] = \frac{sparsity[k]}{\sum_{all\ i} sparsity[i]} \quad (4)$$

Where the $sample\_weight[k]$ variable is the weight of k-th cluster that has been assigned, the $\sum_{all\ i} sparsity[i]$ variable is the sparsity total of the i-th cluster.

3.2.5 SVM-SMOTE resampled

SVM-SMOTE is an oversampling method to overcome unbalanced classes, how to generate new synthesis data by taking samples in the minority class that is close to the supporting vector in determining the decision limit (SVM) using Eq. (5) [59].

4. Result and discussion

This section will explain experimental results applied to datasets.

4.1 Effectivity of Scenario on ML Classifier

To overcome the availability of small datasets, most researches show that small datasets on class-imbalanced can damage the performance of the ML classifier [60][61]. Scenario models are employed in the pre-processing data stage to deal with the small datasets that lead to uneven class classifications. The following is a proposed scenario for validating datasets to enhance data quality, including:

- 1st scenario: Raw dataset
- 2nd scenario: Normalized raw dataset
- 3rd scenario: SMOTE raw dataset
- 4th scenario: SMOTE normalized dataset

The results of the scenario model in the pre-processing stage of the proposed data are shown in Figures 2 and 3. They can improve the performance of the basic ML singular classifier method (KNN, SVM, MLP, NB, LR, RF). The scenario model results, along with their detailed explanations, can be found in Tables 2 through 5. The process of optimizing hyperparameters will use GA search cross-validation, and these tables will be a part of that (section 4.2).

The experimental results in Figures 2 and 3 state that to obtain a superior classification value, a small unbalanced dataset can be pre-processing data using scenarios 3rd or 4th. The test results state that the KNN method in 5-fold GA cross-validation is superior to the arterial and collector datasets, as shown in Figure 4.

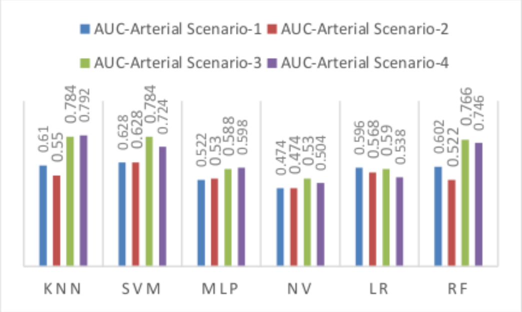


Figure 2. Scenarios for prediction modeling PRTA on arterial roads datasets

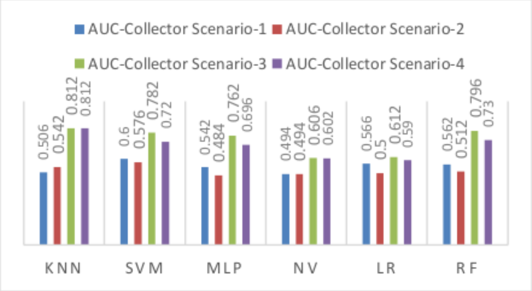


Figure 3. Scenarios for prediction modeling PRTA on collector roads datasets

The results in Table 2 are the performance evaluation values for the 1st scenario for the experiment using raw dataset types tested on two datasets, namely the PRTA classification for arterial and collector road types. The SVM method got the highest AUC values, namely 62.8% and 60%, respectively, while the lowest AUC values were in the NB method, which was 47.4% and 49.4%, respectively.

Table 2. 1st scenario model performance evaluation using raw dataset type

Methods	Performance evaluation				
	Accuracy	Precision	Recall	F1score	AUC
Arterial datasets					
KNN	0.780	0.780	1.000	0.880	0.610
SVM	0.780	0.780	1.000	0.880	**0.628**
MLP	0.762	0.774	0.978	0.868	0.522
NB	0.734	0.796	0.884	0.838	*0.474*

Received: October 18, 2020. Revised: December 20, 2020.

Methods	Performance evaluation				
	Accuracy	Precision	Recall	F1score	AUC
LR	0.768	0.776	0.986	0.872	0.596
RF	0.780	0.780	1.000	0.880	0.602
Collector datasets					
KNN	0.788	0.788	1.000	0.880	0.506
SVM	0.694	0.808	0.804	0.806	**0.600**
MLP	0.788	0.788	1.000	0.880	0.542
NB	0.314	0.292	0.18	0.186	*0.494*
LR	0.788	0.788	1.000	0.880	0.566
RF	0.788	0.788	1.000	0.880	0.562

Table 3 contains the value for the performance evaluation that was determined for the 2nd scenario of the experiment, which made use of the Normalized raw dataset type. These values were validated using two distinct datasets, especially the PRTA categorization for arterial and collector road types. The AUC values obtained using the SVM approach were the highest, coming in at 62.8% and 57.6%, respectively. In contrast, the NB and MLP methods produced the lowest AUC values, which came in at 47.4% and 48.4%, respectively.

Table 3. 2nd Scenario model performance evaluation using normalized raw dataset type

Methods	Performance evaluation				
	Accuracy	Precision	Recall	F1score	AUC
Arterial datasets					
KNN	0.780	0.780	1.000	0.880	0.610
SVM	0.780	0.780	1.000	0.880	**0.628**
MLP	0.762	0.774	0.978	0.868	0.522
NB	0.734	0.796	0.884	0.838	*0.474*
LR	0.768	0.776	0.986	0.872	0.596
RF	0.780	0.780	1.000	0.880	0.602
Collector datasets					
KNN	0.792	0.790	1.000	0.884	0.542
SVM	0.788	0.788	1.000	0.880	**0.576**
MLP	0.788	0.788	1.000	0.880	*0.484*
NB	0.320	0.392	0.186	0.198	0.494
LR	0.216	0.000	0.000	0.000	0.500
RF	0.788	0.788	1.000	0.880	0.512

The result of the performance evaluation value in the 3rd scenario model using the raw dataset type processed with the SMOTE algorithm can be seen in Table 4. The PRTA classification values were tested on the arterial and collector road datasets. The AUC values obtained for the arterial roads using the SVM and KNN methods are equally superior, 78.4% and 57.6%, respectively. The KNN method on the collector dataset is also superior, with an AUC value of 81.2%. In contrast, the NB method on both datasets produces the lowest AUC values, 53%, and 60.6%, respectively.

Table 4. 3rd Scenario model performance evaluation using SMOTE raw dataset type

Methods	Performance evaluation				
	Accuracy	Precision	Recall	F1score	AUC
Arterial datasets					
KNN	0.698	0.78	0.58	0.658	**0.784**
SVM	0.718	0.736	0.692	0.71	**0.784**
MLP	0.554	0.566	0.522	0.508	0.588
NB	0.476	0.39	0.208	0.252	*0.530*
LR	0.546	0.538	0.628	0.58	0.590
RF	0.726	0.716	0.748	0.73	0.766
Collector datasets					
KNN	0.702	0.788	0.562	0.654	**0.812**
SVM	0.712	0.724	0.702	0.708	0.782
MLP	0.674	0.682	0.654	0.666	0.762
NB	0.486	0.394	0.168	0.142	*0.606*
LR	0.542	0.780	0.120	0.202	0.612
RF	0.718	0.750	0.66	0.700	0.796

The experimental results in scenario 4 using the normalized min-max dataset type processed with the smote algorithm can be seen in Table 5. These values were tested for the classification of PRTA on two types of arterial and collector road datasets. The AUC value to measure the performance of the classification method, where the KNN method is equally superior in the arterial and collector datasets, is 79.2% and 81.2%, respectively. In contrast, the NB and LR methods produce the lowest AUC values, 50.4%, and 59%, respectively.

Table 5. 4th Scenario model performance evaluation using min-max normalized and smote datasets type

Methods	Performance evaluation				
	Accuracy	Precision	Recall	F1score	AUC
Arterial datasets					
KNN	0.744	0.794	0.664	0.722	**0.792**
SVM	0.674	0.720	0.592	0.648	0.724
MLP	0.542	0.560	0.484	0.496	0.598
NB	0.474	0.408	0.208	0.250	*0.504*
LR	0.516	0.510	0.556	0.528	0.538
RF	0.692	0.682	0.728	0.700	0.746
Collector datasets					
KNN	0.724	0.786	0.614	0.684	**0.812**
SVM	0.666	0.670	0.678	0.666	0.720
MLP	0.652	0.658	0.660	0.654	0.696
NB	0.498	0.480	0.198	0.200	0.602
LR	0.536	0.358	0.216	0.270	*0.590*
RF	0.656	0.690	0.584	0.628	0.730

4.2 Effect of synthetic data on ML classifier performance using hyper-parameter optimization

This section presents the results of experiments using selected datasets at the pre-processing data stage. Table 6 shows the best scenario model proposed to validate the data set in improving the data quality. Table 6 uses synthetic data generated by several developments of SMOTE methods, including ADASYN, Borderline SMOTE, K-Means-SMOTE, and SVM-SMOTE, to determine the effectiveness of the performance of ML classifier hyper-parameter optimization using GA search cross-validation. GA is a metaheuristic optimization method that has been developed in several domains [62]–[64]. The best ROC-AUC value is declared in bold, whereas the worst value is declared in italics.

The experimental results in Table 6 were tested on two road types of datasets: arterial and collector. Variant SMOTE algorithm using ADASYN method respectively showed that the highest AUC values in the RF method were 79% and KNN 78%. The NB method obtained the lowest values of 50% and 55.8%, respectively. The highest AUC value in the Borderline SMOTE method for the SMOTE algorithm variant shows that the highest AUC value in SVM is 80.6% and KNN at 82%. In comparison, the Naïve Bayes method obtained the lowest values of 55.8% and 57.4%. The variant SMOTE algorithm with the K-Means-SMOTE method shows that the highest AUC values are 89% and 86.6%, respectively, for the KNN method. The lowest values are Logistic Regression 82.6% and Random Forest 75%, respectively. Variant SMOTE algorithm uses the SVM-SMOTE method with the highest AUC values in the Random Forest method of 78.6% and KNN 77%. The lowest values are obtained by the Naïve Bayes method of 50% and 50%. Whereas for some other algorithms, the AUC value is almost the same

The results of voting on the experiment as a KNN is the method that has the highest significant effect in improving the performance of the ML classifier. It has an average AUC value of 0.83 and tends to use balanced raw datasets using SMOTE on small dataset types that are not balanced. As a whole, KNN is the method that has the highest significant effect. KNN is one of the simplest algorithms for looking at the nearest neighbor value [65], even though KNN is considered a poor test on the IRIS dataset [66].

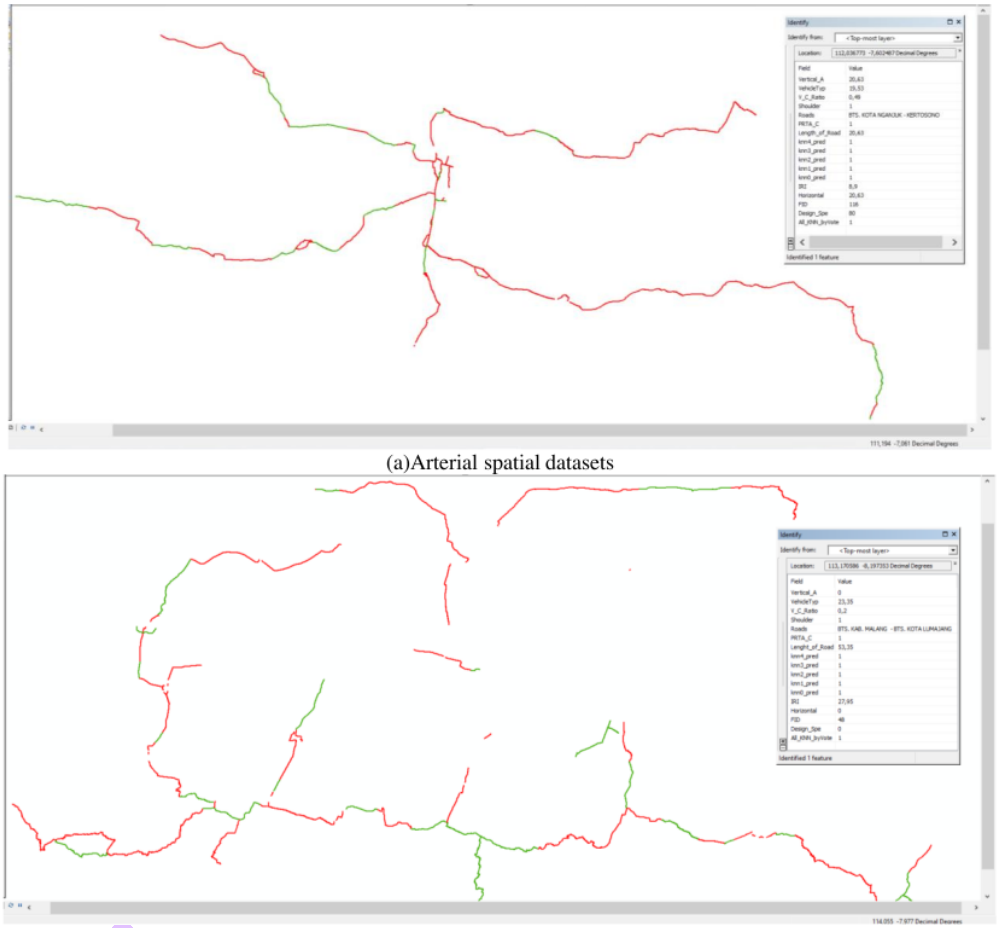
Table 6. The result of balancing datasets and ML classifier hyper-parameter optimization using GA search cross-validation

SMOTE algorithms	Classifiers	Arterial datasets					Collector datasets				
		Accuracy	Precision	Recall	F1score	AUC	Accuracy	Precision	Recall	F1score	AUC
ADASYN	KNN	0.744	0.800	0.65	0.714	0.772(2)	0.69	0.746	0.552	0.628	**0.780(1)**
	SVM	0.696	0.686	0.722	0.702	0.760(3)	0.696	0.688	0.684	0.686	0.748(2)
	MLP	0.518	0.510	0.872	0.642	0.574(5)	0.684	0.714	0.594	0.642	0.696(4)
	NB	0.482	0.414	0.186	0.24	*0.500(6)*	0.522	0.208	0.144	0.132	*0.558(6)*
	LR	0.500	0.500	1.000	0.67	0.580(4)	0.516	0.000	0.000	0.000	0.596(5)
	RF	0.734	0.75	0.722	0.732	**0.790(1)**	0.666	0.726	0.516	0.596	0.728(3)
Borderline SMOTE	KNN	0.714	0.754	0.642	0.69	0.790(2)	0.702	0.776	0.578	0.654	**0.820(1)**
	SVM	0.728	0.714	0.756	0.732	**0.806(1)**	0.714	0.734	0.698	0.710	0.754(3)
	MLP	0.614	0.612	0.642	0.618	0.678(5)	0.680	0.700	0.632	0.664	0.734(4)
	NB	0.478	0.406	0.25	0.296	*0.558(6)*	0.526	0.602	0.212	0.226	*0.574(6)*
	LR	0.612	0.580	0.814	0.68	0.682(4)	0.598	0.638	0.488	0.548	0.636(5)
	RF	0.724	0.730	0.73	0.726	0.764(3)	0.724	0.82	0.576	0.674	0.800(2)
K-Means-SMOTE	KNN	0.814	0.842	0.778	0.804	**0.890(1)**	0.788	0.832	0.722	0.774	**0.866(1)**
	SVM	0.862	0.824	0.92	0.868	0.888(2)	0.830	0.832	0.824	0.826	0.846(3)
	MLP	0.800	0.772	0.858	0.812	0.844(4)	0.502	0.000	0.000	0.000	0.764(5)
	NB	0.774	0.768	0.786	0.776	0.832(5)	0.796	0.812	0.766	0.788	0.862(2)
	LR	0.806	0.778	0.856	0.814	*0.826(6)*	0.796	0.828	0.752	0.786	0.836(4)
	RF	0.858	0.826	0.912	0.866	0.880(3)	0.574	0.766	0.25	0.302	*0.750(6)*
SVM-SMOTE	KNN	0.744	0.800	0.65	0.714	0.772(2)	0.658	0.696	0.506	0.584	**0.770(1)**
	SVM	0.700	0.678	0.766	0.714	0.766(3)	0.694	0.676	0.710	0.690	0.744(2)
	MLP	0.554	0.536	0.792	0.638	0.568(5)	0.682	0.700	0.608	0.646	0.692(4)
	NB	0.482	0.414	0.186	0.240	*0.500(6)*	0.482	0.414	0.186	0.240	*0.500(6)*
	LR	0.500	0.500	1.000	0.670	0.580(4)	0.516	0.000	0.000	0.000	0.596(5)
	RF	0.706	0.708	0.72	0.710	**0.786(1)**	0.660	0.708	0.510	0.586	0.736(3)

4.3 Spatial cross-validation

Cross-validation is an approach in statistical methods that works to evaluate the performance of the model that is built, and this can reduce the bias of the results obtained from the model's performance [67]. The process of cross-validation will divide the data into two parts which are used for learning and validation. For the purpose of cross-validation in hyperparameter tuning of spatial data, spatial partitioning can be utilized [68].

Figure 4 is the result of the 5-fold GA spatial cross-validation in the spatial analysis for the classification of PRTA on the arterial in Figure 4 (a) and collector datasets in Figure 4 (a) using the KNN method, which is a good classifier category based on Table 6 with the K-Means-SMOTE variant dataset. The arterial road dataset type consists of 281 data (178 roads from the real world and 103 data generated from synthetic data in section 4.1 and 4.2) and uses 5-fold GA spatial cross-validation resulting in total predictions of 1410, correct predictions 1302 with an accurate prediction rate of 92.34%. The collector road dataset type includes 316 data, including 201 roads from the real world and 115 data generated from synthetic data in section 4.1. It also employs 5-fold GA spatial cross-validation, which results in a total of 1585 predictions, of which 1378 are accurate, for an accuracy rate of 86.94%.



Identify

Identify from: <Top-most layer>

Location: 112,036773 -7,602487 Decimal Degrees

Field	Value
Vertical_A	20,63
VehicleTyp	19,53
V_C_Ratio	0,49
Shoulder	1
Roads	BTS. KOTA NGANJUK - KERTOSONO
PRTA_C	1
Length_of_Road	20,63
knn4_pred	1
knn3_pred	1
knn2_pred	1
knn1_pred	1
knn0_pred	1
IRI	8,9
Horizontal	20,63
FID	116
Design_Spe	80
All_KNN_byVote	1

Identified 1 feature

111,194 -7,081 Decimal Degrees

(a)Arterial spatial datasets

Identify

Identify from: <Top-most layer>

Location: 113,170586 -8,197353 Decimal Degrees

Field	Value
Vertical_A	0
VehicleTyp	23,35
V_C_Ratio	0,2
Shoulder	1
Roads	BTS. KAB. MALANG - BTS. KOTA LUMAJANG
PRTA_C	1
Lenght_of_Road	53,35
knn4_pred	1
knn3_pred	1
knn2_pred	1
knn1_pred	1
knn0_pred	1
IRI	27,95
Horizontal	0
FID	48
Design_Spe	0
All_KNN_byVote	1

Identified 1 feature

114,055 -7,977 Decimal Degrees

(b)Collector spatial datasets

Figure 4. The results of the spatial analysis of the PRTA classification using the KNN method with 5-fold GA spatial cross-validation for arterial and collector datasets

5. Conclusion

The result of 48 experiments, the raw dataset produced the highest AUC on 11 experiments for 1st and 3rd scenarios. The models results using a dataset based on the 4th scenario are less superior than the 3rd scenario in the 50%-81% ROC-AUC score range. While the dataset that needs to be normalized only

Received: October 18, 2020. Revised: December 20, 2020.

produces the highest AUC in 3 experiments of 4th scenario. While the prediction model using 1st and 2nd scenario datasets all underperformed is 47-62% ROC-AUC score. Regarding balance datasets, the 11 experiments that produced the highest ROC-AUC, namely ten experiments, were in the 3rd scenario for SMOTE raw datasets. The results of the classifier with the SMOTE variant are very satisfying, especially for KNN-SMOTE with a ROC of almost 90% on the other metrics for more balanced accuracy, precision, recall, and f1score values. The conclusion is that to produce an optimal classifier, and the tendency is to use a balanced raw dataset using SMOTE.

Based on the results of the singular classifier experiment by adding hyper-parameter optimization using ga search cross-validation on each variant of SMOTE for the PRTA classification. The KNN method as a single algorithm for classification based on the distance between samples is superior to using datasets on K-Means SMOTE from all algorithms to handle small data sets in unbalanced classes with an average AUC value of 83%, including the good classification category. Where, Complex algorithms such as SVM and RF for the singular classifier rank 2nd with average AUC values of 80% and 79%, respectively, is a moderately classified category. The complex algorithms such as MLP, NB, and LR are not good enough in PRTA classification compared to other single algorithms such as KNN, SVM, and random forest. The overall conclusion from all experiments is that a simple, computationally light algorithm can produce a PRTA classification with a good classification category based on the condition that the dataset must be balanced using the SMOTE variant first. However, the overall performance of several compared algorithms has almost the same or close performance. In further research, it is necessary to conduct studies based on empirical studies to add other scenario models that can handle small dataset types that cause unbalanced classes using specific regions on different private datasets samples. The AUC value can increase the performance of the classification method in ML from 91 to 100% which is a very good classification category using ensemble learning in ML through bagging, boosting, and stacking model experiments.

Conflicts of Interest

The author's affirmation has stated that they have no conflicts of interest.

Author Contributions

Conceptualization of paper topics, A. V. Vitianingsih; research methodology, A. V. Vitianingsih; validation of research results, Z. Othman, S. S. K. Baharin, and A. Suraji; the formal analysis, A. V. Vitianingsih; the research investigation, A. V. Vitianingsih; the resources, A. V. Vitianingsih, Z. Othman, and S. S. K. Baharin; data curation, A. V. Vitianingsih, and A. Suraji; writing—original draft preparation, A. V. Vitianingsih; writing—review and editing, Z. Othman, and S. S. K. Baharin; visualization data and the research results, A. V. Vitianingsih; supervision, Z. Othman, and S. S. K. Baharin.

Acknowledgments

These study findings are part of a thesis at Universiti Teknikal Malaysia Melaka, Malaysia. Development of research results supported by the Indonesian Ministry of Research, Technology, and Higher Education in 2015-2016 and Universitas Dr. Soetomo.

References

[1] O. Tayan, A. M. Al BinAli, and M. N. Kabir, "Analytical and Computer Modelling of Transportation Systems for Traffic Bottleneck Resolution: A Hajj Case Study," *Arab. J. Sci. Eng.*, vol. 39, no. 10, pp. 7013–7037, 2014.

[2] G. Pauer, "Development potentials and strategic objectives of intelligent transport systems improving road safety," *Transp. Telecommun.*, vol. 18, no. 1, pp. 15–24, 2017.

[3] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, 2019.

[4] N. O. Alsrehin, A. F. Klaib, and A. Magableh, "Intelligent Transportation and Control Systems Using Data Mining and Machine Learning Techniques: A Comprehensive Study," *IEEE Access*, vol. 7, no. c, pp. 49830–49857, 2019.

[5] W. Chen *et al.*, "A novel fuzzy deep-learning approach to traffic flow prediction with uncertain spatial–temporal data features," *Futur. Gener. Comput. Syst.*, vol. 89, no. June, pp. 78–88, 2018.

[6] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Phys. A Stat. Mech. its Appl.*, vol. 517, no. March, pp. 29–35, 2019.

[7] S. Chen, Z. Wang, J. Liang, and X. Yuan, "Uncertainty-aware visual analytics for

Received: October 18, 2020. Revised: December 20, 2020.

exploring human behaviors from heterogeneous spatial temporal data," *J. Vis. Lang. Comput.*, vol. 48, no. September 2016, pp. 187–198, 2018.

[8] A. V. Vitianingsih, N. Suryana, and Z. Othman, "Spatial analysis model for traffic accident-prone roads classification: A proposed framework," *IAES Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 365–373, 2021.

[9] A. V. Vitianingsih, Z. Othman, S. Suhana, and K. Baharin, "Spatial Analysis for the Classification of Prone Roads Traffic Accidents: A Systematic Literature Review," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 2, pp. 583–599, 2021.

[10] K. Borowska and J. Stepaniuk, "A rough-granular approach to the imbalanced data classification problem," *Appl. Soft Comput. J.*, vol. 83, p. 105607, 2019.

[11] T. Osayomi, "Regional determinants of road traffic accidents in Nigeria: identifying risk areas in need of intervention," *African Geogr. Rev.*, vol. 32, no. 1, pp. 88–99, 2013.

[12] K. Van Raemdonck and C. Macharis, "The Road Accident Analyzer: A Tool to Identify High-Risk Road Locations," *J. Transp. Saf. Secur.*, vol. 6, no. 2, pp. 130–151, 2014.

[13] K. Geurts, I. Thomas, and G. Wets, "Understanding spatial concentrations of road accidents using frequent item sets," *Accid. Anal. Prev.*, vol. 37, no. 4, pp. 787–799, 2005.

[14] P. Xu and H. Huang, "Modeling crash spatial heterogeneity: Random parameter versus geographically weighting," *Accid. Anal. Prev.*, vol. 75, no. February, pp. 16–25, 2015.

[15] F. Torrieri and A. Batà, "Spatial multi-Criteria decision support system and strategic environmental assessment: A case study," *Buildings*, vol. 7, no. 4, 2017.

[16] C. Aubrecht, P. Meier, and H. Taubenböck, "Speeding up the clock in remote sensing: identifying the 'black spots' in exposure dynamics by capitalizing on the full spectrum of joint high spatial and temporal resolution," *Nat. Hazards*, vol. 86, no. 1, pp. 177–182, 2017.

[17] Á. Briz-Redón, F. Martínez-Ruiz, and F. Montes, "Spatial analysis of traffic accidents near and between road intersections in a directed linear network," *Accid. Anal. Prev.*, vol. 132, no. April, p. 105252, 2019.

[18] L. Ding, G. Xiao, D. Calvanese, and L. Meng, "Consistency assessment for open geodata integration: an ontology-based approach," *Geoinformatica*, vol. 25, no. 1, 2019.

[19] H. Al Majzoub, I. Elgedawy, Ö. Akaydın, and M. Köse Ulukök, "HCAB-SMOTE: A Hybrid Clustered Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification," *Arab. J. Sci. Eng.*, vol. 45, no. 4, pp. 3205–3222, 2020.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.

[21] A. Banerjee and S. Ray, "Spatial models and geographic information systems," *Encyclopedia of Ecology, 2nd Edition*. Elsevier Inc., pp. 1–10, 2018.

[22] L. Zhao, L. Chen, R. Ranjan, K. K. R. Choo, and J. He, "Geographical information system parallelization for spatial big data processing: a review," *Cluster Comput.*, vol. 19, no. 1, pp. 139–152, 2016.

[23] K. E. Brassel and R. Weibel, "A review and conceptual framework of automated map generalization," *Int. J. Geogr. Inf. Syst.*, vol. 2, no. 3, pp. 229–244, 1988.

[24] S. Wang, Y. Zhong, and E. Wang, "An integrated GIS platform architecture for spatiotemporal big data," *Futur. Gener. Comput. Syst.*, vol. 94, no. May, pp. 160–172, 2019.

[25] J. Lacasta, F. J. Lopez-Pellicer, B. Espejo-García, J. Nogueras-Iso, and F. J. Zarazaga-Soria, "Aggregation-based information retrieval system for geospatial data catalogs," *Int. J. Geogr. Inf. Sci.*, vol. 31, no. 8, pp. 1583–1605, 2017.

[26] D. Li, S. Wang, H. Yuan, and D. Li, "Software and applications of spatial data mining," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 3, pp. 84–114, 2016.

[27] N. F. Hordri, A. Samar, S. S. Yuhaniz, and S. M. Shamsuddin, "A systematic literature review on features of deep learning in big data analytics," *Int. J. Adv. Soft Comput. its Appl.*, vol. 9, no. 1, pp. 32–49, 2017.

[28] L. Li, B. Du, Y. Wang, L. Qin, and H. Tan, "Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model," *Knowledge-Based Syst.*, vol. 194, p. 105592, 2020.

[29] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020.

[30] Q.-J. K. and Y. L. Jianli Xiao, Xiang Gao, "More robust and better: a multiple kernel support vector machine ensemble approach for traffic incident detection," *J. Adv. Transp.*, vol. 48, no.

Received: October 18, 2020. Revised: December 20, 2020.

7, pp. 858–875, 2014.

[31] S. A. Manaf, N. Mustapha, M. N. Sulaiman, N. A. Husin, H. Z. M. Shafri, and M. N. Razali, "Hybridization of SLIC and extra tree for object based image analysis in extracting shoreline from medium resolution Satellite images," *Int. J. Intell. Eng. Syst.*, vol. 11, no. 1, 2018.

[32] Z. Cai, Y. Long, and L. Shao, "Classification complexity assessment for hyper-parameter optimization," *Pattern Recognit. Lett.*, vol. 125, no. July, pp. 396–403, 2019.

[33] X. Wang, X. Guan, J. Cao, N. Zhang, and H. Wu, "Forecast network-wide traffic states for multiple steps ahead: A deep learning approach considering dynamic non-local spatial correlation and non-stationary temporal dependency," *Transp. Res. Part C Emerg. Technol.*, vol. 119, no. June, p. 102763, 2020.

[34] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, no. November, pp. 295–316, 2020.

[35] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," *J. Transp. Saf. Secur.*, vol. 9, no. 2, pp. 146–166, 2017.

[36] N. Tran, J. G. Schneider, I. Weber, and A. K. Qin, "Hyper-parameter optimization in classification: To-do or not-to-do," *Pattern Recognit.*, vol. 103, no. July, pp. 1–12, 2020.

[37] W. Dai, D. Li, D. Tang, H. Wang, and Y. Peng, "Deep learning approach for defective spot welds classification using small and class-imbalanced datasets," *Neurocomputing*, vol. 477, pp. 46–60, 2022.

[38] F. A. Mousa and I. E. Fattoh, "Hybrid Chicken Swarm Optimization (CSO) and Fuzzy Logic (FL) Model for Handling Imbalanced Datasets," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 6, pp. 10–19, 2021.

[39] R. Goel, "Modelling of road traffic fatalities in India," *Accid. Anal. Prev.*, vol. 112, no. October, pp. 105–115, 2018.

[40] C. Han, H. Huang, J. Lee, and J. Wang, "Investigating varying effect of road-level factors on crash frequency across regions: A Bayesian hierarchical random parameter modeling approach," *Anal. Methods Accid. Res.*, vol. 20, no. Desember, pp. 81–91, 2018.

[41] S. Unhapipat, M. Tiensuwan, and N. Pal, "Bayesian Predictive Inference for Zero-Inflated Poisson (ZIP) Distribution with Applications," *Am. J. Math. Manag. Sci.*, vol. 37, no. 1, pp. 66–79, 2018.

[42] D. A. Anggoro and S. S. Mukti, "Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 6, pp. 198–207, 2021.

[43] M. R. Jabbarpour, H. Zarrabi, R. H. Khokhar, S. Shamshirband, and K. K. R. Choo, "Applications of computational intelligence in vehicle traffic congestion problem: a survey," *Soft Comput.*, vol. 22, no. 7, pp. 2299–2320, 2018.

[44] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Comput. Oper. Res.*, vol. 106, no. June, pp. 210–224, 2019.

[45] A. L. I. Oliveira, P. L. Braga, R. M. F. Lima, and M. L. Cornelio, "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation," *Inf. Softw. Technol.*, vol. 52, no. 11, pp. 1155–1166, 2010.

[46] J. Murillo-Morera, C. Castro-Herrera, J. Arroyo, and R. Fuentes-Fernández, "An automated defect prediction framework using genetic algorithms: A validation of empirical studies," *Iberamia*, vol. 19, no. 57, pp. 114–137, 2016.

[47] C. J. Burgess and M. Lefley, "Can genetic programming improve software effort estimation? A comparative evaluation," *Inf. Softw. Technol.*, vol. 43, no. 14, pp. 863–873, 2001.

[48] Florin Gorunescu, *Data Mining: Concept, Models, Techniques*. Springer, 2011.

[49] A. V. Vitianingsih, S. S. K. Baharin, O. Othman, and A. Suraji, "Empirical Study of a Spatial Analysis for Prone Road Traffic Accident Classification based on MCDM Method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 665–679, 2022.

[50] Direktorat Jenderal Bina Marga, "Highway Capacity Manual Project (HCM)," *Man. Kapasitas Jalan Indones.*, vol. 1, no. I, p. 564, 1997.

[51] S. Cost and S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features," *Mach. Learn.*, vol. 10, no. 1, pp. 57–78, 1993.

[52] M. Wasikowski, "Combating the Class Imbalance Problem in Small Sample Data Sets," pp. 1–97, 2009.

[53] S. He, H., Bai, Y., Garcia, E., & Li, "ADASYN:

Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008," in *IEEE World Congress on Computational Intelligence*, 2008, no. 3, pp. 1322– 1328.

[54] R. Malhotra and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," *Neurocomputing*, vol. 343, pp. 120–140, 2019.

[55] and B.-H. M. Hui Han, Wen-Yuan Wang, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *the Lecture Notes in Computer Science*, 2005, pp. 878 – 887.

[56] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018.

[57] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Saf. Sci.*, vol. 125, no. January, p. 104616, 2020.

[58] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Inf. Sci. (Ny).*, vol. 572, no. September, pp. 574–589, 2021.

[59] Y. Tang, Y. Q. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 1, pp. 281–288, 2009.

[60] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[61] C. F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci. (Ny).*, vol. 477, pp. 47–54, 2019.

[62] S. Nikbakht, C. Anitescu, and T. Rabczuk, "Optimizing the neural network hyperparameters utilizing genetic algorithm," *J. Zhejiang Univ. Sci. A*, vol. 22, no. 6, pp. 407–426, 2021.

[63] J. H. Han, D. J. Choi, S. U. Park, and S. K. Hong, "Hyperparameter Optimization for Multi-Layer Data Input Using Genetic Algorithm," *2020 IEEE 7th Int. Conf. Ind. Eng. Appl. ICIEA 2020*, pp. 701–704, 2020.

[64] J. H. Han, D. J. Choi, S. U. Park, and S. K. Hong, "Hyperparameter Optimization Using a Genetic Algorithm Considering Verification Time in a Convolutional Neural Network," *J. Electr. Eng. Technol.*, vol. 15, no. 2, pp. 721–726, 2020.

[65] M. A. Ferraciolli, F. F. Bocca, and L. H. A. Rodrigues, "Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models," *Comput. Electron. Agric.*, vol. 161, no. December 2017, pp. 233–240, 2019.

[66] M. Q. Bashabsheh, L. Abualigah, and M. Alshinwan, "Big Data Analysis Using Hybrid Meta-Heuristic Optimization Algorithm and MapReduce Framework BT - Integrating Meta-Heuristics and Machine Learning for Real-World Optimization Problems," in *Integrating Meta-Heuristics and Machine Learning for Real-World Optimization Problems*, E. H. Houssein, M. Abd Elaziz, D. Oliva, and L. Abualigah, Eds. Cham: Springer International Publishing, 2022, pp. 181–223.

[67] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecol. Modell.*, vol. 406, no. April 2018, pp. 109–120, 2019.

[68] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data," 2018.

IJIES-Application Of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets

ORIGINALITY REPORT

14%	11%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Negeri Surabaya The State University of Surabaya Student Paper	3%
2	ijai.iaescore.com Internet Source	2%
3	thesai.org Internet Source	1%
4	www.hindawi.com Internet Source	1%
5	www.igi-global.com Internet Source	<1%
6	oaji.net Internet Source	<1%
7	inass.org Internet Source	<1%
8	www.inass.org Internet Source	<1%

9	Zhaozhao Xu, Derong Shen, Tiezheng Nie, Yue Kou, Nan Yin, Xi Han. "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data", Information Sciences, 2021 Publication	<1 %
10	Anik Vega Vitianingsih, Nanna Suryana, Zahriah Othman. "Spatial analysis model for traffic accident-prone roads classification: a proposed framework", IAES International Journal of Artificial Intelligence (IJ-AI), 2021 Publication	<1 %
11	dokumen.pub Internet Source	<1 %
12	www.mdpi.com Internet Source	<1 %
13	www.acarindex.com Internet Source	<1 %
14	Burgess, C.J.. "Can genetic programming improve software effort estimation? A comparative evaluation", Information and Software Technology, 20011215 Publication	<1 %
15	Mengfei Wu, Ximing Li. "Unbalanced Data Classification Algorithm Based on Hybrid Sampling and Ensemble Learning", 2021 16th International Conference on Intelligent	<1 %

	Systems and Knowledge Engineering (ISKE), 2021 Publication	
16	academic-accelerator.com Internet Source	<1%
17	www.preprints.org Internet Source	<1%
18	Buket Geyik, Medine Kara. "Severity Prediction with Machine Learning Methods", 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020 Publication	<1%
19	D N S Putra, I N Yulita. "Multilayer Perceptron for Activity Recognition Using a Batteryless Wearable Sensor", IOP Conference Series: Earth and Environmental Science, 2019 Publication	<1%
20	Information Management & Computer Security, Volume 22, Issue 3 (2014-09-16) Publication	<1%
21	Oliveira, A.L.I.. "GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation", Information and Software Technology, 201011 Publication	<1%

22	Xinkai Yi, Yingying Xu, Qian Hu, Sujatha Krishnamoorthy, Wei Li, Zhenzhou Tang. "ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection", Complex & Intelligent Systems, 2022 Publication	<1%
23	aip.scitation.org Internet Source	<1%
24	di.ubi.pt Internet Source	<1%
25	"Encyclopedia of Big Data Technologies", Springer Science and Business Media LLC, 2019 Publication	<1%
26	He, Haibo, Sheng Chen, Hong Man, Sachi Desai, and Shafik Quoraishee. "", Unmanned/Unattended Sensors and Sensor Networks VII, 2010. Publication	<1%
27	Tuanfei Zhu, Yaping Lin, Yonghe Liu. "Improving interpolation-based oversampling for imbalanced data learning", Knowledge-Based Systems, 2019 Publication	<1%
28	Zhi-Hua Zhou. "Training cost-sensitive neural networks with methods addressing the class	<1%

	imbalance problem", IEEE Transactions on Knowledge and Data Engineering, 1/2006 Publication	
29	research.aimultiple.com Internet Source	<1 %
30	safetylit.org Internet Source	<1 %
31	"Information Technology and Systems", Springer Science and Business Media LLC, 2021 Publication	<1 %
32	"Intelligent Data Engineering and Automated Learning – IDEAL 2019", Springer Science and Business Media LLC, 2019 Publication	<1 %
33	Faiza Khan, Summrina Kanwal, Sultan Alamri, Bushra Mumtaz. "Hyper-Parameter Optimization of Classifiers, Using an Artificial Immune Network and Its Application to Software Bug Prediction", IEEE Access, 2020 Publication	<1 %
34	Maulida Ayu Fitriani, Dany Candra Febrianto. "Data Mining for Potential Customer Segmentation in the Marketing Bank Dataset", JUITA: Jurnal Informatika, 2021 Publication	<1 %

35	Sobhan Sarkar, Anima Pramanik, J. Maiti, Genserik Reniers. "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data", Safety Science, 2020 Publication	<1 %
36	Zhi Chen, Jiang Duan, Cheng Yang, Li Kang, Guoping Qiu. "SMLBoost-adopting a soft-margin like strategy in boosting", Knowledge-Based Systems, 2020 Publication	<1 %
37	downloads.hindawi.com Internet Source	<1 %
38	eprints.utm.my Internet Source	<1 %
39	link.springer.com Internet Source	<1 %
40	research.gold.ac.uk Internet Source	<1 %
41	www.springerprofessional.de Internet Source	<1 %
42	www.wwu.de Internet Source	<1 %
43	Lecture Notes in Computer Science, 2015. Publication	<1 %

44	"Data Engineering and Communication Technology", Springer Science and Business Media LLC, 2020 Publication	<1 %
45	"Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2019 Publication	<1 %
46	Anik Vega Vitianingsih, Zahriah Othman, Safiza Suhana Kamal Baharin, Aji Suraji. "Empirical Study of a Spatial Analysis for Prone Road Traffic Accident Classification based on MCDM Method", International Journal of Advanced Computer Science and Applications, 2022 Publication	<1 %
47	Wei Dai, Dayong Li, Ding Tang, Huamiao Wang, Yinghong Peng. "Deep learning approach for defective spot welds classification using small and class-imbalanced datasets", Neurocomputing, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

IJIES-Application Of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets

GRADEMARK REPORT

FINAL GRADE	GENERAL COMMENTS
/0	**Instructor**

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13