International Journal of
Intelligent Engineering & Systems

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[1,2]* Zahriah Othman[2] Safiza Suhana Kamal Baharin[2]
Aji Suraji[3] Anastasia Lidya Maukar[4]

[1]*Informatics Department, Universitas Dr. Soetomo, Surabaya, Indonesia*
[2]*Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia*
[3]*Department of Civil Engineering, University of Widyagama Malang, Malang, Indonesia*
[4]*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 over-sampling 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. The experiment was 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, with an AUC value of 0.89, which is included in the good classification category of all ML classifiers proposed to handle small data sets imbalanced classes using SMOTE raw datasets for K-Means variants SMOTE.

Keywords: Spatial analysis, Imbalance 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]. Research with good performance for PRTA classification has 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, 8] because 0.96 of the data uses private spatial datasets [9, 10]. The HSTA data types occur because of GIS characteristics used to solve specific region problems [11-17]. The term specific region can be interpreted as a linear network in spatial statistics [18]. The heterogeneous geospatial data results from the

DOI: 10.22266/ijies2022.1031.58

acquisition of data from several different government agencies, which causes the data to have many other formats with various structures [19]. The public ontology is used to resolve government-owned public data types to show inconsistencies in acquiring data sources [19].

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 imbalanced if the minority class reaches 0.40, 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 [20]. An imbalanced dataset will affect the classification sensitivity value in the minority class if it is not represented evenly [21, 22].

Many studies have handled the problem of road accident data imbalance using the variant SMOTE method, a popular method in the field of software engineering. The results of the research [23] use SMOTE, Borderline SMOTE, and SVM SMOTE methods. The research result in paper [24, 25] uses SMOTE, random over-sampling (ROS), and random under-sampling (RUS) method. Meanwhile, SMOTE-over-sampling and random under-sampling methods [26], and balanced bagging [27] were used to overcome imbalanced datasets obtained in real-time (time/hour). The random under-sampling, SMOTE over-sampling, and mixed technique methods overcome imbalanced data [28].

The process of constructing an Artificial Intelligence (AI) model and combining it with experiments on spatial datasets is known as spatial analysis modeling [29]. It was collecting spatial knowledge through spatial datasets and supplying knowledge of models used in the framework through the use of artificial intelligence approaches based on machine learning models from various sources. In GIS, spatial datasets take on the role of the fundamental framework for developing spatial analysis algorithms, investigating algorithmic principles, or modifying pre-existing algorithms. [30]. The objective of the spatial analysis model is to describe the GIS software that will be produced and to conduct simulations to put 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 and how the data are processed through spatial analysis to become information that can be used in a decision support system [31]. 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 [32] as well as an information retrieval system that is based on an aggregation of spatial datasets [33].

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 [34, 32]. In the discipline of machine learning, the categorization model is widely used [35] to be used to research in the field of geographic information system (GIS) spatial analysis. However, since 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, it is dependent on the field of study, which is never the same as the subject of the research carried out.

Over the last three decades, many ML techniques have been proposed to improve GIS-Spatial accuracy for the PRTA classification on imbalanced dataset types [23, 28]. The study literature for the PRTA classification in terms of handling imbalanced class in small spatial datasets through the synthetic over-sampling method that integrates with hyperparameters through accuracy testing the ML method in the highway safety domain with testing data on road accidents. 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 [36]. The ML integrates various algorithms with combined machine learning models to complete tasks in the data mining field, including their classification, clustering, prediction, etc. [37] to improve the robustness [38, 4]. The methods for the single ML classifiers are widely used to determine the effectiveness of the proposed model, including LR, DT, RF, and AdaBoost [24]; RF, NB, KNN, and ANNs [25]; Bayes classifiers [28]; binary classifiers [26, 27]; SVMs and Probabilistic Neural Network (PNN) [23]; 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 [39].

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 [40]. The choice of tuning parameter technique in the ML model used has an essential role in determining the resulting sensitivity analysis [41] and validating the increased best performance model [42, 43]. A hyper-parameter optimization technique is a process of tuning parameters that is suitable for training on the ML model [44]. The classification technique in the ML model has a complexity assessment level for hyper-parameter optimization [40]. 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 [44].

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 imbalanced 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 0.89 to all single competition algorithms to handle small datasets in pre-processing data using 3rd scenario and imbalanced datasets process using KMeans-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 methods 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 following 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

This section will review several previous studies on imbalanced data techniques on small road accident datasets to create new synthetic data, the PRTA classification method approach used to test the best-imbalanced data techniques, and hyperparameter tuning methods to improve the performance of the classification method.

In the paper [23], the researchers convey the results using the SMOTE method to overcome the imbalanced dataset that caused the number of accidents in the dataset to be insufficient (small dataset category). The dataset was tested using the ML classifier method, namely SVM and PNN; the results of the tests stated that the PNN method was superior to SVM, with AUC values of 0.90 [23]. The weakness of the results of this study [23] is that there is no data pre-processing process to select the best data to be tested on the ML classifier; this is because the dataset used is based on real-time traffic condition data (depending on weather conditions, accident, and loop detector data).

The authors of [25] develop a sampling technique scenario of RAW, RUS, ROS, and SMOTE to overcome the imbalance road accidents dataset with hyperparameter optimization using two techniques, namely random hold-out, and 5-fold CV, which were tested on the ML classifier RF, NB, KNN, and ANN. The AUC values were 0.83, 0.68, 0.76, and 0.78, respectively, based on the RUS, RUS, ROS, and ROS sampling techniques. The weakness of this study is that high accuracy values will be achieved if there are many features of the accident data extracted.

Researchers in the paper [26] proposed the SMOTE-over-sampling and random under-sampling method to handle imbalanced data based on a time-dependent accident event dataset (hours); the ML binary classifier NB, LR, MARS, and RF methods were used to test the dataset with AUC values of 0.50, 0.66, 0.65, 0.65, respectively. The limitation of this study is that the variables used to depend on the data set that displays hourly values on the determinants associated with the incidence of PRTA dataset; this creates uncertainty in the results due to increased missing information because it does not consider the Spatio-temporal so that the AUC value low [26].

In [27], researchers used a generative adversarial networks (GANs) model to overcome small imbalanced datasets (i.e., traffic incidents datasets, including traffic flow volume, traffic speed, and building), which were tested on the SVM ML classifier with an accuracy value of 0.89. The weakness of this study's results is the dataset's behavior that will affect the performance of the ML classifier, where training data samples affect the overall detection performance [27].

The results of the study researchers [28] used RUS, SMOTE over-sampling, and mix technique methods to overcome imbalanced data, the results of the study [28] stated that the over-sampling technique with test data processed using the Bayes classifier between NB and Bayesian networks improved the performance of the proposed model with an AUC value of 0.68, while for the original pre-processing dataset, the AUC value is 0.64, and the mix-model is 0.66.

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 [44, 42]. The researchers [45] used the Hyper-parameter technique to determine the model's sensitivity 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 got 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 a small sample size and is only suitable for hierarchical model structures [46, 47]. The most widely used hyper-parameter methods are random search, grid search, and manual search. However, this method is computationally impractical [40]. The GA [48] and PSO methods are popular methods used for hyper-parameter techniques [42, 49, 50].

Based on the literature study conducted, some of the models produced experienced several shortcomings and, for now, have not made a satisfactory accuracy value. This study will propose a scenario model to overcome imbalanced 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 imbalanced class classifications in spatial datasets for attribute data categories.

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

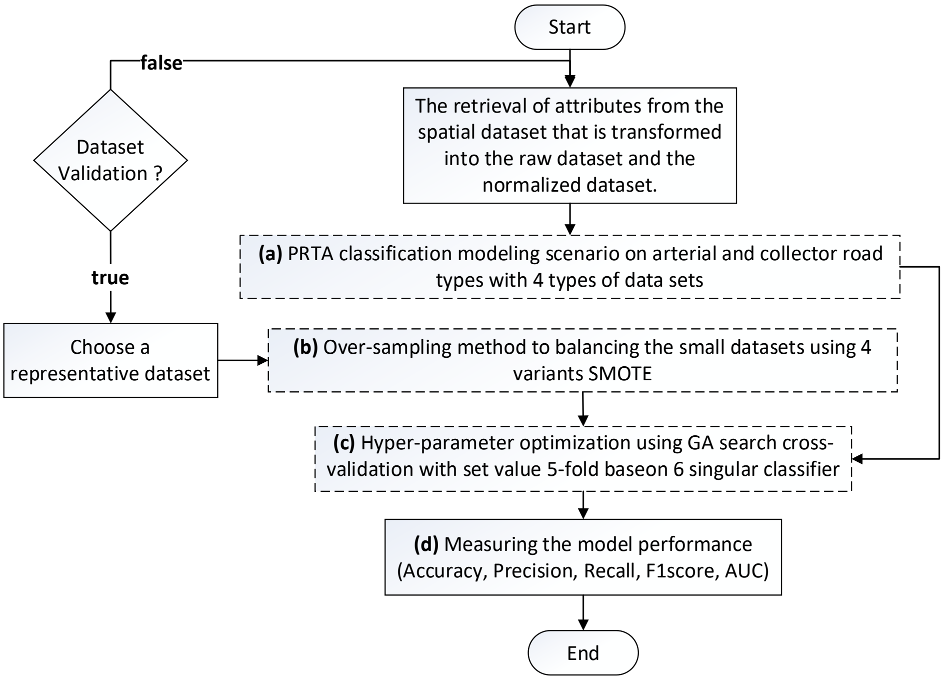

Start
false
The retrieval of attributes from the spatial dataset that is transformed into the raw dataset and the normalized dataset.
Dataset Validation ?
true
(a) PRTA classification modeling scenario on arterial and collector road types with 4 types of data sets
Choose a representative dataset
(b) 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

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 [51, 52]. 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 [53].

(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% (very good), 81-90% (good), 71-80%(fair), and 61-70% (poor), and values below 60% are considered to be false classifications [54].

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 Centre 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 [55, 56].

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 [21] 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) [57]. The data set points from the SMOTE method are placed at any point on the extrapolation line [58].

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

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 (section 3.1), 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. The value of $\delta(x_i, y_i)$ is the distance between X and Y vector observations on each explanatory variable based on Eq. (2) [57].

$$\delta(x_i, y_i) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k \quad (2)$$

Where variable n is the number of class categories in the 1st variable, variable C_1 is the number of values of the 1st category in each x_i occurs, variable C_{1_i} is the number of the 1st category in each x_i which belongs to the i-th class, variable C_2 is the number of values in the 2nd category in each y_i that occurs, variable C_{2_i} is the number of the 2nd category in each y_i which belongs to the i-th class, variable k is a constant value usually set to 1.

3.2.2. ADASYN (adaptive synthetic sampling) resampled

ADASYN is a method for balancing data by approaching it through sampling from imbalanced datasets [59]. 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 [59] using Eq. (3) 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 [60].

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \quad (3)$$

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 [61]. The synthetic data generator is only carried out in the example borderline [61] to generate a new instance using Eq. (4) by measuring between borderline instances and minority instants using k-nearest neighbors [20].

$$New\ instace = P_i + gap * (distance\ (P_i, P_j)) \quad (4)$$

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

K-Means SMOTE is an oversampling technique that handles imbalanced classes, consisting of clustering, filtering, and oversampling [62]. Identifying locations in the input space generates synthetic data [63] 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 [64].

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

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 imbalanced 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) [65].

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 [66, 67]. 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 proposed model for using data on a small imbalanced dataset type was tested using a scenario model at the pre-processing data stage, with the results in Fig. 2 and 3. These results show the performance of scenarios 1 to 4 tested on the ML classifier method, including KNN, SVM, MLP, NB, LR, and RF. Scenario 3 is superior to all the ML classifiers, the details can be found in Table 1 to 4, where the mean AUC values for scenarios 1 to 4 are 0.55, 0.53, 0.70, and 0.67, respectively. The ML classifier hyperparameter optimization will be processed using GA search cross-validation, and this table will be part of Section 4.2.

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

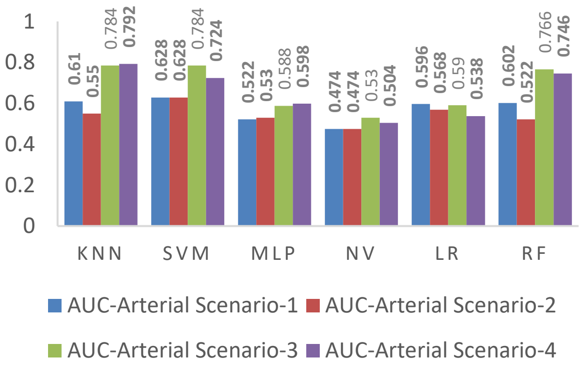


Figure. 2 Scenarios for prediction modeling PRTA on arterial road datasets

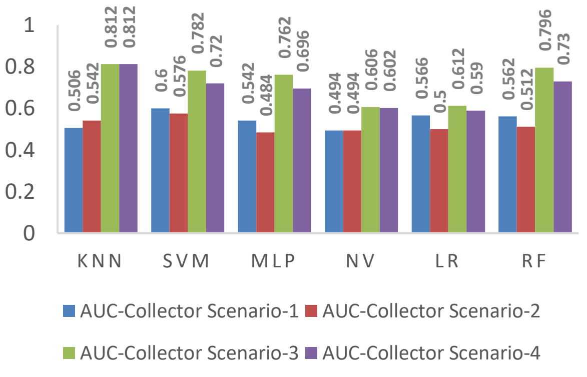


Figure. 3 Scenarios for prediction modeling PRTA on collector road datasets

The results in Table 1 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 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

Table 1. 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*
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 2. 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

Table 3. 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

produced the lowest AUC values, which came in at 47.4% and 48.4%, respectively.

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 3. 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.

The experimental results in scenario 4 using the normalized min-max dataset type processed with the smote algorithm can be seen in Table 4. 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 4. 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

In this section, the impact of using a new synthetic dataset in the field of road safety will be explained which was obtained at the pre-processing stage of the data through a four-scenario approach to be tested on the ML classifier. Test data based on parameters in section 3.1 [55, 56] with experimental results using a data set selected at the pre-processing data stage. Table 5 shows the best scenario model proposed to validate the data set to improve the data quality. Table 5 uses synthetic data generated by several developments of synthetic data generated using the variant SMOTE over-sampling method, including ADASYN, Borderline SMOTE, K-Means-SMOTE, and SVM-SMOTE will be combined with the raw dataset to determine the effectiveness of the performance of ML classifier hyper-parameter optimization using GA search cross-validation. The GA is a metaheuristic optimization method that has been developed in several domains [68-70]. The best ROC-AUC value is declared in bold, whereas the worst value is declared in italics underline.

The experimental results in Table 5 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%. The NB 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 Naïve Bayes method obtained the lowest values of 50% and 50%. Whereas for some other algorithms, the AUC value is almost the same.

Table 5. The result of balancing datasets and ML classifier hyper-parameter optimization using GA search cross-validation

Imbalance Data Method	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)

Based on the literature study in Section 2, the experimental results in Table 5 refer to the research development [25]. The suggestions given for future work are to observe and apply the performance of the Nominal and Continuous-SMOTE (SMOTE-NC) oversampling techniques, Borderline-SMOTE (BL-SMOTE), K-Means-SMOTE (KM-SMOTE), and SVM- SMOTE as an effort to improve the performance of the ML classifier by testing the accident dataset, and in this case study we apply it to the road traffic accident dataset.

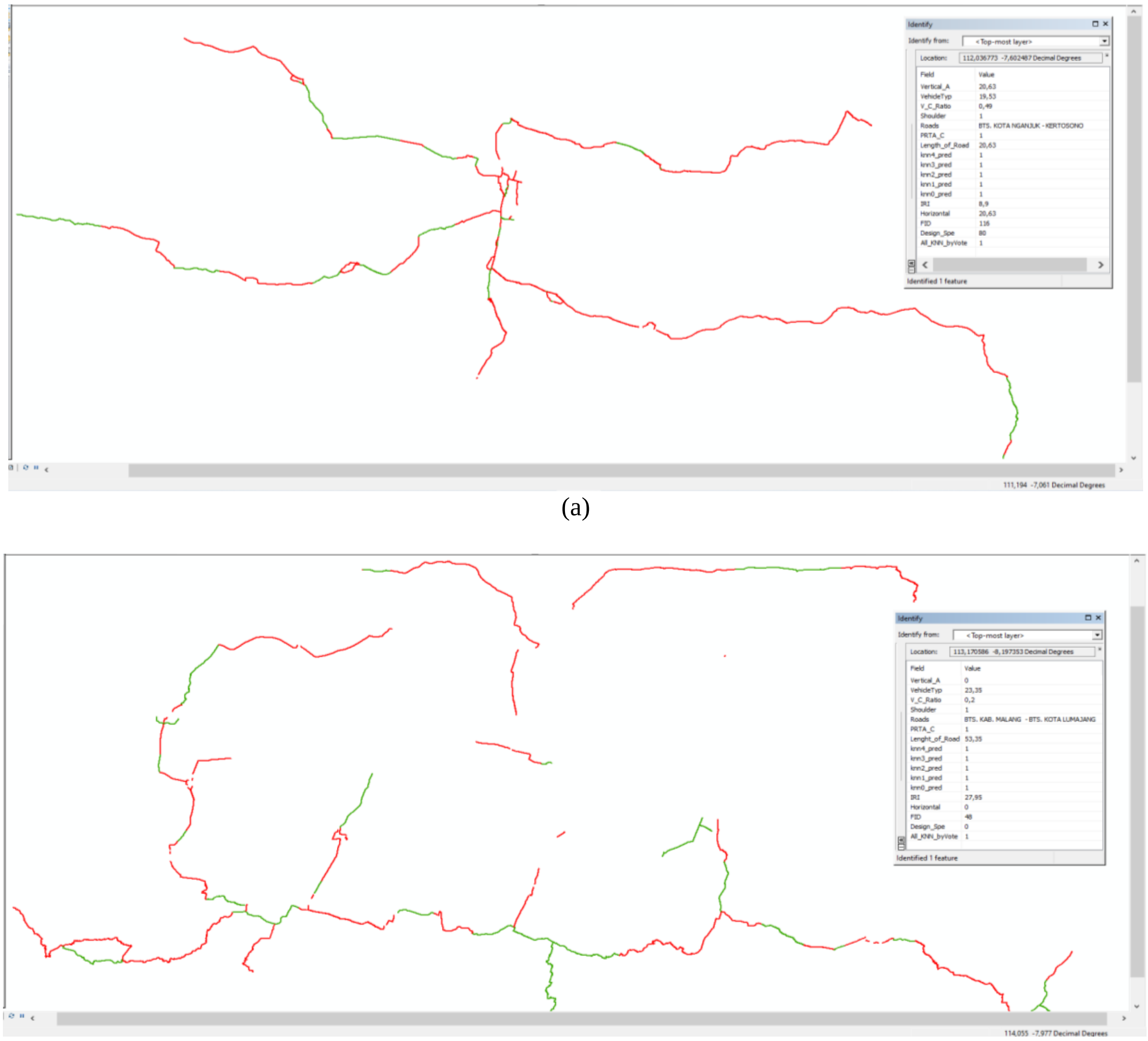
The overall experimental results for dealing with the imbalance of small spatial datasets that use traffic road accidents as datasets show that KNN-KM-SMOTE is the method that has the highest performance in improving the performance of the ML classifier with an AUC value rating of 0.89 (good classification [56]). The SVM-KM-SMOTE, KNN-BL-SMOTE, NB-KM-SMOTE, MLP-KM-SMOTE, LR-KM-SMOTE, and RF-ADASYN-SVM-SMOTE methods are 0.88, 0.82, 0.86, 0.84, 0.84, 0.83 and 0.79, respectively. 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 [71], even though KNN is considered a poor test on the IRIS dataset [72].

The performance of the ML classifier for the AUC value of KNN-KM-SMOTE (0.89), and KNN-BL-SMOTE (0.82) is superior to the AUC value of KNN-RUS 0.68 [25]. The AUC value of RF-ADASYN-SVM-SMOTE (0.79) decreased by 0.04 from the AUC value of 0.83 of RF-RUS [25]. In comparison, the AUC value of 0.86 for NB-KM-SMOTE is better than the AUC NB-ROS value, which only reaches an AUC value of 0.76 [25]. The ML classifier and oversampling method SVM-KM-SMOTE with an AUC value of 0.88 are better than the SVM-SMOTE, which only reaches the AUC value of 0.74 [23], while with the ML classifier SVM and GANs oversampling method (SVM-GANs) the difference is 0.01 for the AUC value of 0.89 [27]. The ML classifier method with a combination of oversampling methods for NB-KM-SMOTE, LR-KM-SMOTE, RF-ADASYN-SVM-SMOTE with AUC values of 0.86, 0.83, and 0.79, respectively, is far superior to the results of research [26] for NB, LR, RF-SMOTE-maximum dissimilarity sampling AUC values are 0.50, 0.66, and 0.65, respectively.

4.3 Spatial cross-validation

A cross-validation is an approach in statistical methods that works to evaluate the performance of

Received: June 13, 2022. Revised: August 3, 2022.



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

(a)

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

(b)

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: (a) Arterial spatial datasets and (b) Collector spatial datasets

the model that is built, and this can reduce the bias of the results obtained from the model's performance [73]. The process of cross-validation will divide the data into two parts which are used for learning and validation. The purpose of cross-validation in hyperparameter tuning of spatial data, spatial partitioning can be utilized [74].

Fig. 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 Fig. 4 (a) and collector datasets in Fig. 4 (b) using the KNN method, which is a good classifier category based on Table 5 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 sections 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%. Comparison of the size of the dataset rate in previous studies, including papers with as many as 85.182 datasets [23], 120.277 datasets [25], 16.728 datasets [26], and 1.560 datasets [27].

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 0.50% to 0.81 ROC-AUC score range. While the dataset that needs to be normalized only produces the highest AUC in 3 experiments of the 4th scenario. While the prediction model using 1st and 2nd scenario datasets all underperformed is 0.47 to 0.62 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 0.90 on the other metrics for more balanced accuracy, precision, recall, and F1score values. The conclusion is that to produce an optimal classifier, 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 ML classifier based on the distance between samples is superior to using the dataset in the 3rd scenario and using K-Means SMOTE oversampling technique (KNN-KM-SMOTE) from all algorithms to handle small data sets in unbalanced classes with an AUC value is 0.89, including the good classification category based on experiments using the traffic accident dataset. Complex algorithms on the ML classifier SVM, NB, MLP, KNN, and LR to rank 2nd with average AUC values of 0.88, 0.86, 0.86, 0.84, 0.82, and 0.83, respectively, is a good ML classification category. The RF classifier gets the smallest AUC value, which is 0.79 is a fair ML classification category. 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 (3rd scenario).

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 imbalanced 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 ML ensemble learning 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; accuration spatial datasets, 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; spatial and attribute data collector, A. V. Vitianingsih, and A. L. Maukar.

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. A. B. Ali, 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, "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 exploring human behaviors from heterogeneous spatial temporal data", *J. Vis. Lang. Comput.*, Vol. 48, No. September 2016, pp. 187-198, 2018.

[8] V. Sharmila, "Multi-Class Arrhythmia Detection using a Hybrid Spatial-Temporal Feature Extraction Method and Stacked Auto Encoder", *Int. J. Intell. Eng. Syst.*, Vol. 14, No. 2, pp. 82-94, 2021, doi: 10.22266/ijies2021.0430.08.

[9] 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.

[10] 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.

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

[12] 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.

[13] K. V. 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] Á. B. Redón, F. M. 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.

[19] 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.

[20] H. A. Majzoub, I. Elgedawy, Ö. Akaydın, and M. K. 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.

[21] 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. September, pp. 321-357, 2002.

[22] Y. Liu, W. Han, X. Wang, and Q. Li, "Oversampling algorithm based on spatial distribution of data sets for imbalance learning", In: *Proc. of 2020 5th Int. Conf. Comput. Commun. Syst.*, pp. 45-49, 2020.

[23] A. B. Parsa, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Real-time accident detection: Coping with imbalanced data", *Accid. Anal. Prev.*, Vol. 129, No. January, pp. 202-210, 2019.

[24] J. Choi, B. Gu, S. Chin, and J. S. Lee, "Machine learning predictive model based on national data for fatal accidents of construction workers", *Autom. Constr.*, Vol. 110, No. September, pp. 1-14, 2020.

[25] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, "Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods", *Eng. Constr. Archit. Manag.*, 2022.

[26] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset", *Accid. Anal. Prev.*, Vol. 127, No. February, pp. 134-149, 2019.

[27] M. Schlögl, "A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach", *Accid. Anal. Prev.*, Vol. 136, No. March, pp. 1-12, 2020.

[28] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets", *Accid. Anal. Prev.*, Vol. 88, No. March, pp. 37-51, 2016.

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

[30] 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.

[31] 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.

[32] 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.

[33] J. Lacasta, F. J. L. Pellicer, B. E. García, J. N. Iso, and F. J. Z. Soria, "Aggregation-based information retrieval system for geospatial data catalogs", *Int. J. Geogr. Inf. Sci.*, Vol. 31, No. 8, pp. 1583-1605, 2017.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Q. J. K. and Y. L. J. Xiao, and X. Gao, "More robust and better: a multiple kernel support vector machine ensemble approach for traffic incident detection", *J. Adv. Transp.*, Vol. 48, No. 7, pp. 858-875, 2014.

[39] 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, doi: 10.22266/ijies2018.0228.07.

[40] 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.

[41] 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.

[42] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice", *Neurocomputing*, Vol. 415, No. November, pp. 295-316, 2020.

[43] 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.

[44] 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.

[45] R. Goel, "Modelling of road traffic fatalities in India", *Accid. Anal. Prev.*, Vol. 112, No. October, pp. 105-115, 2018.

[46] 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.

[47] 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.

[48] 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, doi: 10.22266/ijies2021.1231.19.

[49] 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.

[50] 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.

[51] 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.

[52] J. M. Morera, C. C. Herrera, J. Arroyo, and R. F.

Fernández, "An automated defect prediction framework using genetic algorithms: A validation of empirical studies", *Iberamia*, Vol. 19, No. 57, pp. 114-137, 2016.

[53] 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.

[54] F. Gorunescu, *Data Mining: Concept, Models, Techniques*, 2011.

[55] 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.

[56] D. J. B. Marga, "Highway Capacity Manual Project (HCM)", *Man. Kapasitas Jalan Indones.*, Vol. 1, No. I, p. 564, 1997.

[57] 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.

[58] M. Wasikowski, "Combating the Class Imbalance Problem in Small Sample Data Sets", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10, pp. 1388-1400, 2010, doi: 10.1109/TKDE.2009.187.

[59] S. He, H. Bai, Y. Garcia, and E. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008", In: *Proc. of IEEE World Congress on Computational Intelligence*, 2008, No. 3, pp. 1322-1328.

[60] 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.

[61] B. H. M. H. Han, W. Y. Wang, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", *in the Lecture Notes in Computer Science*, pp. 878-887, 2005.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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, No. November, pp. 113-141, 2013.

[67] 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.

[68] 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.

[69] 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*, pp. 701-704, 2020.

[70] 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.

[71] 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.

[72] 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, pp. 181-223, 2022.

[73] 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.

[74] 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", *arXiv:1803.11266*, 2018.