

# Weighted nearest neighbors and radius oversampling for imbalanced data classification

Gede Angga Pradipta<sup>1</sup>, Putu Desiana Wulaning Ayu<sup>1</sup>, Made Liandana<sup>2</sup>, Dandy Pramana Hostiadi<sup>1</sup>

<sup>1</sup>Department of Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia

<sup>2</sup>Department of Informatics and Computer, Institut Teknologi dan Bisnis STIKOM Bali, Denpasar, Indonesia

## Article Info

### Article history:

Received Aug 19, 2023

Revised Mar 7, 2024

Accepted Aug 30, 2024

### Keywords:

High dimensional data

Imbalanced data

K-nearest neighbors

Synthetic minority

oversampling technique

Weighted feature

## ABSTRACT

The challenges associated with high-dimensional and imbalanced datasets were observed to often lead to a degradation in the performance of classical machine learning algorithms. In the case of high dimensional data, not all features contribute significantly and are considered relevant to the performance of the model. Therefore, this study introduced a novel method called feature weighted variance analysis-nearest neighbors (WFVANN) which was developed on the foundation of k-nearest neighbors (KNN). The process involved modifying the calculation of the Euclidean distance by fully considering the relevance and contribution levels of features based on their F-value. WFVANN at the algorithmic level processing and radius-synthetic minority oversampling technique (R-SMOTE) at the data level processing used as the oversampling method later became the proposed model to solve the aforementioned issues. Moreover, extensive experiments were conducted on two distinct types of data including the high-dimensional and imbalanced by comparing WFVANN with the state-of-art KNN-based and synthetic minority oversampling technique (SMOTE)-based methods. The results showed that the proposed method had the highest accuracy, precision, recall, and F1-measure values across the majority of test datasets and outperformed the other methods.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Gede Angga Pradipta

Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM

Bali, Indonesia

Email: angga\_pradipta@stikom-bali.ac.id

## 1. INTRODUCTION

The dataset with significant disparity during the process of distributing data into different classes is technically stated to be imbalanced. This significant disparity, often extreme, exists among classes or labels of data cases, thereby making imbalanced datasets common in real-world applications and concrete fields such as medical diagnosis [1]-[3], fault diagnosis [4], [5], anomaly detection [6]-[8], intrusion detection [9]-[11], and several others. Meanwhile, the minority class holds a higher level of importance and interest during the recognition process by machine learning models. For instance, identifying and recognizing patterns of rare diseases in medical diagnosis is crucial but the actual data count for normal conditions far outweighs those linked to the diseases. A critical challenge encountered when dealing with class imbalance during learning is the failure of most standard machine learning algorithms to project accurate boundary lines for each class within the dataset in some cases. This is because machine learning predominantly learns patterns from the majority class, introducing bias toward the minority class and leading to class overlapping. This overlapping, also referred to as class complexity or separability, signifies the degree of separation between classes in the data. Consequently, standard machine learning algorithms

struggle to define and determine discriminative rules for class separation. This overlapping feature space leads to the loss of intrinsic properties within the data, rendering it redundant or irrelevant in the process of recognizing good decision boundaries between classes.

Several solutions are discovered to have been proposed to address this issue over time and those applied in previous studies can be categorized into three groups including data sampling, algorithmic modification, and cost-sensitive learning [12]–[14]. This is because some studies found methods by preprocessing the data, particularly by resampling the minority data to alter the class distribution and tackle imbalanced datasets. One of the most widely used methods is oversampling and this involves creating a superset of the original dataset by replicating some instances or developing new instances from existing ones. Studies on data oversampling widely used the synthetic minority oversampling technique (SMOTE) [15]. The main idea generally associated with this method is the creation of new examples for the minority class by interpolating several instances from the class. However, SMOTE has some drawbacks despite its ability to improve the distribution of examples in each class. One of the drawbacks is related to blind oversampling which involves focusing only on the information from its nearest data or nearest positive example without considering the spatial information of the neighbors [16]. This usually results in several newly generated data points falling into the areas of the negative or minority class, leading to the creation of noisy data and disruption in the inter-class areas within the dataset. Furthermore, the overlapping feature space causes the features to lose their intrinsic property, leading to redundancy or irrelevance in recognizing good decision boundaries between classes.

One category of solutions developed to address these challenges focuses on the type of interpolation used and the determination of the regions in which new data are formed using SMOTE method. The interpolation mechanism can take various forms such as the range-restricted which involves considering the information of both the nearest positive and negative neighbors. Moreover, some studies used multiple interpolations [17], involving more than two examples or following topologies based on geometric shapes such as ellipses [18], Voronoi diagrams [19], and graphs [20]. Several studies [21]–[23] also applied clustering-based interpolation with each new example limited to being formed in the same cluster area from the sample point in addition to the combination with the clustering method.

The determination of the distance between a positive sample point and its nearest neighbors for interpolation in SMOTE method is based on the calculation of Euclidean distance. Several SMOTE developments also concentrate on appropriate data sampling to reduce the occurrence of overlapping regions and prevent the generation of noisy new data. To identify the best data samples, several methods have been used to select candidate samples. This was indicated in previous studies where samples were categorized into safe and dangerous zones [24] selected border regions [25], and determined difficulty weights for each instance [26], [27]. The two exceptions identified were the generation of synthetic examples after a learning vector quantization (LVQ) optimization process [28] and the selection of initial points from the support vectors obtained by an SVM [29]. In general, the distance calculations used in both traditional SMOTE and its developments were observed to rely on  $k$ -nearest neighbors (KNN) algorithm used in selecting data samples. However, the use of KNN also has several shortcomings such as the sensitivity to the neighborhood size  $k$  [30] and the distance function applied to select KNN. The identification of the most suitable distance formula for all training samples was found to be a challenging exercise. KNN also has high complexity due to the need to search for nearest neighbors and is considered less effective for imbalanced class datasets.

Most SMOTE methods were continuously being developed based on traditional KNN method, thereby leading to the persistence of the limitations. This led previous studies to propose improvements to KNN method, particularly focusing on the issue of sensitivity to the  $k$  value. The local mean factor was applied to mitigate the effects of  $k$  sensitivity while several other methods including  $k$ -harmonic nearest neighbors (KHNN) [31], local mean-based  $k$ -nearest neighbors (LMKNN) [32], local mean-based pseudo nearest neighbors (LPMNN) [33], and multi-local means-based nearest neighbors (MLNN) [34] concentrated on reducing outliers in the vicinity of the sample points. Several improvements were also made by assigning weights to each data point within the neighborhood such as pseudo nearest neighbors (PNN) [35], weighted representation-based  $k$ -nearest neighbors (WRKNN), and weighted local mean representation-based  $k$ -nearest neighbors (WLMRKNN) [30]. The development of these weighting methods was based on the observation that each nearest neighbor often contributes differently to the classification outcome in real-world problem data. KNN method was observed to have been developed based on the distance and position of nearest neighbors while considering the weights of each data point. The purpose was to account for potential outliers in the surrounding area without considering the contribution of each feature in determining nearest neighbors distance for each data point. The basic and most common distance measurement in KNN was found to be typically performed using Euclidean distance calculation. However, there was the possibility of each feature having varying contributions to the classification outcome. The proximity of each data point was likely to be influenced by several features contributing to and correlating with the class label. Therefore, this study proposed a new method based on KNN called the feature weighted variance analysis-nearest neighbors

(WVANN). The process involved modifying the distance calculation of data points by adding a weighting feature to the existing data features. The weight values depended on the correlation and contribution level of each feature. Moreover, the F-value obtained through analysis of variance (ANOVA) method was used to compute the feature contribution values and further combined with radius-synthetic minority oversampling technique (R-SMOTE) modification, an oversampling method, to solve the imbalanced data problem. WVANN applied at the algorithmic level and R-SMOTE at data level processing were later designed as the proposed model to solve the challenges associated with building a robust machine learning mode in two data conditions, including high dimensional and imbalanced.

## 2. METHOD

### 2.1. Weighted feature variance analysis-nearest neighbors

The algorithm modified to calculate the Euclidean distance in KNN was observed to rely on the F-value obtained from the feature selection process using ANOVA. The biggest challenge in machine learning was the selection of the best features to train the model. Therefore, this study aimed to select the features considered highly dependent on the response variable. This was because the variance of a feature usually determines the level of its impact on the response variable based on the criterion that a low variance indicates a lack of impact and vice versa. ANOVA was defined as a statistical method normally used to check the means of two or more groups that are significantly different from each other. Similarly, in KNN method, the calculation of proximity between sample points and their nearest neighbors should be heavily influenced by features selected based on high relevance to the dependent variable. A higher F-value for a feature usually signifies a greater weight assigned in determining the Euclidean distance. It was also noted that not all features possess significance in shaping the decision boundary between classes in the case of high dimensional data. Some irrelevant features used in data pattern determination could reduce the performance of the machine learning model developed. Therefore, the proposed WVANN method assigned weights to each feature based on their respective variance analyses as indicated in the flow process presented in Figure 1.

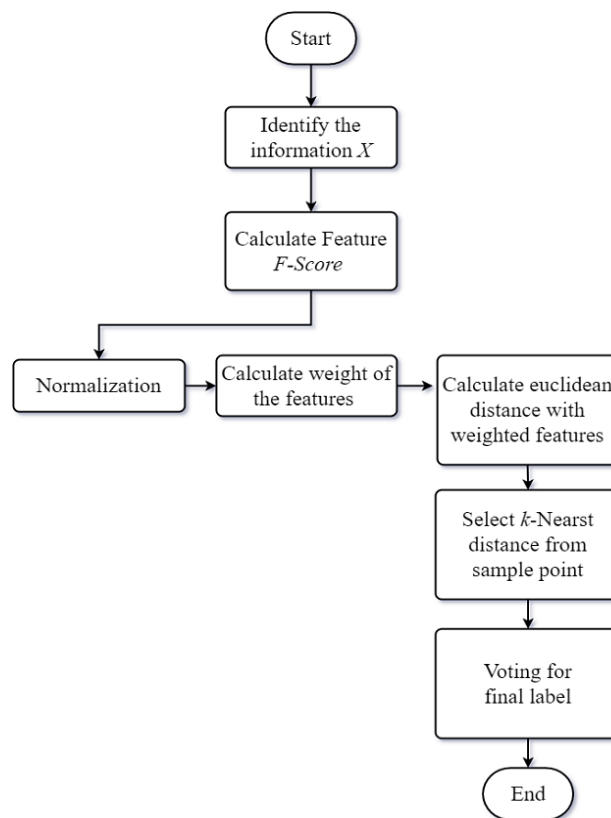


Figure 1. The flow of WVANN process

The weights were obtained from the F-value determined using ANOVA in the feature selection method. The features with a high F-value were assigned proportionally higher weights while those with a low or even zero F-value had no impact on the distance calculation in KNN. Moreover, WFFVANN algorithm was formalized in Algorithm 1 and the weighting for each feature in the data was determined using ANOVA by calculating the F-value for each feature. ANOVA ranked the features by calculating the variance ratio between and within groups. Furthermore, the F-value was computed by finding the ratio of mean square between (MSB) to mean square within (MSW). In Step 1, the variance value for each group or label was calculated. For each feature  $i$  within a label, its average value ( $\bar{X}_i$ ) was determined and subtracted from the total average value of the feature ( $\bar{X}$ ). The subtraction result was later multiplied by the number of labels in the data, denoted as  $k$ . The result was subsequently divided by the degree of freedom for MSB, represented as  $k - 1$ . Step 2 involved calculating the MSW value. Each data point within label  $i$  ( $X_{ij}$ ) was subtracted from the average value of the label ( $\bar{X}_i$ ) and the result was divided by the degree of freedom for MSW, represented as  $K - k$ . Step 3 was used to divide MSB by MSW to obtain the F-value for feature  $i$ . In steps 4-7, the resulting F-value was normalized using the Min-Max normalization formula. The method was used to scale the feature values to a range between 0 and 1 and this was achieved by subtracting the minimum value of the feature from each value and dividing by the range of the feature weighted. Subsequently, the distance for the new data point was calculated against all training data points  $X$  using the Euclidean distance formula by adding the effect of weight on each feature. Finally, the process continued in steps 8-11 by sorting with nearest neighbors value from the calculation results.

Algorithm 1: Weighted feature by variance analysis algorithm using F-value

Input:

$X$ : training data,  $Y$ : label of  $x$ ,  $m$ : Number Nearest Neighbors ( $a_1, a_2, \dots, a_m$ ),  $X_{ij}$ : index  $j$  in group class  $I$ ,  $X_i$ : instances index  $I$ ,  $\bar{X}_i$ : mean within-group class,  $\bar{X}$ : mean all data,  $D_i$ : number of class,  $f$ : feature in row data,  $n_i$ : number of class,  $n$ : number of features,  $K$  = count of all data, and  $k$  = count of groups.

Output: Labels Class of  $x$  Samples

1. Calculating mean square between (MSB):  

$$\left( \sum_i^k k \times (X_i - \bar{X})^2 \right) / (k - 1)$$
2. Calculating mean square within (MSW):  

$$\left( \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right) / (K - k)$$
3. Calculating F-value  

$$F_{value} = \left( \sum_i^k k \times (X_i - \bar{X})^2 \right) / (k - 1) \times \left( \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right) / (K - k)$$
4.  $F_{norm}$  = Min-Max normalization  $F_{value}$
5. While ( $f \leq n$ )  
 Calculate distance:  

$$d(x, X)_i = \sqrt{\sum_{i=1}^n (x_f - X_{fi})^2 \times F_{norm}}$$

$$f = f + i$$
6. End
7. End
8.  $asc\ d(x, X)_i$
9. Head ( $m$ )
10. Classify  $x$ :  $C(x_i) = argmax_k \sum x_j \in knn\ C(X_j, Y_k)$
11. End

## 2.2. Radius synthetic minority oversampling technique algorithm

The issues of overlapping, noise, and small disjunct cases were discovered to be emerging from the random selection of samples within the minority class data. The noise within the minority class data could lead to the creation of new noisy data, resulting in conflicts between the regions of each class. This challenge could be tackled using R-SMOTE method by filtering the sample data to ensure a more precise sample selection process. Therefore, the proposed modified SMOTE model was initiated by categorizing minority class data points into three groups including safe, noise, or small disjunct. The data selection or filtering process was performed using KNN method based on the position and proximity of the data to other classes. Each minority data point was selected using KNN parameter set to 5, and those correctly classified were labeled as safe while

those classified to be a majority class data point were tagged noise or small disjunct. The categorization process was followed by the generation of new synthetic data which was limited to the safe category. Similar to SMOTE method, synthetic data were generated by identifying the nearest minority data points and drawing interpolation lines between them. The determination of the number of nearest data points in SMOTE method was also based on KNN with parameter  $k$  representing the number of nearest data points. However, the use of this  $k$  parameter as described earlier posed the risk of generating synthetic data that could cause overlapping between the minority and majority classes. Therefore, this study proposed the use of a radius parameter instead which was determined by finding the distance to the nearest majority data point from the sample to be used as radius value. All the new data points were generated within this radius boundary using the circle equation presented in (1) and exemplified in a two-dimensional vector.

$$\begin{aligned} \|\tilde{b} - \tilde{p}\| &\leq r^2 \\ \|\tilde{a} - \tilde{p}\| &\leq r^2 \end{aligned} \quad (1)$$

$$\sum_{i=1}^n (b_{ij} - p_{ij})^2 \leq r^2 \quad (2)$$

$$r^2 = \sum_{j=1}^n (p_j - t_j)^2 \quad (3)$$

Where  $p$  represents the center point of the circle (minority sample point), with  $(p_1, p_2, p_3, \dots, p_n)$  and  $t$  ( $t_1, t_2, t_3, \dots, t_n$ ) being the nearest majority points to the center of the circle,  $b_i$  is a new data point below radius with  $(b_1, b_2, b_3, \dots, b_n)$  where  $i=1 \dots n$ , then  $r^2$  is the distance between  $p$  and  $t$  as in (3). The proposed model is presented in Figure 2.

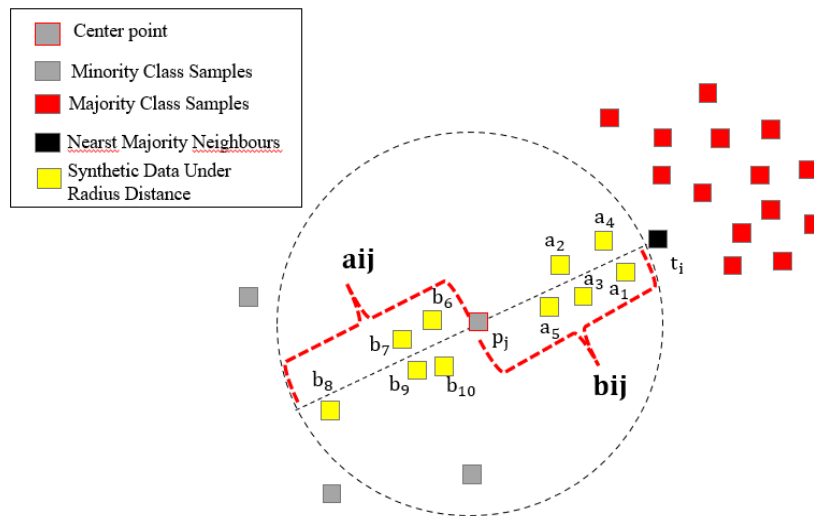


Figure 2. R-SMOTE algorithm scheme

The distance of each minority sample from the majority class was also calculated using the Euclidean distance method. The nearest majority data point was the one with the minimum value to all distances from minority data points, as shown in (4).

$$r_{ij} = \min \sum_{i=1}^n \sum_{j=1}^n \sqrt{(p_j - t_i)^2} \quad (4)$$

Where,  $r_{ij}$  represents the shortest distance between minority data to  $j$  and majority data to  $i$ . The identification of the majority data point was followed by the synthesis of new data through interpolation between these two points. The synthetic data were created along two directions of lines, including  $r_{ij}$  and  $-r_{ij}$  based on the (5) and (6):

$$a_{ij} = p_j + (\text{rand}(0,1) \times (r_{ij} - p_j)) \quad (5)$$

$$b_{ij} = p_j + (\text{rand}(0,1) \times (p_j - r_{ij})) \quad (6)$$

The area to produce these new data was limited to reduce the occurrence of overlapping as recorded in SMOTE method.

### 3. RESULTS AND DISCUSSION

This section was used to implement and test WfVANN method using several high-dimensional datasets. A comparative analysis was also conducted between WfVANN and two previous methods including the original KNN and LMKNN. Subsequently, a second round of experiments was applied to assess the effect of employing WfVANN as a classifier on oversampled data with some SMOTE development method, including Borderline-SMOTE, adaptive synthetic sampling (ADAYSN), safe level SMOTE, and SMOTE-IPF.

#### 3.1. Experimental framework and dataset characteristics

The experiment test was divided into two parts including the high-dimensional and imbalanced datasets. High-dimensional datasets were used to assess the effectiveness and application of the proposed method, WfVANN, in weighting each feature while imbalanced datasets were employed to examine its effect as a classifier when integrated into oversampled and non-oversampled data. Table 1 provides an overview of relevant metadata for the high-dimensional datasets, including attributes, sample counts, classes, and disease types. These microarray datasets were sourced from R packages designed to evaluate machine learning algorithms and models.

Table 1. Metadata for high-dimensional datasets

No	Dataset	#Samples	#Attr	Classes	Disease
1	Alon	62	2,000	2	Colon cancer
2	Borovecki	31	22,283	2	Huntington's Disease
3	Chiaretti	111	12,625	2	Leukemia
4	Chin	118	22,215	2	Breast cancer
5	Chowdary	104	22,283	2	Breast cancer
6	Christensen	217	1,413	3	-
7	Golub	72	7,129	3	Leukimia
8	Gordon	181	12,533	2	Lung Cancer
9	Gravier	168	2,905	2	Breast cancer
10	Khan	63	2,308	4	SRBCT

A total of 13 different imbalanced datasets were obtained from different application areas on binary and multiclass classification problems. The datasets used had a different number of features and a different imbalance ratio and were obtained from UCI machine learning repository[36] and knowledge extraction based on evolutionary learning (KEEL) repository [37]. Table 2 shows their characteristics with a focus on the imbalanced ratio value (IR), which represents the value of the ratio between negative and positive classes, the number of features (#Attr) in each dataset, the number of data or instances (#samples) in each dataset, as well as the number of comparisons of positive and negative instances in percent size.

Table 2. Metadata for imbalanced datasets

No	Name	IR	#Attr	#Samples	Positive instances (%)	Negative instances (%)
1	03subcl5-600-5-70-BI	5	2	600	16.67	83.30
2	04clover5z-600-5-70-BI	5	2	600	16.67	83.30
3	ecoli-0-1-3-7_vs_2-6	39.14	7	281	2.49	97.51
4	glass1	1.82	9	214	35.46	64.54
5	new thyroid	4.84	5	215	17.12	82.88
6	paw02a-600-5-70-BI	5	2	600	16.67	83.30
7	wine	1.5	13	178	40.00	60.00
8	yeast-1-4-5-8_vs_7	22.10	8	693	4.330	95.67
9	Umbilical Cord	18.87	5	151	5.300	94.70
10	Breast	2.36	9	286	29.12	70.38
11	Haberman	2.78	3	306	26.39	73.61
12	Pima	1.87	8	768	34.86	65.14
13	Bupa	1.38	6	345	42.19	57.81

The validation mechanism used was k-fold cross-validation with a total of 10 folds. Moreover, the performance of the classification model was tested using four metrics including accuracy, precision, recall, and F-Measure. In machine learning classification tasks, confusion matrix including true positive (TP), true negative (TN), false positive (FP), and false negative (FN) were the main parameters from which other performance metrics such as precision, recall, and F1 scores were computed. The accuracy was used to measure the amount of data correctly classified according to the ground-truth label divided by the total data used for the test. Precision was the rate of correct predictions among all samples predicted to belong to the minority class and indicated the number of positive predictions considered to be correct. Meanwhile, recall focused on the proportion of minority-class samples labeled as positive.

### 3.2. Performance analysis of high dimensional data

The performance of WfvANN model was observed to have excelled across most of the used high-dimensional datasets as presented in Table 3. In terms of accuracy, the model exhibited superior performance in datasets such as Borovecki, Chin, Chowdary, Christensen, Golub, and Khan. A substantial improvement was also observed in the Borovecki dataset compared to the other three methods as indicated by the difference of 33% recorded with KNN and LMKNN achieving an accuracy of 56% while WfvANN had 89%. The recall measurement showed that WfvANN had enhanced values in 6 out of the total 10 datasets with a remarkable 35% increase specifically recorded in Borovecki dataset. Furthermore, the assessment of the precision and F1-measure values indicated that WfvANN model showed commendable performance with increased values recorded across 6 datasets.

Table 3. The highest precision, recall, and F1 score (%) produced by KNN, LMKNN, and WfvANN

Dataset	Accuracy			Recall			Precision			F-1 Score		
	KNN	LMK	WfVA	KNN	LMK	WfVA	KNN	LMK	WfVA	KNN	LMK	WfVA
Alon	63	89	78	65	90	80	78	91	85	59	89	78
Borovecki	56	56	89	60	55	90	75	55	90	50	55	90
Chiaretti	54	77	64	25	53	51	21	49	45	22	51	46
Chin	83	89	92	79	87	89	89	90	94	80	88	91
Chowdary	94	97	100	95	96	100	93	97	100	93	97	100
Christensen	100	97	100	100	98	100	100	98	100	100	98	100
Golub	86	95	100	82	88	100	77	97	100	79	92	100
Gordon	85	100	96	56	100	89	92	100	98	56	100	93
Gravier	73	98	75	50	96	56	36	99	71	42	97	54
Khan	84	100	100	81	100	100	87	100	100	77	100	100

The effect of the  $k$  value is presented in Figure 3 with the proposed model denoted by square shape and its stability was found to be quite stable compared to other models. In case Alon dataset Figure 3(a), The accuracy remains somewhat consistent as the number of neighbors ( $k$ ) increases. There is some fluctuation between different values of  $k$ , but it stabilizes at around  $k=5$ . Furthermore, borovecki Figure 3(b), The accuracy remains stable and high across most values of  $k$ . The highest performance is observed with  $k=2$  and continues across different  $k$  values. In the chiaretti dataset Figure 3(c) shows a more fluctuating behavior with KNN, with a significant dip around  $k=3$  but stabilizing at higher values of  $k$ . Otherwise chin dataset Figure 3(d) shows accuracy fluctuates quite significantly for different values of  $k$ . It shows that the choice of  $k$  can drastically affect performance on this dataset. Chowdary dataset Figure 3(e), the accuracy is quite stable across different values of  $k$ , showing only a slight dip at some  $k$  values. Overall, proposed method performs consistently well on this dataset. Subsequently in Christensen Figure 3(f), shows high accuracy until around  $k=6$ , where there is a sharp decline in performance. After this point, the accuracy remains low. Golub Figure 3(g), there is notable fluctuation in accuracy for different values of  $k$ , indicating that this dataset is sensitive to the choice of  $k$ . The performance shows an upward trend after  $k=3$ . Furthermore, Gordon dataset Figure 3(h), shows performance on this dataset is generally stable, though there is a slight drop at specific  $k$  values. However, it remains high across the board. Gravier dataset in Figure 3(i) shows fluctuating accuracy, with a notable dip at  $k=4$ , but the performance bounces back with higher values of  $k$ . Then the last one is khan dataset in Figure 3(j) shows irregular performance, with large fluctuations in accuracy as  $k$  changes. The performance is inconsistent across different values of  $k$ . However, the optimal  $k$  value was different from those used in the accuracy metric for high-dimensional datasets. WfvANN showed reliability on high-dimensional datasets with 9 out of 10 datasets discovered to have achieved accuracy, precision, and F1-measure values that surpassed (or equal to) those of other methods, except in the Gravier dataset where LMKNN method had a higher accuracy value of 98%.

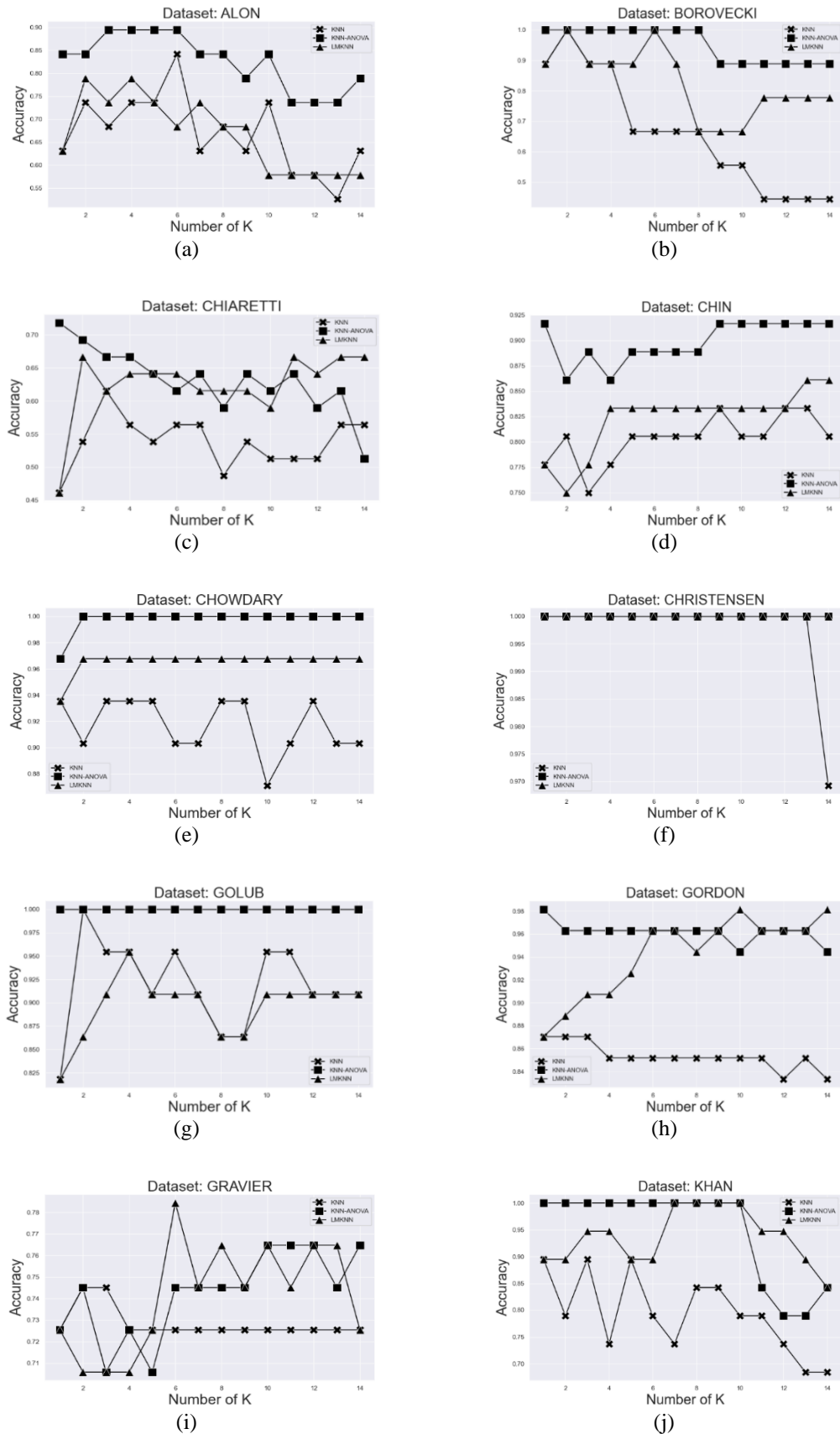


Figure 3. Accuracy performance using different k for high-dimensional datasets in dataset: (a) alon, (b) borovecki, (c) chiaretti, (d) chin, (e) chowdary, (f) christensen, (g) golub, (h) gordon, (i) gravier, and (j) khan



### 3.3. Performance analysis of the imbalanced data

The purpose of this experiment was to understand the performance of WfVANN model during the process of handling imbalanced datasets. This was considered necessary due to the ability of the imbalanced dataset conditions to cause a decline in the accuracy of conventional machine learning methods, particularly on minority data. Oversampling method was discovered to have become a trending solution to tackle these conditions and SMOTE was observed to have been widely used and developed in previous studies. Therefore, a comparison was made between the proposed R-SMOTE and two previous state-of-the-art methods including SMOTE and Borderline-SMOTE. The performance evaluation was conducted using four metrics including accuracy, precision, recall, and F1-measure as indicated in the following Table 4 where the highest results are highlighted in bold font. R-SMOTE was observed to have the best performance by achieving the highest accuracy values in 9 datasets including 04clover5z-600-5-70-BI, new-thyroid, wine, yeast-1-4-5-8\_vs\_7, umbilical cord, breast, haberman, pima, and bupa. The most significant difference in accuracy was recorded in the Bupa dataset where R-SMOTE had 83.2% while Borderline and SMOTE attained 57.3% and 64.3%, respectively. Moreover, from the analysis of recall values, R-SMOTE consistently outperformed the other three methods.

Table 4. Result of combinations of WFANN algorithm with oversampling SMOTE, Borderline-SMOTE, and R-SMOTE

Dataset	Accuracy			Recall			Precision			F-1 Score		
	SMT	BDR	RSMT	SMT	BDR	RSMT	SMT	BDR	RSMT	SMT	BDR	RSMT
03subcl5-600-5-70-BI	66.4	67.2	65.1	66.2	68.1	64.9	66.3	67.2	66.2	66.2	68.3	66.4
04clover5z-600-5-70-BI	76.1	73.2	82.5	76.1	73.5	82.5	75.6	71.2	82.5	76.1	71.3	83.2
Ecoli-0-1-3-7_vs_2-6	97.1	99.1	98.6	97.1	99.5	97.2	98.2	99.1	98.5	98.1	99.1	98.5
Glass1	79.1	88.3	74.2	79.2	87.3	74.3	79.1	87.3	73.3	80.1	86.8	74.1
New thyroid	97.5	99.5	99.5	97.5	99.5	99.5	97.2	99.3	99.2	97.1	98.9	98.9
Paw02a-600-5-70-BI	76.2	74.6	77.2	76.1	75.2	78.2	75.9	75.2	77.2	76.7	74.9	78.2
Wine	95.4	95.4	100	95.2	95.5	100	94.8	94.8	100	94.8	94.8	100
Yeast-1-4-5-8_vs_7	89.2	92.1	96.1	89.2	92.5	95.8	89.2	91.5	96.1	88.9	92.5	95.1
Umbilical Cord	96.3	96.2	96.3	96.3	96.2	96.3	96.3	96.4	97.2	96.3	97.7	98.1
Breast	74.3	76.2	87.2	75.1	76.2	86.4	74.3	76.2	86.2	73.4	76.3	88.3
Haberman	71.2	65.1	78.3	72.1	65.1	79.2	73.2	66.7	80.2	73.2	66.8	81.6
Pima	79.2	75.6	87.3	79.1	75.3	86.9	79.2	76.3	87.1	80.1	77.8	88.7
Bupa	67.3	64.3	83.2	67.3	64.2	82.1	67.2	64.2	82.3	66.8	63.4	81.8

SMT=SMOTE; BDR=Borderline-SMOTE; RSMT=R-SMOTE

The combination method of WfVANN and R-SMOTE was observed to have produced the highest values in 10 out of 13 datasets including 04clover5z-600-5-70-BI, Paw02a-600-5-70-BI, New-thyroid, Wine, Yeast-1-4-5-8\_vs\_7, Umbilical Cord, Breast, Haberman, Pima, and Bupa. This underscored the alignment between the test data facts and the prediction outcomes of WfVANN and R-SMOTE, thereby producing satisfying performance improvements. Moreover, the precision results also mirrored this trend, with 10 out of 13 datasets attaining the highest values through WfVANN R-SMOTE method. The most significant increase in precision was observed in the Bupa dataset, with 82.3% recorded for R-SMOTE, 64.2% for Borderline-SMOTE, and 67.2% for SMOTE. The results showed the alignment between the prediction outcome of the proposed model and the actual data and were found to be satisfactory for both the negative and positive classes. Similarly, the results were reflected in the F1-Measure metric with the proposed combination model discovered to have achieved the highest value in 10 datasets including 04clover5z-600-5-70-BI, Paw02a-600-5-70-BI, New-thyroid, Wine, Yeast-1-4-5-8\_vs\_7, Umbilical Cord, Breast, Haberman, Pima, and Bupa. The F1-measure also implied the simultaneous maximization of both precision and recall that offered a trade-off with one metric coming at the cost of another. More precision involved a harsher critic or classifier that doubts even the actual positive samples from the dataset, thereby reducing the recall score. Meanwhile, more recall entailed lax critic which allowed any sample resembling a positive class to pass and made border-case negative samples classified as “positive”, reducing the precision. The combination of WfVANN and R-SMOTE models effectively balanced and maximized precision and recall values based on the experiment results shown in Tables 3 and 4. Furthermore, Figure 4 shows the comparison of accuracy, precision, recall, and F1-measure outcomes for the proposed method on the Umbilical cord, indicating the most significant performance enhancement compared to other data and methods. This plot shows the accuracy in Figure 4(a), recall in Figure 4(b), precision in Figure 4(c), and F1-measure in Figure 4(d) of WfVANN model using different oversampling techniques as the number of k neighbors changes. SMOTE generally performs well across different values of k, maintaining higher accuracy compared to the other methods. Borderline SMOTE experiences significant dips at k=6 and k=10, showing unstable behavior with different values of k. R-SMOTE remains quite stable but tends to slightly underperform compared to SMOTE at higher k values.

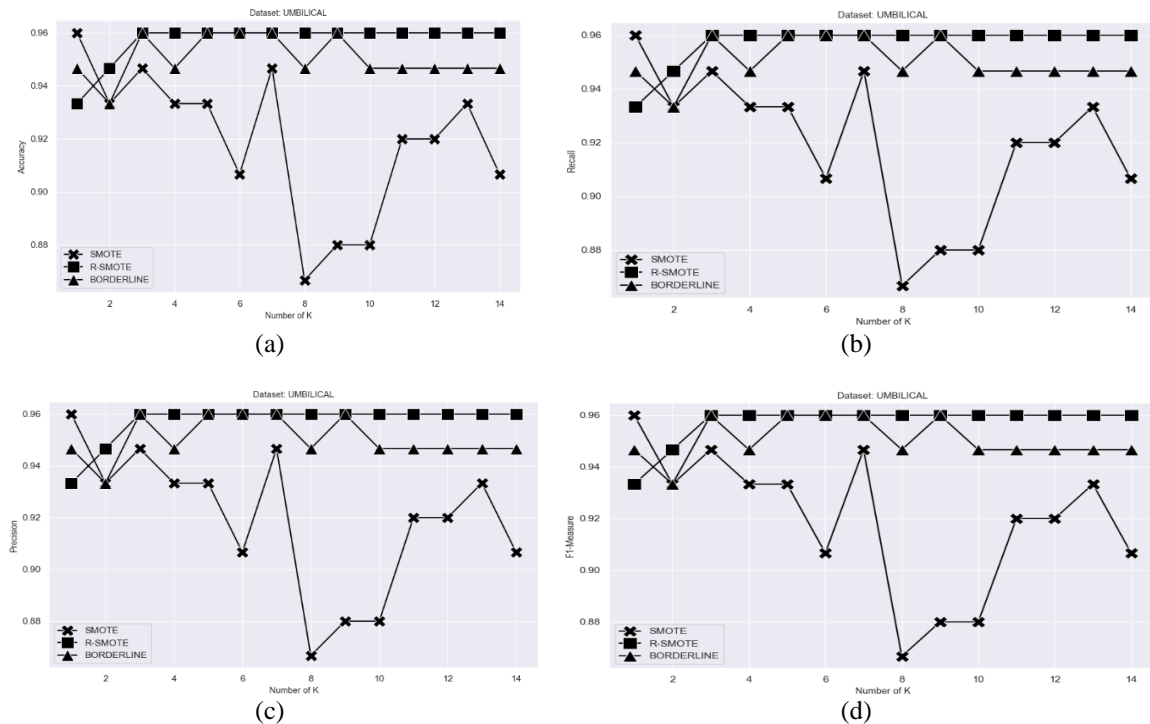


Figure 4. Comparison of the results of WFVANN combination with the oversampling algorithm in (a) accuracy, (b) recall, (c) precision, and (d) F1-measure on the umbilical cord dataset

#### 4. CONCLUSION

In conclusion, this study introduced an innovative KNN variant named WFVANN. The essence of this method depended on the observation of the relevance and contribution of each feature toward the calculation of Euclidean distance in KNN method. The feature relevance and contribution were measured using the F-value and weight of each feature, which depended on the magnitude of the resulting F-value. The evaluation was conducted using 10-fold cross-validation (10-FCV) with experiments applied to the two types of data including high-dimensional and imbalanced. The experiments on high-dimensional datasets showed that WFVANN outperformed other methods including KNN and LMKNN. This was confirmed by the fact that WFVANN model yielded satisfactory results with 6 out of 10 datasets achieving the highest values compared to other methods. The phenomenon indicated the effectiveness of using weights for relevant features in determining prediction outcomes. The results also showed that not all features contributed valuable information in determining data patterns in high-dimensional datasets, but some had the capacity to disrupt the learning process. The combination of R-SMOTE oversampling method at the data level and WFVANN method at the algorithmic level was proposed in the test of imbalanced datasets and indicated satisfying accuracy, precision, recall, and F1-measure values. R-SMOTE method showed superior performance metrics compared to SMOTE and Borderline-SMOTE. The results validated the effectiveness of constraining the area in R-SMOTE and modifying feature weights in WFVANN to enhance robustness against imbalanced data conditions. The limitations of this study were also acknowledged. The computational time was relatively high due to the calculation of feature weights for each feature, particularly in high-dimensional data. Therefore, special attention was required to address computational time constraints to further refine this model. In the future, feature weighting development should be combined with other feature selection methods, and distance calculations explored using alternative methods such as Minkowski and Manhattan distances. Future developments should also focus on determining the most optimal  $k$  value automatically.

#### ACKNOWLEDGEMENTS

The authors are grateful to the Directorate of Research, Technology, and Community Service (DPRM) Indonesia for funding this study through Program Funding (Fundamental Reguler) for the 2023 Fiscal Year. Next, thank you to the research assistants, namely Arya Faisal Akbar and Hendra Wijaya who have helped carry out this research in the intelligent systems laboratory of Institut Teknologi dan Bisnis STIKOM Bali.




## REFERENCES

- [1] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316.
- [2] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Improving classification performance of fetal umbilical cord using combination of SMOTE method and multiclassifier voting in imbalanced data and small dataset," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 441–454, 2020, doi: 10.22266/ijies2020.1031.39.
- [3] R. Wardoyo, I. M. A. Wirawan, and I. G. A. Pradipta, "Oversampling approach using radius-SMOTE for imbalance electroencephalography datasets," *Emerging Science Journal*, vol. 6, no. 2, pp. 382–398, 2022, doi: 10.28991/ESJ-2022-06-02-013.
- [4] H. Zhang, W. Yang, W. Yi, J. B. Lim, Z. An, and C. Li, "Imbalanced data based fault diagnosis of the chiller via integrating a new resampling technique with an improved ensemble extreme learning machine," *Journal of Building Engineering*, vol. 70, 2023, doi: 10.1016/j.jobbe.2023.106338.
- [5] S. Sun, T. Wang, and F. Chu, "A multi-learner neural network approach to wind turbine fault diagnosis with imbalanced data," *Renewable Energy*, vol. 208, pp. 420–430, 2023, doi: 10.1016/j.renene.2023.03.097.
- [6] D. Liu, S. Zhong, L. Lin, M. Zhao, X. Fu, and X. Liu, "Deep attention SMOTE: Data augmentation with a learnable interpolation factor for imbalanced anomaly detection of gas turbines," *Computers in Industry*, vol. 151, 2023, doi: 10.1016/j.compind.2023.103972.
- [7] Y. Gao, X. Yin, Z. He, and X. Wang, "A deep learning process anomaly detection approach with representative latent features for low discriminative and insufficient abnormal data," *Computers and Industrial Engineering*, vol. 176, 2023, doi: 10.1016/j.cie.2022.108936.
- [8] J. Jiang *et al.*, "A dynamic ensemble algorithm for anomaly detection in IoT imbalanced data streams," *Computer Communications*, vol. 194, pp. 250–257, 2022, doi: 10.1016/j.comcom.2022.07.034.
- [9] H. Ding, L. Chen, L. Dong, Z. Fu, and X. Cui, "Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection," *Future Generation Computer Systems*, vol. 131, pp. 240–254, 2022, doi: 10.1016/j.future.2022.01.026.
- [10] G. Mohiuddin *et al.*, "Intrusion detection using hybridized meta-heuristic techniques with weighted XGBoost classifier," *Expert Systems with Applications*, vol. 232, 2023, doi: 10.1016/j.eswa.2023.120596.
- [11] M. S. Milosevic and V. M. Ciric, "Extreme minority class detection in imbalanced data for network intrusion," *Computers and Security*, vol. 123, 2022, doi: 10.1016/j.cose.2022.102940.
- [12] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, 2021, doi: 10.1016/j.patcog.2021.107965.
- [13] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, 2023, doi: 10.1016/j.asoc.2023.110415.
- [14] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem: a review," *2021 6th International Conference on Informatics and Computing, ICIC 2021*, 2021, doi: 10.1109/ICIC54025.2021.9632912.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [16] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences*, vol. 291, pp. 184–203, 2015, doi: 10.1016/j.ins.2014.08.051.
- [17] S. Gazzah and N. E. Ben Amara, "New oversampling approaches based on polynomial fitting for imbalanced data sets," *DAS 2008 - Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, pp. 677–684, 2008, doi: 10.1109/DAS.2008.74.
- [18] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 238–251, 2016, doi: 10.1109/TKDE.2015.2458858.
- [19] W. A. Young, S. L. Nykl, G. R. Weckman, and D. M. Chelberg, "Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets," *Neural Computing and Applications*, vol. 26, no. 5, pp. 1041–1054, 2015, doi: 10.1007/s00521-014-1780-0.
- [20] C. Bunkhumpornpat and S. Subpaiboonkit, "Safe level graph for synthetic minority over-sampling techniques," *13th International Symposium on Communications and Information Technologies: Communication and Information Technology for New Life Style Beyond the Cloud, ISCIT 2013*, pp. 570–575, 2013, doi: 10.1109/ISCIT.2013.6645923.
- [21] Q. Liu *et al.*, "Application of KM-SMOTE for rockburst intelligent prediction," *Tunnelling and Underground Space Technology*, vol. 138, 2023, doi: 10.1016/j.tust.2023.105180.
- [22] 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," *Information Sciences*, vol. 572, pp. 574–589, 2021, doi: 10.1016/j.ins.2021.02.056.
- [23] Z. Xiang, Y. Su, J. Lan, D. Li, Y. Hu, and Z. Li, "An improved SMOTE algorithm using clustering," *Proceedings - 2020 Chinese Automation Congress, CAC 2020*, pp. 1986–1991, 2020, doi: 10.1109/CAC51589.2020.9327176.
- [24] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," *Advances in Knowledge Discovery and Data Mining*, pp. 475–482, 2009, doi: 10.1007/978-3-642-01307-2\_43.
- [25] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *Advances in Intelligent Computing*, pp. 878–887, 2005, doi: 10.1007/11538059\_91.
- [26] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.
- [27] R. Alejo, V. García, and J. H. Pacheco-Sánchez, "An efficient over-sampling approach based on mean square error back-propagation for dealing with the multi-class imbalance problem," *Neural Processing Letters*, vol. 42, no. 3, pp. 603–617, 2015, doi: 10.1007/s11063-014-9376-3.
- [28] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, "LVQ-SMOTE - learning vector quantization based synthetic minority over-sampling technique for biomedical data," *BioData Mining*, vol. 6, no. 1, 2013, doi: 10.1186/1756-0381-6-16.
- [29] J. B. Wang, C. A. Zou, and G. H. Fu, "AWSMOTE: An SVM-based adaptive weighted SMOTE for class-imbalance learning," *Scientific Programming*, vol. 2021, 2021, doi: 10.1155/2021/9947621.




- [30] J. Gou, W. Qiu, Z. Yi, X. Shen, Y. Zhan, and W. Ou, "Locality constrained representation-based K-nearest neighbor classification," *Knowledge-Based Systems*, vol. 167, pp. 38–52, 2019, doi: 10.1016/j.knsys.2019.01.016.
- [31] Z. Pan, Y. Wang, and W. Ku, "A new k-harmonic nearest neighbor classifier based on the multi-local means," *Expert Systems with Applications*, vol. 67, pp. 115–125, 2017, doi: 10.1016/j.eswa.2016.09.031.
- [32] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," *Pattern Recognition Letters*, vol. 27, no. 10, pp. 1151–1159, 2006, doi: 10.1016/j.patrec.2005.12.016.
- [33] J. Gou, Y. Zhan, Y. Rao, X. Shen, X. Wang, and W. He, "Improved pseudo nearest neighbor classification," *Knowledge-Based Systems*, vol. 70, pp. 361–375, 2014, doi: 10.1016/j.knsys.2014.07.020.
- [34] J. Gou, W. Qiu, Q. Mao, Y. Zhan, X. Shen, and Y. Rao, "A multi-local means based nearest neighbor classifier," *International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2017, pp. 448–452, 2017, doi: 10.1109/ICTAI.2017.00075.
- [35] Y. Zeng, Y. Yang, and L. Zhao, "Pseudo nearest neighbor rule for pattern classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3587–3595, 2009, doi: 10.1016/j.eswa.2008.02.003.
- [36] "About," *UC Irvine Machine Learning Repository*. [Online]. Available: <https://archive.ics.uci.edu/about>
- [37] "KEEL-dataset repository," *KEEL-Knowledge Extraction based on Evolutionary Learning*. [Online]. Available: <https://sci2s.ugr.es/keel/datasets.php>

## BIOGRAPHIES OF AUTHORS






**Gede Angga Pradipta**    holds a Doctor of Computer Science from Department of Computer Science and Electronics, Faculty of Natural Sciences, Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2021. He also received a bachelor's degree in computer informatics from Universitas Atma Jaya (UAJY), Yogyakarta, Indonesia, in 2012 and a master's degree in information technology from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2014. His research interests include machine learning, pattern recognition, and image processing. He is currently lecturing with Department of Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. He can be contacted at email: [angga\\_pradipta@stikom-bali.ac.id](mailto:angga_pradipta@stikom-bali.ac.id).






**Putu Desiana Wulaning Ayu**    received the Dr. (Doctor) in Computer Science from The Department of Computer Science and Electronics, Faculty of Mathematics and Natural Science Universitas Gadjah Mada, with the dissertation "Segmentation and feature extraction model on 2-D ultrasonography images for amniotic fluid classification". Her research interests are medical image processing, machine learning, deep learning, and computer vision. She is lecturing in Magister Information Systems, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. She is a member Indonesian Computer, Electronics, and Instrumentation Support Society. She can be contacted at email: [wulaning\\_ayu@stikom-bali.ac.id](mailto:wulaning_ayu@stikom-bali.ac.id).



**Made Liandana**    holds a Master of Engineering from the Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2014. He also received a bachelor's degree in computer systems from STMIK STIKOM Bali, Denpasar, Indonesia, in 2011. His research interests include machine learning, the internet of things, and wearable device. He is currently lecturing with the Department of Informatics and Computer, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia. He can be contacted at email: [liandana@stikom-bali.ac.id](mailto:liandana@stikom-bali.ac.id).



**Dandy Pramana Hostyadi**    received a bachelor's degree from Institut Teknologi dan Bisnis STIKOM Bali master's degree from Udayana University, and a Doctoral degree from Institut Teknologi Sepuluh Nopember, all in computer science. He is now an Assistant Professor and head of the Cyber and defense technology division at the Center of Excellence directorate in Institut Teknologi dan Bisnis STIKOM Bali. Also, he manages the network cyber and malware (NCM) lab at Institut Teknologi dan Bisnis STIKOM Bali. His research interests include network security, AI, information security, and computer networks. He can be contacted at email: [dandy@stikom-bali.ac.id](mailto:dandy@stikom-bali.ac.id).