A Global Weighting Scheme Based on Intra-Class and Inter-Class Term Distributions in Bag-of-Visual Words Image Classification

Catur Supriyanto, Hanung Adi Nugroho, Teguh Bharata Adji

***Abstract*—Bag-of-visual words (BoVW) is one of the most popular image representations in the image classification. Many features (also called visual words or codebooks) are generated to create the vector of images. However, similar to *term* in the text classification, these features may be relevant or irrelevant which affect the accuracy of classification. Many global weighting schemes (e.g. inverse document frequency) have been proposed to detect the relevant features. These global weighting schemes are based on document frequency (DF) in which the features will have the same weight when they have the same DF. This condition leads to reduce the discriminative power of features. Therefore, this study proposes a global weighting scheme based on intra-class and inter-class term distributions. The experiment was conducted by comparing the proposed method with the state-of-the-art global weighting schemes called inverse gravity moment (IGM) and modified inverse document frequency (mIDF). The evaluation of these weighting schemes is performed on the BoVW based image classification. Support vector machine (SVM) is used as a classifier to evaluate the methods on several benchmark datasets. By using the statistical analysis, such as Friedman nonparametric test, the proposed global weighting scheme outperforms the state-of-the-art global weighting schemes.**

***Index Terms*—bag-of-visual words, image classification, global weighting scheme, term distribution.**

I. Introduction

CONTENT-based image retrieval (CBIR) has become an interest research area in the last decade. Searching, classification, and clustering are widely applied in CBIR. In the large-scale image collection, CBIR is more feasible than text-based image retrieval [1]. Manually labeling is the drawback of text-based image retrieval which is very time-consuming. Moreover, the subjectivity of each person is different to label each image. Therefore, the same image may have different label from the different person. CBIR has been proposed to extract the features independently without human intervention.

Similar to text mining, an image is represented as a vector. The length of vector shows the number of features. Each image may have dozens or hundreds of features and some of the features may not useful for image retrieval. They may be redundant, noise or even irrelevant features. Feature weighting and feature selection are the preprocessing steps which can handle these problems. Both feature weighting and feature selection are used to weight the features. However, feature weighting is more flexible than feature selection [2]. Feature weighting assigns weights to the features with continues value and feature selection assigns weights to the features with binary value 0 and 1. Features with 0 value will be removed from the dataset. Feature selection will reduce the number of features in the dataset, but still maintaining the high accuracy [3]. Feature selection produces the selected features which are considered as the relevant features, while feature weighting produces the new weight of each feature in the document which is a multiplication of local and global weight. The relevant features should have high class distinguishing power to improve the accuracy of classification.

In the weighting scheme, the weight of feature can be composed of local weight and global weight [4]. The use of term frequency as local weighting fails to yield the accurate classification. Therefore, several global weighting schemes have been proposed. The global weighting scheme is an approach to detect the relevant and irrelevant features. Based on the existence of class information of the dataset, global weighting schemes are classified into supervised and unsupervised weighting schemes [5] [6] [7]. Unsupervised weighting schemes do not require any class information on the dataset, for example inverse document frequency (IDF). Other methods such as chi-square (CHI), information gain (IG), odds ratio (OR), inverse class-space-density-frequency (IDF-ICSDF) are classified into supervised weighting schemes which need class information on the dataset. These global weighting schemes are based on document frequency (DF), including inverse gravity moment (IGM) [4] and modified inverse document frequency (mIDF) [8]. As the newly global weighting schemes, IGM and mIDF have been proven to be more accurate than the other DF based global weighting schemes.

A problem may arise when we use DF in the global weighting scheme. The terms will have the same weight when they have the same DF in a category, even though they may have a different term frequency (TF) distribution. This condition leads to reduce the discriminative power of the features. Thus, intra-class and inter-class term distributions defined in Zhou *et al.* [9] is adopted. In [9], Zhou *et al.* proposed the concept of the intra-class and inter-class term distributions for feature selection.

The objective of this study is to improve the accuracy of BoVW image classification by using intra-class and inter-class term distributions based global weighting scheme. The rest of the paper is organized as follows. Section 2 presents

Manuscript received May 31, 2017; revised February 3, 2018

Catur Supriyanto (Corresponding author) is with the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia; and Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia. Email: catur.supriyanto@mail.ugm.ac.id.

Hanung Adi Nugroho is with the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia. Email: adinugroho@ugm.ac.id.

Teguh Bharata Adji is with the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia. Email: adji@ugm.ac.id.

the theory of current weighting schemes, Section 3 describes our proposed weighting scheme, Section 4 presents the result of the experiment, and the last Section 5 concludes the results.

II. Current Weighting Scheme Methods

A. Global Weighting Scheme

Weighting scheme is the important task in the process of classification. Both text and image classification utilize the weighting scheme to assign weights to the relevant terms or features in each category. Weighting scheme is able to enhance the discriminative power of the features [10]. In the traditional weighting scheme, such as term frequency inverse document frequency (TF-IDF), the local weight is represented by term frequency (TF) and the global weight is represented by inverse document frequency (IDF). TFIDF of a term t_i is defined in (1).

$$w(t_i, d) = tf_{id} \times IDF_i \tag{1}$$

$$IDF_i = \log\left(\frac{N}{df_i}\right) \tag{2}$$

Here, tf_{id} is the term frequency of term t_i which is occur in the document, N is the total document, and df_i is the number of documents that contain term t_i.

B. Inverse Gravity Moment (IGM)

A newly weighting scheme called term frequency inverse gravity moment (TF-IGM) has been proposed by Chen *et al.* [4]. In their experiment, TF-IGM outperforms several weighting schemes such as TF-IDF, TF-RF, TF-CHI, TF-Prob, and TF-IDF-ICSDF. The TF-IGM value of term t_i in a document d is calculated as follows:

$$w(t_i, d) = tf_{id} \times IGM\,(t_i) \tag{3}$$

$$IGM(t_i) = 1 + \lambda_1 \times \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r} \tag{4}$$

Here, f_{kr} $(r = 1, 2, , m)$ is the total number of documents containing term t_i in the r-th class, which are sorted in descending order. λ_1 is the adjustment coefficient and its value is set to 5 to 9.

C. Modified Inverse Document Frequency (mIDF)

Similar to IGM, mIDF uses DF to measure the global weight of a term, which is proposed by Sabbah *et al.* [8]. The weight of term t_i in the document d can be calculated by (5). The global weight mIDF of term t_i is represented as (6). Differ from traditional IDF, mIDF calculates the number of documents where term t does not appear in the document $(N - DF_t)$.

$$w(t_i, d) = tf_{id} \times mIDF_i \tag{5}$$

$$mIDF_i = log\left[\frac{N}{1/((N - df_i) + 1)}\right] \tag{6}$$

III. The Proposed Feature Weighting Scheme

As we stated in the Introduction section, global weighting scheme based on DF has a problem, suppose we have t_1 and t_2 with their DFs in two categories being $t_1\{3, 3\}$ and $t_2\{3, 3\}$, therefore these terms will have the same global weight. However, if these terms have their TF being $C_1 \rightarrow t_1\{1, 2, 1\}, t_2\{2, 1, 9\}, C_2 \rightarrow t_1\{9, 8, 8\}, t_2\{3, 7, 1\}$, apparently the global weight of t_1 is higher than t_2. Since the distribution of TF in t_1 is more homogeneous than t_2. This concept is based on the method proposed by Zhou *et al.* [9].

Based on the problem mentioned above, we propose global weighting scheme based on intra-class and inter-class term distributions. The concept of intra-class and inter-class term distributions which has been proposed by Zhou *et al.* [9] is computed by the following equations:

$$s(t_i)^2 = \frac{1}{K}\sum_{k=1}^{K}(\overline{tf_{ki}} - \overline{tf_i})^2 \tag{7}$$

$$s(t_{ki})^2 = \frac{1}{|C_k|}\sum_{j \in C_k}^{K}(tf_{ij} - \overline{tf_{ki}})^2 \tag{8}$$

$$F(t_{ki}) = \frac{s(t_i)^2}{s(t_{ki})^2} \times \frac{\overline{tf_{ki}}}{\overline{tf_i}} \tag{9}$$

$$\lambda_2 = \frac{K!}{(K-2)! \cdot 2!} \tag{10}$$

$$G(t_i) = \frac{1}{\lambda_2} \cdot \sum_{1 \le q < r \le K} |F(t_{q,i}) - F(t_{r,i})| \tag{11}$$

Here, f_{ij} is term frequency of term t_i in document j, $\overline{tf_i}$ is the average term frequency of term t_i in the collection of documents, $\overline{tf_{ik}}$ is the average term frequency of term t_i in the category k, $|C_k|$ is the document frequency of term t_i in the category k, and K is the number of categories.

In order to adopt the concept of inter-class and intra-class term distributions in the proposed weighting scheme, we use G (11) to replace IGM (4) or replace $mIDF$ (6) as global weighting scheme, hence the weight of term is calculated by (12).

$$w(t_i, d) = tf_{id} \times G\,(t_i) \tag{12}$$

In our case, some features do not exist in some categories. Therefore, the value of intra-class term distribution $s(t_{ki})^2$ is 0 and the global weight of term $G(t_i)$ return the NaN value. In order to overcome that problem, we add one $(+1)$ to the intra-class term distribution as shown in the following formula.

$$F(t_{ki}) = \frac{s(t_i)^2}{s(t_{ki})^2 + 1} \times \frac{\overline{tf_{ki}}}{\overline{tf_i}} \tag{13}$$

An example of the proposed weighting scheme written as below.

1) Assume that we have two features t_1 and t_2 in the three classes. Each class consists of two documents, as shown in Table I.
2) Calculate $\overline{tf_i}$
   $\overline{tf_{t1}} = \frac{2+1+1+2+0+0}{6} = 1$

Fig. 1: Example images from Coil-100 dataset

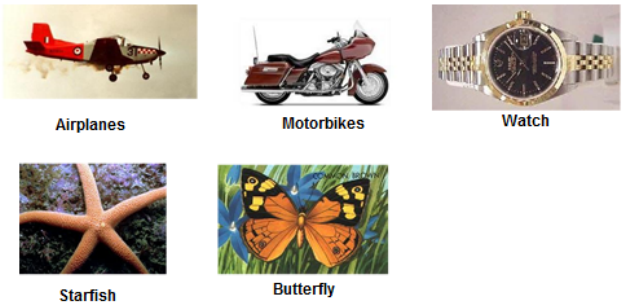

Fig. 2: Example images from Caltech-101 dataset

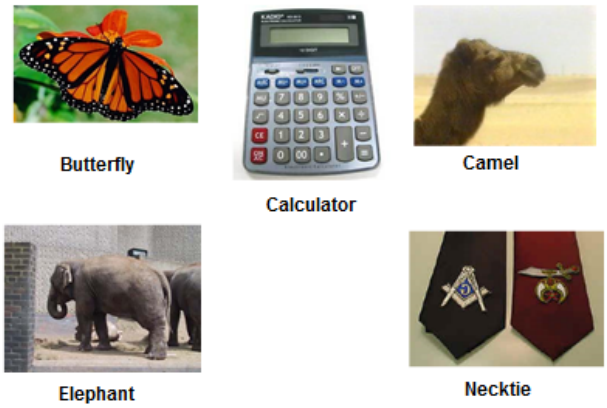

Fig. 3: Example images from Caltech-256 dataset

$\overline{tf_{t2}} = \frac{2+1+0+0+0+0}{6} = 0.5$

3) Calculate $\overline{tf_{ki}}$

$\overline{tf_{C_1,t_1}} = \frac{2+1}{2} = 1.5$

$\overline{tf_{C_2,t_1}} = \frac{1+2}{2} = 1.5$

$\overline{tf_{C_3,t_1}} = \frac{0+0}{2} = 0$

$\overline{tf_{C_1,t_2}} = \frac{2+1}{2} = 1.5$

$\overline{tf_{C_2,t_2}} = \frac{0+0}{2} = 0$

$\overline{tf_{C_3,t_2}} = \frac{0+0}{2} = 0$

4) Calculate $s(t_i)^2$ using (7)

$s(t_1)^2 = \frac{1}{3}((1.5-1)^2 + (1.5-1)^2 + (0-1)^2) = 0.5$

$s(t_2)^2 = \frac{1}{3}((1.5-0.5)^2 + (0-0.5)^2 + (0-0.5)^2) = 0.5$

5) Calculate $s(t_{ki})^2$ using (8)

$s(t_{C_1,t_1})^2 = \frac{1}{2}((2-1.5)^2 + (1-1.5)^2) = 0.25$

$s(t_{C_1,t_2})^2 = \frac{1}{2}((1-1.5)^2 + (2-1.5)^2) = 0.25$

$s(t_{C_1,t_3})^2 = \frac{1}{2}((0-0)^2 + (0-0)^2) = 0$

$s(t_{C_2,t_1})^2 = \frac{1}{2}((2-1.5)^2 + (1-1.5)^2) = 0.25$

$s(t_{C_2,t_2})^2 = \frac{1}{2}((0-0)^2 + (0-0)^2) = 0$

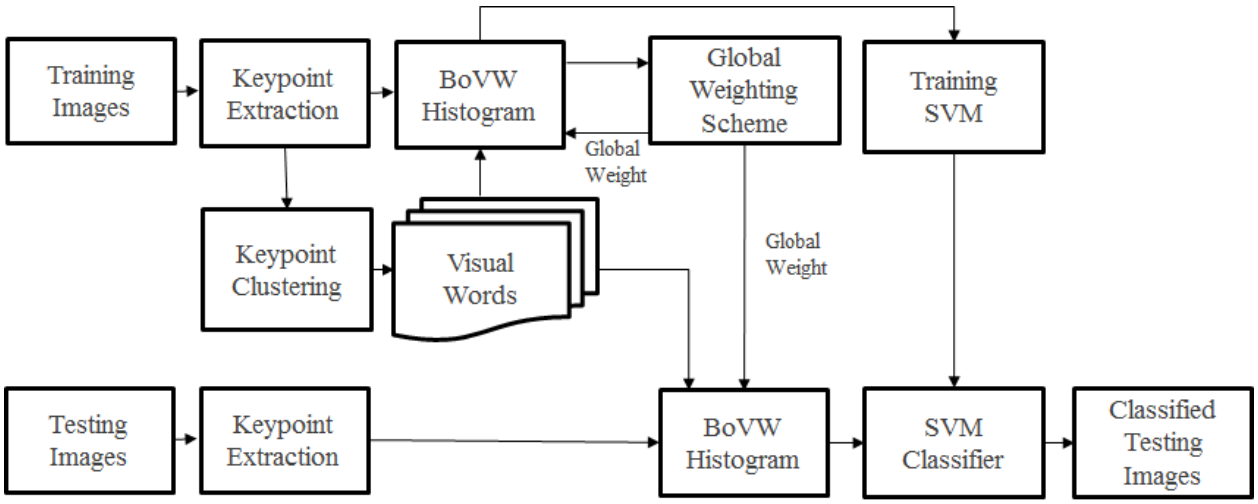

Fig. 4: The Process of BoVW Image Classification

$$s(t_{C_2,t_3})^2 = \frac{1}{2}((0-0)^2 + (0-0)^2) = 0$$

6) Calculate $F(t_{ki})$ using (9)

$$F(t_{C_1,t_1}) = \frac{0.5}{0.25} \times \frac{1.5}{1} = 3$$

$$F(t_{C_1,t_2}) = \frac{0.5}{0.25} \times \frac{1.5}{1} = 3$$

$$F(t_{C_1,t_3}) = \frac{0.5}{0} \times \frac{0}{1} = NaN$$

$$F(t_{C_2,t_1}) = \frac{0.5}{0.25} \times \frac{1.5}{0.5} = 6$$

$$F(t_{C_2,t_2}) = \frac{0.5}{0} \times \frac{0}{0.5} = NaN$$

$$F(t_{C_2,t_3}) = \frac{0.5}{0} \times \frac{0}{0.5} = NaN$$

Since the calculations produce NaN values which are caused by the absence of term in the class, we add one (+1) as shown in (13).

7) Calculate $F(t_{ki})$ using (13)

$$F(t_{C_1,t_1}) = \frac{0.5}{0.25+1} \times \frac{1.5}{1} = 0.6$$

$$F(t_{C_1,t_2}) = \frac{0.5}{0.25+1} \times \frac{1.5}{1} = 0.6$$

$$F(t_{C_1,t_3}) = \frac{0.5}{0+1} \times \frac{0}{1} = 0$$

$$F(t_{C_2,t_1}) = \frac{0.5}{0.25+1} \times \frac{1.5}{0.5} = 1.2$$

$$F(t_{C_2,t_2}) = \frac{0.5}{0+1} \times \frac{0}{0.5} = 0$$

$$F(t_{C_2,t_3}) = \frac{0.5}{0+1} \times \frac{0}{0.5} = 0$$

8) Calculate λ_2 using (10)

$$\lambda_2 = \frac{3!}{(3-2)!.2!} = \frac{6}{2} = 3$$

9) Calculate the global weight $G(t_i)$ using (11)

$$G(t_1) = \frac{1}{3} \times (|0.6-0.6| + |0.6-0| + |0.6-0|) = 0.4$$

TABLE I: The example of term-document matrix

	C1		C2		C3	
	d1	d2	d3	d4	d5	d6
t1	2	1	1	2	0	0
t2	2	1	0	0	0	0

TABLE II: The result of the proposed global weighting scheme for the term-document matrix

	C1		C2		C3	
	d1	d2	d3	d4	d5	d6
t1	0.8	0.4	0.4	0.8	0	0
t2	1.6	0.8	0	0	0	0

$$G(t_2) = \frac{1}{3} \times (|1.2-0| + |1.2-0| + |0-0|) = 0.8$$

10) Multiply the local weight from Table I with the global weight using (12) and the result is shown in Table II

IV. Experiments

A. Dataset

In order to evaluate the proposed global weighting scheme, the commonly benchmark datasets i.e. Coil-100 [11], Caltech-101 [12], and Caltech-256 [13] are used in the image classification task. Coil-100 contains 100 object categories and each category contains 72 images. Each dataset is split into 70% for training images and 30% for testing images. Therefore, in each category, 50 images are used for training and 22 images for testing.

Caltech-101 has 101 object categories. Each category contains 40-800 images. Those images are randomly selected until only 80 images, which are divided into 50 training images and 30 testing images. The third dataset, Caltech-256 has 256 categories and contains 30,607 images.

In this experiment, six categories of Coil-100, five categories of the Caltech-101 (airplanes, butterfly, motorbikes, starfish, and watch), and five categories of Caltech-256 (butterfly, calculator, camel, elephant, and necktie) are selected randomly (see Fig. 1, Fig. 2, and Fig. 3 respectively). The selection of these datasets, the number of classes and the number of images of each category are based on Wang *et al.* [14].

TABLE III: Ranks of global weighting schemes based on classification accuracy expressed in % and Shapiro-Wilks test through different amount of visual words on Coil-100 dataset

Centroid Initialization	100 Visual Words			200 Visual Words			500 Visual Words		
	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method
1	75.76 (2)	63.64 (1)	91.67 (3)	63.64 (2)	60.61 (1)	78.79 (3)	56.06 (1)	62.88 (2)	87.12 (3)
2	91.67 (2)	85.61 (1)	97.73 (3)	68.18 (2)	61.36 (1)	96.97 (3)	57.58 (1)	61.36 (2)	82.58 (3)
3	75.76 (2)	63.64 (1)	91.67 (3)	65.91 (2)	63.64 (1)	94.70 (3)	50.00 (1)	62.88 (2)	84.09 (3)
4	75.76 (2)	60.61 (1)	83.33 (3)	57.58 (1)	62.12 (2)	81.06 (3)	61.36 (1)	65.15 (2)	80.30 (3)
5	72.73 (2)	60.61 (1)	81.06 (3)	58.33 (1)	63.64 (2)	97.73 (3)	56.82 (1)	63.64 (2)	85.61 (3)
6	78.03 (2)	65.91 (1)	94.70 (3)	78.79 (2)	61.36 (1)	81.82 (3)	62.12 (1)	66.67 (2)	86.36 (3)
7	76.52 (2)	63.64 (1)	**99.24 (3)**	53.79 (1)	65.15 (2)	**98.48 (3)**	75.76 (2)	65.91 (1)	84.85 (3)
8	62.88 (1.5)	62.88 (1.5)	97.73 (3)	56.82 (1)	61.36 (2)	96.21 (3)	40.91 (1)	62.12 (2)	87.12 (3)
9	70.45 (2)	62.12 (1)	87.12 (3)	52.27 (1)	63.64 (2)	92.42 (3)	81.82 (2)	62.88 (1)	**87.88 (3)**
10	62.88 (2)	62.12 (1)	96.97 (3)	76.52 (2)	65.91 (1)	95.45 (3)	70.45 (2)	63.64 (1)	83.33 (3)
Average Accuracy	74.24	65.08	**92.12**	63.18	62.88	**91.36**	61.29	63.71	**84.92**
Shapiro-Wilk p-value	0.15941	**0.00001**	0.21631	0.34324	0.29117	**0.01231**	0.91824	0.46685	0.67176

TABLE IV: Ranks of global weighting schemes based on classification accuracy expressed in % and Shapiro-Wilks test through different amount of visual words on Caltech-101 dataset

Centroid Initialization	100 Visual Words			200 Visual Words			500 Visual Words		
	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method
1	23.33 (2)	20.67 (1)	64.00 (3)	33.33 (2)	20.67 (1)	65.33 (3)	28.67 (2)	20.00 (1)	**54.00 (3)**
2	34.67 (2)	20.67 (1)	64.00 (3)	31.33 (2)	20.67 (1)	62.00 (3)	33.33 (2)	20.00 (1)	50.67 (3)
3	32.00 (2)	20.67 (1)	62.00 (3)	32.67 (2)	20.67 (1)	66.00 (3)	34.67 (2)	21.33 (1)	**54.00 (3)**
4	32.00 (2)	22.00 (1)	65.33 (3)	31.33 (2)	20.00 (1)	65.33 (3)	32.00 (2)	22.00 (1)	49.33 (3)
5	32.00 (2)	20.67 (1)	**68.00 (3)**	35.33 (2)	21.33 (1)	66.00 (3)	30.67 (2)	21.33 (1)	51.33 (3)
6	33.33 (2)	24.00 (1)	62.67 (3)	34.00 (2)	22.00 (1)	64.67 (3)	30.00 (2)	20.00 (1)	50.00 (3)
7	30.67 (2)	21.33 (1)	66.67 (3)	29.33 (2)	21.33 (1)	62.00 (3)	34.67 (2)	21.33 (1)	**54.00 (3)**
8	31.33 (2)	20.00 (1)	64.67 (3)	34.67 (2)	20.67 (1)	**66.67 (3)**	28.67 (2)	21.33 (1)	**54.00 (3)**
9	33.33 (2)	20.00 (1)	63.33 (3)	35.33 (2)	21.33 (1)	64.67 (3)	26.67 (2)	21.33 (1)	52.00 (3)
10	34.00 (2)	21.33 (1)	64.00 (3)	34.00 (2)	22.00 (1)	64.00 (3)	32.00 (2)	22.00 (1)	49.33 (3)
Average Accuracy	31.67	21.13	**64.47**	33.13	21.07	**64.67**	31.14	21.07	**51.87**
Shapiro-Wilk p-value	**0.00257**	**0.01829**	0.60840	0.38818	0.24680	0.16213	0.69498	**0.01233**	**0.03957**

TABLE V: Ranks of global weighting schemes based on classification accuracy expressed in % and Shapiro-Wilks test through different amount of visual words on Caltech-256 dataset

Centroid Initialization	100 Visual Words			200 Visual Words			500 Visual Words		
	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method	IGM	mIDF	Proposed Method
1	34.00 (2)	25.33 (1)	43.33 (3)	31.33 (2)	25.33 (1)	42.67 (3)	31.33 (2)	27.33 (1)	38.67 (3)
2	32.00 (2)	24.67 (1)	44.67 (3)	28.67 (2)	24.00 (1)	45.33 (3)	30.67 (2)	26.00 (1)	39.33 (3)
3	34.67 (2)	26.00 (1)	44.00 (3)	28.67 (2)	25.33 (1)	44.67 (3)	30.67 (2)	25.33 (1)	38.00 (3)
4	30.67 (2)	24.67 (1)	**45.33 (3)**	31.33 (2)	25.33 (1)	45.33 (3)	31.33 (2)	24.67 (1)	**42.00 (3)**
5	32.00 (2)	24.00 (1)	43.33 (3)	29.33 (2)	25.33 (1)	44.67 (3)	30.67 (2)	24.67 (1)	38.00 (3)
6	30.00 (2)	26.67 (1)	44.00 (3)	30.00 (2)	24.67 (1)	44.67 (3)	32.67 (2)	24.67 (1)	37.33 (3)
7	30.00 (2)	24.67 (1)	34.00 (3)	30.00 (2)	25.33 (1)	**46.00 (3)**	31.33 (2)	25.33 (1)	38.67 (3)
8	28.67 (2)	24.67 (1)	44.00 (3)	31.33 (2)	23.33 (1)	42.67 (3)	31.33 (2)	25.33 (1)	38.67 (3)
9	34.67 (2)	26.00 (1)	44.00 (3)	32.00 (2)	25.33 (1)	42.67 (3)	31.33 (2)	24.67 (1)	37.33 (3)
10	34.00 (2)	25.33 (1)	43.33 (3)	29.33 (2)	25.33 (1)	43.33 (3)	31.33 (2)	25.33 (1)	39.33 (3)
Average Accuracy	32.07	25.20	**43.00**	30.20	24.93	**44.20**	31.27	25.33	**38.73**
Shapiro-Wilk p-value	0.25135	0.39056	**0.00001**	0.20251	**0.00020**	0.10123	**0.00276**	**0.00660**	**0.03987**

The weighting schemes are implemented in the image classification task. Fig. 4 shows the flow of the classification of the images. The images are split into training and testing images. Keypoints are extracted by using Scale Invariant Feature Transform (SIFT) for each training and testing image. The extracted keypoints of SIFT are robust to changes in viewpoint, illumination and affine distortion [15]. The superiority of SIFT over the other keypoint extraction methods has been shown by Mikolajczyk and Schmid [16] by comparing ten different keypoint extraction methods and the result shows that SIFT performs best.

In the training process, k-means algorithm is used to group the keypoints and to generate a set of centroids. These centroids are used as visual words to construct the histogram or feature vector of each image. Therefore, the number of visual words is similar to the number of clusters or centroids. The distance between each keypoint and each visual word was then measured. The keypoint which has the minimum distance with the visual word is assigned. The frequency of keypoints in each visual word is counted to construct the histogram of visual words.

TABLE VI: Number of keypoints in the training images

	Number of Keypoints	Number of Training Images	Avg. Number of Keypoints
Coil-100	21,566	300	71.89
Caltech-101	115,984	250	463.94
Caltech-256	317,697	250	1,270.79

In the developed BoVW model, the histogram of visual words is constructed from the training images, and not from

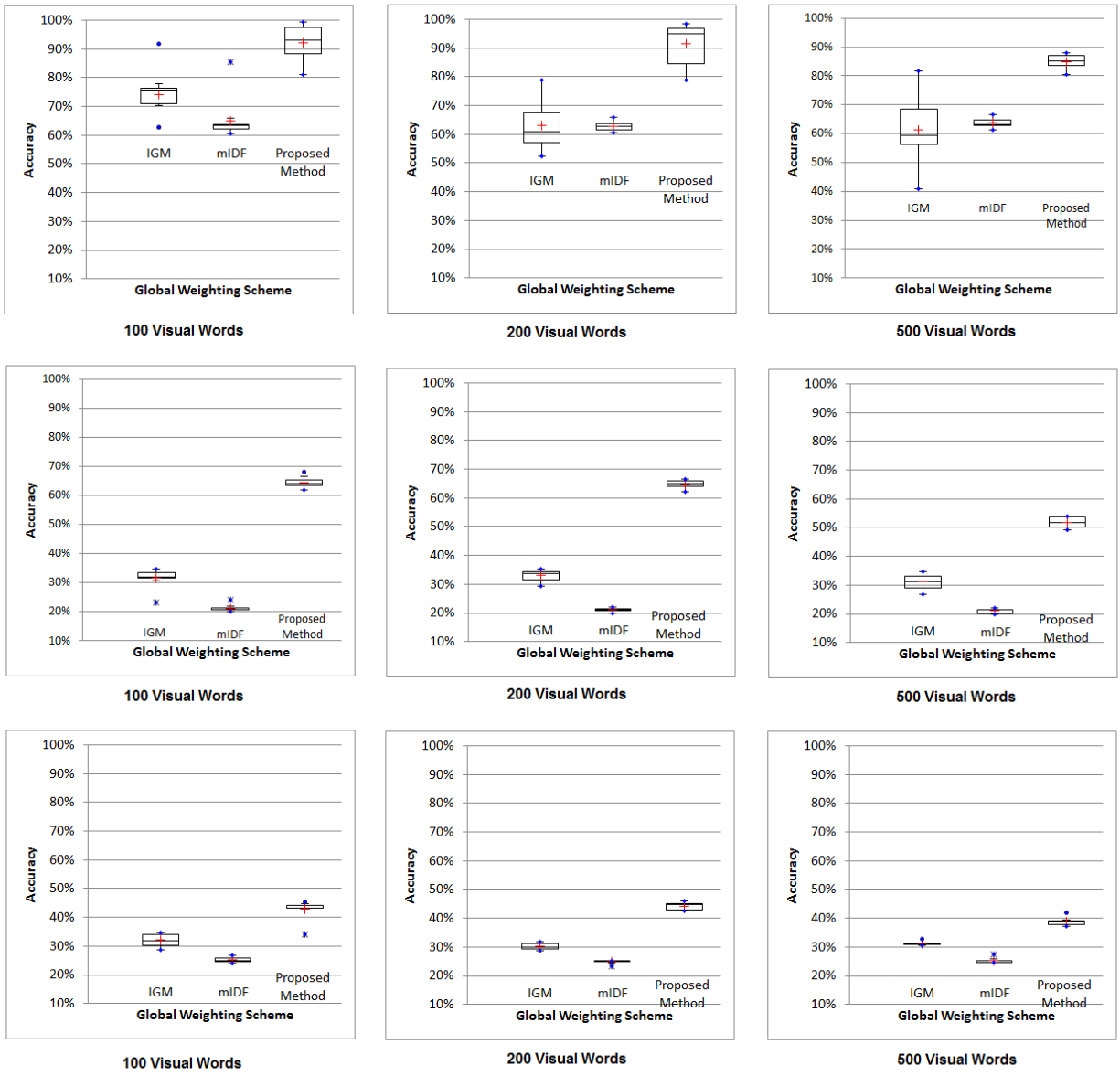

Fig. 5: The boxplots of classification accuracy on (First Row) Coil-100 dataset, (Second Row) Caltech-101 dataset, and (Third Row) Caltech-256 dataset

TABLE VII: Friedman test and average ranks of global weighting schemes to be used in Nemenyi test

Dataset	Visual Words	χ_F^2	Average Ranks		
			IGM	mIDF	Proposed Method
Coil-100	100	19.05	1.95	1.05	3
	200	21.4	1.7	1.5	3
	500	15.8	1.3	1.7	3
Caltech-101	100	20	2	1	3
	200	20	2	1	3
	500	20	2	1	3
Caltech-256	100	20	2	1	3
	200	20	2	1	3
	500	20	2	1	3

the total images, which is similar to the training process by Zhuo *et al.* [17]. Based on the histogram of visual words in the training images, the global weight of each visual words is computed by using the global weighting scheme. Then, the global weight is used to update the histogram of visual words in both training and testing images. These histograms are able to be used as input for the classifier. Support vector machine (SVM) is proposed as the classifier since this algorithm is popular in image retrieval [18]. In the process of classification, we apply LibSVM with the standard parameter in the Rapidminer.

Global weighting schemes are applied to the histogram of visual words to improve the performance of classification. IGM and mIDF as the newly current global weighting schemes are compared to our proposed weighting scheme. In the experiment, we set some parameters to analyze these weighting schemes. We set the number of visual words into 100, 200, and 500 visual words. These numbers are based on the previous work in [1] [14] [17] [19] that use hundreds visual words in their experiments. We also set $\lambda_1 = 7$ for the parameter of IGM as shown in the (4). This value is suggested by Chen *et al.* [4], since the number produces

TABLE VIII: Critical value in the Chi-Square table

degree of freedom	2	3	4	5
$q_{0.05}$	5.991	7.814	9.487	11.070
$q_{0.1}$	4.605	6.251	7.779	9.236

TABLE IX: Critical value for the Nemenyi test

#methods	2	3	4	5
$q_{0.05}$	1.960	2.343	2.569	2.728
$q_{0.1}$	1.645	2.052	2.291	2.459

the optimal performance in their experiments. The process of image classification is divided into feature extraction and classification. The feature extraction is conducted by using Matlab, such as keypoint extraction, keypoint clustering, generation of visual words histogram and global weighting scheme. Since the centroid initialization of k-means is random; we initialize the centroids ten times. The other task such as classification task is conducted by using Rapidminer. All of the experiments are executed on a system with the hardware specification: Intel i7 and 8GB of RAM.

B. Evaluation Methods

To measure the performance of global weighting schemes, we use accuracy, which is the total number of correctly classified testing images divided by the total number of testing images. We also use statistical analysis such as Friedman test [20] and Nemenyi post-hoc test [21] to evaluate the significant difference of accuracies among IGM, mIDF, and our proposed global weighting scheme. These statistical analysis tests are suggested by Demsar [22] to identify the significant difference between several classifiers on several datasets. In our case, we use these statistical analysis tests to identify the significant difference between global weighting schemes on ten times centroid initialization, thus we collect the average accuracy of each global weighting scheme. Friedman test is used to computes the average rank of global weighting scheme. Average rank of global weighting scheme is calculated using (14).

$$R_i = \frac{1}{N}\sum_j r_j^i \tag{14}$$

Here, r_j^i is the rank of the i-th of k number of global weighting schemes on j-th of N number of centroid initializations.

The Friedman statistics χ_F^2 is used to measure whether there is a statistical difference between the global weighting schemes or not [23]. Followed by Nemenyi test to identify which global weighting scheme perform best.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_i^k R_i^2 - \frac{k(k+1)^2}{4}\right] \tag{15}$$

The χ_F^2 is calculated by using (15). If the value of χ_F^2 is greater than the critical value in the Chi-Square table, null hypothesis can be rejected means that there is a significant difference among the global weighting schemes. To look up the critical value in the Chi-Square table, we need to specify the significance levels α and calculate the degree of freedom $df = k - 1$. Here, k is the number of groups or global weighting schemes.

The Nemenyi post-hoc test is calculated if the null hypothesis is rejected. It uses the critical difference to evaluate the ranks of global weighting schemes. The critical difference is obtained from (16).

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{16}$$

Here, q_α is the critical value. If the difference of mean ranks between two global weighting schemes is greater than the value of CD, it means that there is a significant difference on their performance.

C. Experiment Result

The accuracy of image classification on the Coil-100, Caltech-101, and Caltech-256 datasets using global weighting schemes are shown in Tables III-V respectively. Based on the average accuracy, the performance of the proposed method is better than the other global weighting schemes for any amount of visual words on all datasets. The tables also show the ranks (in the bracket) of each global weighting scheme on each centroid initialization. The highest rank belongs to the best (highest accuracy) global weighting scheme [23]. The ranks can be the average if they have the same accuracies. For example, in the 8th order centroid initialization (see Table III), the accuracy of the proposed method is 97.73%. Meanwhile, IGM and mIDF have the same accuracy 62.88%. Since the proposed method has the highest accuracy, it is ranked 3rd. However, IGM and mIDF have the same average rank $\frac{1+2}{2} = 1.5$. These ranks are used in the Nemenyi post-hoc test.

In Tables III-V, a large number of visual words seem to slightly decrease the performance of classification. It is shown in the boxplots diagrams (see Fig. 5) which visualizes the distribution of accuracies from Tables III-V. A boxplot split into quartiles. The central line indicates the median or second quartile (Q2). The box surrounds the median indicates the interquartile range (IQR), the lower and upper bounds of IQR are the first quartile (Q1) and third quartile (Q3). In the comparison among datasets, the accuracies of the three global weighting schemes on Caltech-256 dataset is the worst; the median value of accuracies is less than 50%. Low accuracies on Caltech-256 dataset due to the large number of generated keypoints. Based on Table VI, Caltech-256 produces the largest number of keypoints, with a total of 317,697 keypoints, followed by Caltech-101 (115,984 keypoints) and Coil-100 (21,566 keypoints). A large number of keypoints on Caltech-256 also lead to time-consuming for visual word generation. In Caltech-101 dataset, the proposed weighting scheme performs better than the other weighting schemes. The median value of accuracies of the proposed weighting scheme is more than 50%. It also happens on Coil-100 dataset, the performance of the proposed weighting scheme is better than the others. The median value of accuracies of the proposed weighting scheme is higher than 80% accurate.

In the boxplot diagram, the small size of the box reveals the consistency of the performance of the methods. In the Caltech-101 and Caltech-256, most weighting schemes have the consistent performance of accuracy. However, the accuracies of IGM and the proposed weighting scheme in Coil-100 dataset seem to generate inconsistent performance.

TABLE X: The difference of average ranks

Dataset	Visual Words	IGM vs mIDF	IGM vs Proposed Method	mIDF vs Proposed Method
Coil-100	100	0.9	**1.05**	**1.95**
	200	0.2	**1.3**	**1.5**
	500	0.4	**1.7**	**1.3**
Caltech-101	100	1	1	**2**
	200	1	1	**2**
	500	1	1	**2**
Caltech-256	100	1	1	**2**
	200	1	1	**2**
	500	1	1	**2**

TABLE XI: Confusion matrix of the proposed method on Coil-100 dataset with 99.24% accuracy rate

		Actual						
		Obj2	Obj5	Obj31	Obj35	Obj41	Obj59	class precision
Pred. Class	Obj2	22	0	0	0	**1**	0	95.65%
	Obj5	0	22	0	0	0	0	100%
	Obj31	0	0	22	0	0	0	100%
	Obj35	0	0	0	22	0	0	100%
	Obj41	0	0	0	0	21	0	100%
	Obj59	0	0	0	0	0	22	100%
	class recall	100%	100%	100%	100%	95.45%	100%	

TABLE XII: Confusion matrix of the proposed method on Caltech-101 dataset with 68% accuracy rate

		Actual Class					
		Airplanes	Butterfly	Motorbikes	Starfish	Watch	class precision
Pred. Class	Airplanes	27	5	1	0	6	69.23%
	Butterfly	2	13	2	3	**9**	44.83%
	Motorbikes	1	0	26	2	0	89.66%
	Starfish	0	**12**	0	23	2	62.16%
	Watch	0	0	1	2	13	81.25%
	class recall	90.00%	43.33%	86.67%	76.67%	43.33%	

TABLE XIII: Confusion matrix of the proposed method on Caltech-256 dataset with 45.33% accuracy rate

		Actual Class					
		Butterfly	Calculator	Camel	Elephant	Necktie	class precision
Pred. Class	Butterfly	20	3	**18**	**15**	**10**	30.30%
	Calculator	1	19	2	1	2	76.00%
	Camel	5	0	3	2	0	30.00%
	Elephant	3	4	6	9	1	39.13%
	Necktie	1	4	1	3	17	65.38%
	class recall	66.67%	63.33%	10.00%	30.00%	56.67%	

mIDF shows the consistent results in all datasets, although the median value of mIDF still smaller than the median values of other weighting schemes.

Based on the results of the Shapiro-Wilk test, some of accuracy distributions are not normal (see Tables III-V for details). The Shapiro-Wilk p-value is less than the confidence level 0.05 highlighted in bold. Therefore, the non-parametric test, i.e. Friedman test is appropriate to know whether the accuracy is significantly different or not. In the statistic analysis, the non-parametric test is appropriate for the small size of the samples [24]. The Friedman test and average rank from Tables III-V are shown in Table VII. In the Coil-100 dataset and 100-visual words, the value of χ_F^2 (15) is

$$\chi_F^2 = \frac{12\times 10}{3(3+1)}\left[(1.95)^2+(1.05)^2+(3)^2-\frac{3(3+1)^2}{4}\right] = 19.05$$

For significant level $\alpha = 0.05$, $k = 3$, and $df = 3 - 1 = 2$, the critical value in Chi-Square table is 5.99 (see Table VIII). Therefore, the value of χ_F^2 is higher than the critical value. It means that the null hypothesis is rejected or there is a significant difference among the accuracies of the three global weighting schemes.

For further analysis, the Nemenyi post-hoc test is performed to decide the significant difference between the global weighting schemes. For the significance level $\alpha = 0.05$ and $k = 3$, the critical value for the Nemenyi test q_α is 2.343 (see Table IX). Therefore, the critical difference of the Nemenyi test (16) is

$$CD = 2.343\sqrt{\frac{3(3+1)}{6\times 10}} = 1.0478$$

To check the performance of the global weighting schemes is significantly different or not, the difference of the average rank of the global weighting schemes is calculated (see Table X). It is highlighted in bold if the value is larger than CD, it means that the performance of the global weighting schemes is significantly different. Based on the table, the accuracies of the proposed method outperform mIDF significantly on all datasets. As well as the accuracies of the proposed method outperform IGM significantly but only on the Coil-100 dataset. While the accuracies between

IGM and mIDF are statistically insignificant, hence they have the same performance.

Tables XI-XIII display the confusion matrix of the proposed method on the Coil-100, Caltech-101, and Caltech-256 datasets. The proposed method produces the highest accuracy of 99.24% on the Coil-100 dataset. A minor confusion only occurs in the class *Obj41* and class *Obj2*. On the Caltech-101 dataset, the proposed method only achieves the accuracy of 68%. The major confusions occur between class *Starfish* and class *Butterfly*, or class *Watch* and class *Butterfly*. Meanwhile, on the Caltech-256 dataset, the proposed method only achieves the accuracy of 45.33%. Many major confusions are made, i.e. between the class *Camel* and class *Butterfly*, or class *Elephant* and class *Butterfly*, or class *Necktie* and class *Butterfly*.

V. Conclusions

This study presents the global weighting scheme based on intra-class and inter-class term distributions. The main issue addressed in this study is DF based global weighting schemes do not concern with the distribution of TF. The proposed method is compared to the state-of-the-art baseline methods, i.e. IGM and mIDF. The methods are tested on BoVW based image classification. Three benchmark datasets, i.e. Coil-100, Caltech-101, and Caltech-256 are used to evaluate the global weighting schemes. The statistical analysis Friedman test and Nemenyi post-hoc test are applied to decide which method performs best with a significant difference in classification accuracy.

Based on this research work, the proposed global weighting scheme outperforms the baseline methods. The accuracy is significantly different on the three datasets. A large amount of extracted keypoints on the dataset give an impact on the accuracy. In Caltech-256 dataset which produces large amount of keypoints than the two datasets, the three global weighting schemes produce low accuracy. In the next work, we can apply keypoint selection method to reduce the number of keypoints but still maintaining the high accuracy. Moreover, keypoint selection will speed up the process of visual word generation.

References

[1] A. Alfanindya, N. Hashim, and C. Eswaran, "Content based image retrieval and classification using speeded-up robust features (SURF) and grouped bag-of-visual-words (GBoVW)," in *International Conference on Technology, Informatics, Management, Engineering and Environment*, 2013, pp. 77–82.

[2] C.-H. Lee, F. Gutierrez, and D. Dou, "Calculating feature weights in naive bayes with kullback-leibler measure," in *11th IEEE International Conference on Data Mining*, 2011, pp. 1146–1151.

[3] C.-J. Tu, L.-Y. Chuang, J.-Y. Chang, and C.-H. Yang, "Feature selection using PSO-SVM," *IAENG International Journal of Computer Science*, vol. 33, no. 1, pp. 111–116, 2007.

[4] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. 66, pp. 245–260, 2016.

[5] S. Plansangket and J. Q. Gan, "A new term weighting scheme based on class specific document frequency for document representation and classification," in *Proceeding of 7th Computer Science and Electronic Engineering Conference (CEEC)*, 2015, pp. 5–8.

[6] M. Lan, C.-L. Tan, and H.-B. Low, "Proposing a new term weighting scheme for text categorization," in *Proceedings of the 21st national conference on artificial intelligence*, 2006, pp. 763–768.

[7] J. Gautam and E. Kumar, "An integrated and improved approach to terms weighting in text classification," *International Journal of Computer Science Issues*, vol. 10, pp. 245–260, 2013.

[8] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma, O. Krejcar, and H. Fujita, "Modified frequency-based term weighting schemes for text classification," *Applied Soft Computing*, vol. 58, pp. 193–206, 2017.

[9] H. Zhou, J. Guo, and Y. Wang, "A feature selection approach based on term distributions," *SpringerPlus*, vol. 5, pp. 245–260, 2016.

[10] M. Okawa, "Offline signature verification based on bag-of-visual words model using KAZE features and weighting schemes," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops Offline*, 2016, pp. 252–258.

[11] J. Geusebroek, G. Burghouts, and A. Smeulders, "A feature selection approach based on term distributions," *The Amsterdam library of object images*, vol. 61, pp. 103–112, 2005.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, pp. 59–70, 2007.

[13] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset (technical report 7694)," in *California Institute of Technology*, 2007.

[14] R. Wang, K. Ding, J. Yang, and L. Xue, "A novel method for image classification based on bag of visual words," *Journal of Visual Communication and Image Representation*, vol. 40, pp. 24–33, 2016.

[15] Y. Xie, S. Jiang, and Q. Huang, "Weighted visual vocabulary to balance the descriptive ability on general dataset," *Neurocomputing*, vol. 119, pp. 478–488, 2013.

[16] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 110, pp. 1615–1630, 2005.

[17] L. Zhuo, Z. Geng, J. Zhang, and X. g. Li, "ORB feature based web pornographic image recognition," *Neurocomputing*, vol. 173, pp. 511–517, 2016.

[18] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *International Scholarly Research Network ISRN Artificial Intelligence*, vol. 2012, pp. 1–19, 2012.

[19] W. Bouachir, M. Kardouchi, and N. Belacel, "Improving bag of visual words image retrieval: A fuzzy weighting scheme for efficient indexation," in *5th International Conference on Signal Image Technology and Internet Based Systems*, 2009, pp. 215–220.

[20] M. Friedman, "A comparison of alternative tests of significance for the problem of *m* rankings," *Ann. Math. Stat.*, vol. 11, pp. 86–92, 1940.

[21] P. Nemenyi, "Distribution-free multiple comparisons," *Ph.D. Thesis Princeton University*, 1963.

[22] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[23] A. Kaur and I. Kaur, "An empirical evaluation of classification algorithms for fault prediction in open source projects," *Journal of King Saud University*, pp. 1–16, 2016.

[24] N. Settouti, M. E. A. Bechar, and M. A. Chikh, "Statistical comparisons of the top 10 algorithms in data mining for classification task," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, pp. 46–51, 2016.