Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052011 doi:10.1088/1757-899X/546/5/052011

Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification

Arfiani[1], Zuherman Rustam[1], Jacub Pandelaki[2], Arga Siahaan[2]

[1]Department of Mathematics, University of Indonesia, Depok 16424, Indonesia
[2]Department of Radiology, RSUPN dr. Cipto Mangunkusumo, Jakarta 10430, Indonesia

rustam@ui.ac.id

Abstract. Acute sinusitis is an inflammation of the sinus which causes the cavity around the sinus to swells due to accumulated mucus. It makes the patient experience difficulty in breathing through the nose. Generally, it is caused by the common cold, and in most cases, the patient recovers within seven to ten days. However, persistent acute sinusitis can cause severe infections and other complications. Therefore, it requires timely detection and more accurate method of classification. Many techniques have been used to classify acute sinusitis but, in this study, the machine learning methods which includes Kernel Spherical K-Means (KSPKM) and Support Vector Machine (SVM) was applied. SPKM is the application of K-Means, in this research, it was modified by changing the inner product with kernel function to ensure linear data separation on higher dimensions for the maximization of SPKM performance. The SVM is a binary classification method that helps to create a model with good generalization ability. We used CT scan result data from RSCM, Central Jakarta. Simulations were performed with different percentage of training data. The results were compared in terms of Accuracy and Running Time. The score showed that the performance of KSPKM attained an accuracy rate of 97%, while SVM reached 90%.

1. Introduction

Sinusitis is an inflammation of the sinus wall [1]. It is a small cavity that is interconnected through the airways in the skull bones [1]. Sinus is located on the back of the forehead bone, inside the cheekbone structure, on both sides of the nose, and behind the eye [1]. The sinus produces mucus which is useful for filtering and cleansing bacteria and other particles in the inhaled air [1]. Furthermore, the sinus also ensures proper regulation of temperature and humidity of the air entering the lungs [1].

In line with the duration, sinusitis is divided into four types, namely acute sinusitis, sub-acute sinusitis, chronic sinusitis, and recurrent sinusitis [1]. The focus of this research is on acute sinusitis. It is the most common type of sinusitis with duration of 2-4 weeks. It is caused by common cold [1]. It causes the sinus cavity to become inflamed and swollen due to the accumulated mucus [2]. The presence of accumulated mucus usually leads to difficulty breathing through the nose [2]. The swelling of the sinus cavity causes headache [2]. In most cases, home remedies are the most effective solution for treating acute sinusitis [2]. However, sinusitis occurs many times and lasts longer causing serious infections and complications such as meningitis, impaired vision, loss of sense of smell, and spread of infection in the sinuses to the bones or skin [1].

Diagnose a patient of acute sinusitis, an examination from the radiology agency is needed. A common method of diagnosis employed during examinations is an imaging test using Computed Tomography Scanning (CT scan) or Magnetic Resonance Imaging (MRI). They are used to obtain a detailed description of the sinuses area and nose including the condition of inflammation or blockage [1]. Most acute sinusitis caused by a virus can heal on their own [1]. However, if it caused by a bacterial infection, the use of antibiotics is required to prevent the spread of the infection [1].

In the health sector, many methods have been carried out to diagnose acute sinusitis. However, this study used computational techniques by applying machine learning. A proposal was made for Kernel Spherical K-Means (KSPKM) and Support Vector Machine (SVM). SPKM is a development method of K-Means for clustering. The SPKM modified by converting the inner products to the kernel function. SVM is known as a binary classification method for maximizing the result of classification. By using both methods, it is expected to help the health sector to be able to diagnose acute sinusitis more efficiently.

Previous research on the classification of sinusitis, including acute sinusitis, has been performed with various methods such as Binary Logistic Regression [3], Automatic Segmentation [4], Imaging Features [5, 6, 7], and Anatomical Based Classification [8]. Then, the SPKM method has been used for Simulation Modeling [9] and Text Clustering [10]. The SVM method has been used for Classification of Schizophrenia [11], Classification of Cancer Data [12, 13], and Classification of Hyper spectral Imagery [14], Traffic Incident Detection [15], Intrusion Detection System [16, 17, 18, 19], Face Recognition [20, 21], Predicting Bank Failures [22], and Evaluating the Internationalization Success of Companies [23].

2. Methods

2.1. Data

The data used in this study obtained from the results of the CT scan of the Department of Radiology dr. RSUPN Cipto Mangkunsumo (RSCM), Central Jakarta, which consists of four features, and they include: Gender, Age of Patient, Air Normal Cavity, HU (Hounsfield Unit) for acute sinusitis, and there is one Prediction Class. The data includes 200 observations which are 100 data labeled acute sinusitis and 100 data with acute non-sinusitis labels. The display of data shown in the following table::

Table 1. Results of CT Scan Acute Sinusitis Data from Department of Radiology, RSCM.

No. Patient	Gender	Age	Air Normal Cavity	HU	Prediction Class
1.	M	47	-1004	114	0
2.	F	45	-663	93	0
3.	M	20	-890	22	1
4.	F	15	-968	17	1
5.	F	18	-1021	36	1

2.2. Kernel Function

The kernel function is given as follows:

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j'}) = \langle \boldsymbol{\varphi}(\boldsymbol{x_i}), \boldsymbol{\varphi}(\boldsymbol{x_j}) \rangle \quad (1)$$

where $\varphi(\boldsymbol{x})$ is a function that maps $\boldsymbol{x} \in R^n$ to the feature space. Every time $\langle \varphi(\boldsymbol{x_i}), \varphi(\boldsymbol{x_j}) \rangle$ appears in the classification algorithm, it can be replaced with $K(\boldsymbol{x_i}, \boldsymbol{x_j'})$ [24]. By using kernel functions, it is

expected that data can be separated linearly on higher dimensions. In this study, the kernel Radial Basis Function (RBF) was used with the following formula [25]:

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j'}) = \mathbf{exp}\left(-\boldsymbol{\gamma}\left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|^{\mathbf{2}}\right) \quad (2)$$

2.3. Spherical K-Means

Spherical K-Means (SPKM) is one of the clustering methods used to group data. According to [9], SPKM algorithm is found by Shi Zhong and is a K-Means algorithm with cosine similarity or measuring the equation between vector data through the inner product. This algorithm aims to maximize objective functions [10]:

$$L = \sum_{x} \boldsymbol{x}^T V_{p(x)}$$

where:

- $\boldsymbol{x}$ = Data in vector.
- $p(x) = \arg max_p\, \boldsymbol{x}^T V_p$.
- V_p = Center cluster on cluster to p.

The Spherical K-Means algorithm is as follows:

Input:

- Training data to be used, i.e. $\boldsymbol{x} = (x_1, x_2, \dots, x_N)$, where N is the number of the data.
- Number of cluster (p).
- Initial cluster center (V), obtained from the mean of each cluster.
- Tolerance ε.

Output: Data vectors with cluster labels $(y_n), n = 1,2, \dots, N$, where $y_n \in \{1,2, \dots, p\}$.

Steps:

1. Determine clusters for each vector of data $\boldsymbol{x}_n, n = 1,2, \dots, N$.

$$\hat{y}_n = \arg max_p\, \boldsymbol{x}^T V_p$$

2. While $x_p = \{\boldsymbol{x}_n | \hat{y}_n = p\}$, determine the estimation of the cluster center with:

$$V_p = \frac{\sum_{x \in x_p} \boldsymbol{x}}{\left\|\sum_{x \in x_p} \boldsymbol{x}\right\|}$$

3. If $\hat{y}_n = y_n$ or $|V_{new} - V_{old}| < \varepsilon$ then the algorithm stops. If not, repeat from step 1.

2.4. Kernel Spherical K-Means

Kernel Spherical K-Means (KSPKM) is a SPKM algorithm that modifies the inner product into the kernel function. The KSPKM aims to maximize objective functions as follows:

$$\sum_{\boldsymbol{x}} \boldsymbol{K}\left(\boldsymbol{x}, \boldsymbol{V_{p(x)}}\right) \quad (3)$$

where:

- $\boldsymbol{x}$ is a vector of data.

- $p(x) = argmax_c \left(exp\left(-\frac{\|x-V_c\|^2}{2\sigma^2}\right)\right)$, is the closest cluster center for each vector.
- $V_c = \frac{\sum_{x \in X_c} x}{\|\sum_{x \in X_c} x\|}$, is the center of the cluster to p.

The following is an algorithm from KSPKM [10]:
Input: $\boldsymbol{x} = \{x_1, x_2, \dots, x_N\}$, where N is the number of the data.
Output: Vector of cluster centre $V = \{V_1, V_2, \dots, V_c\}$.
The steps in the KSPKM algorithm are as follows:

1. Determine:
   a. Number of the cluster c where $1 < c < n$.
   b. The cluster centre $V^0 = \{V_1, V_2, \dots, V_c\}$.
   c. Parameter σ.
   d. Tolerance ε.
2. For $t = 1$ to $t = T$, do 3 until 5.
3. Determine cluster for each vector of data $\boldsymbol{x}$.

$$\hat{\boldsymbol{p}}(\boldsymbol{x}) = \boldsymbol{argmax_c}\left(\boldsymbol{exp}\left(-\frac{\|\boldsymbol{x}-\boldsymbol{V_c}\|^2}{\boldsymbol{2\sigma^2}}\right)\right) \quad (4)$$

4. If $x_c = \{\boldsymbol{x}_i | \hat{p}(\boldsymbol{x}) = p(x)\}$ then enter x_c into cluster to C_p, determine the new centre cluster with:

$$\boldsymbol{V_c} = \frac{\sum_{\boldsymbol{x} \in \boldsymbol{C_p}} \boldsymbol{x}}{\left\|\sum_{\boldsymbol{x} \in \boldsymbol{C_p}} \boldsymbol{x}\right\|} \quad (5)$$

If $\|V^t - V^{t-1}\| < \varepsilon$, then the algorithm stops. If not, then return to step 2.

2.5. Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression introduced by Vapnik in the late 1990s. SVM is related to Structural Risk Minimization (SRM). Initially, SVM was used for binary classification, but now it could be used for multiclass classification. SVM takes the form of mapping input space into higher dimensional space to support nonlinear classification problem where the maximum separation of the hyperplane is constructed. The hyperplane is a linear pattern whose maximum margin provides maximum separation between decision classes.

2.5.1. Characteristic of SVM [26]

Given dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ where N is the number of samples, $\boldsymbol{x}_i \in R^D$ is a feature vectors from sample-i, where D is the number of feature (dimension), and y_i is a class label. For two class classification problem $y_i \in \{-1, +1\}$, but for multiclass classification problem $y_i \in \{1,2, \dots, k\}$ where k is the number of class. The main goal of SVM is to determine the best hyperplane:

$$\boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0} \quad (6)$$

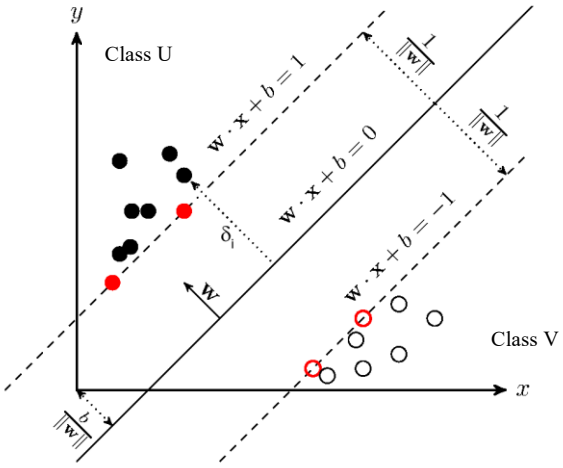

Figure 1. SVM is trying to determine the best hyperplane to separate two classes, class U and V.

The problem of SVM optimization can be summarized as follows:

$$\min \frac{1}{2}\|w\|^2 \tag{7}$$

$$s.t.\, y_i\left(w^T \cdot x_i + b\right) \geq 1, \forall i = 1, \ldots, N \tag{8}$$

Objective function (7) to find $w \in R^n$ dan $b \in R^n$ subject to (8), where w is the weights and b is bias. By completing the equation above, the formula w and b are obtained as follows:

$$w = \sum_{I=1}^{N} a_i y_i x_i \tag{9}$$

$$b = \frac{1}{N_s} \sum_{i \in s} \left(y_i - \sum_{m \in s} a_m y_m x_m\right) \tag{10}$$

and the decision function as follows:

$$f(x) = w \cdot x + b \tag{11}$$

which could maximize the margins.

2.6. Parameter Optimization

In this study, several parameters were optimized with the grid search method. It finds one by one combination of parameters that produces the optimum model [27]. The optimized parameters are:

- $\sigma = 0.1$ and
- $C = 1000$

2.7. Model Performance Validation

To validate the performance of the model, the Hold-Out Validation method was applied. Here, data is separated into two parts. They are; training and testing data. Model evaluation is obtained from testing data. Computationally, this method is easy and fast [28].

In this study, Hold-Out Validation is used with a different percentage of data. To overcome the weaknesses of Hold-Out Validation is very dependent on the data used for training and testing [27]. Simulations were performed nine times with different percentages of the data used.

2.8. Model Performance Validation

In this study, a performance evaluation model is conducted by measuring accuracy and running time. Let TN, TP, FN, FP denote True Negative, True Positive, False Negative, and False Positive, respectively. The following formula below is used:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FN + FP}} \tag{12}$$

3. Experimental Results

This study used software MATLAB R2017a for KSPKM and Python 3.6 for SVM.

3.1. Acute Sinusitis Classification using KSPKM

The results are shown below:

Table 2. Results of Accuracy and Running Time Acute Sinusitis Classification using KSPKM with RBF Kernel, parameter: $\sigma = 0.1$.

Data Training (%)	Accuracy (%)	Running Time (s)
10	94.41	0.02
20	94.97	0.03
30	93.53	0.02
40	94.96	0.03
50	97.00	0.02
60	94.94	0.03
70	94.92	0.03
80	94.87	0.03
90	89.47	0.03

According to Table 2, the data training at 50% with 97% accuracy and running time of 0.02 seconds recorded as the best accuracy. Whereas, the lowest accuracy result was recorded at data training 90% at 89.47% and running time for 0.03 seconds.

3.2. Acute Sinusitis Classification using SVM

Table 3 shows that data training attained best accuracy at 90% with 90% accuracy and running time 0.13 seconds. Whereas, the lowest accuracy result was recorded at data training 10% at 54.44% and running time for 0.07 seconds.

Table 3. Results of Accuracy and Running Time Acute Sinusitis Classification using SVM with RBF Kernel, parameter: $\sigma = 0.1$.

Data Training (%)	Accuracy (%)	Running Time (s)
10	54.44	0.07
20	69.37	0.06
30	63.57	0.07
40	69.17	0.07
50	70.00	0.08
60	76.25	0.08
70	81.67	0.1
80	80.00	0.1
90	90.00	0.13

The following Figures (Figures 2 and 3) showing the accuracy and running time of acute sinusitis classification with KSPKM and SVM:

Figure 2. Graph of Acute Sinusitis Classification using KSPKM with RBF Kernel, parameter: $\sigma = 0.1$.

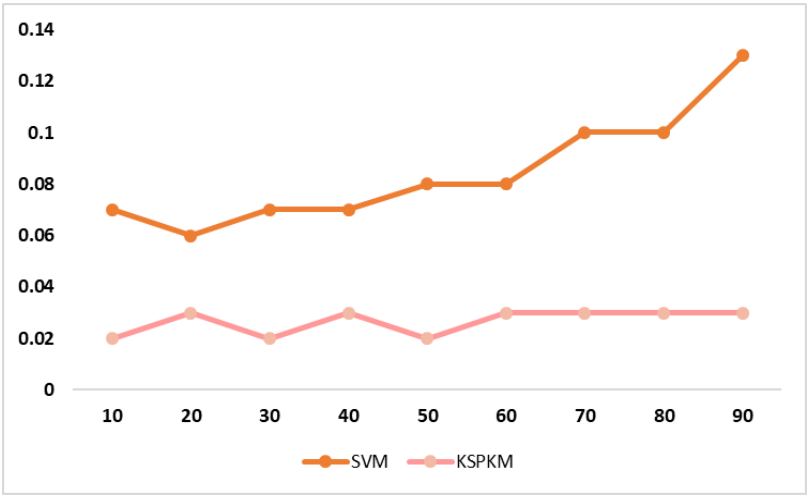

Figure 3. Graph of Acute Sinusitis Classification using SVM with RBF Kernel, parameter: $\sigma = 0.1$.

4. Discussion

In this paper, an examination is conducted on the comparison of KSPKM with SVM to classify acute sinusitis based on accuracy and running time. Figure 2-3 show the performance results of the two methods and a proposal was made using $\sigma = 0.1$ as the parameter which was determined by the grid search method as described previously. According to Figure 2, the results of the study indicate that the SVM produced the accuracy progressively based on the Hold-Out Validation value specified. Meanwhile, KSPKM produced very consistent accuracy. Since KSPKM has consistent accuracy in all conditions, then it is better than SVM. Fig. 3 shows the running time results from the two methods proposed. KSPKM required a shorter running time than SVM. This because of during the modification, the inner product was converted into the kernel function and used the RBF kernel with selected parameters. SVM produced lower accuracy and longer running time than KSPKM. Therefore, KSPKM is better than SVM for acute sinusitis classification. This cannot be generalized using the other data or

the other optimization parameters; consequently, the problems were limited in terms of the data used and the optimization parameters.

5. Conclusion

This research proposed the method of Kernel Spherical K-Means (KSPKM) and Support Vector Machine (SVM) to classify acute sinusitis. Simulation of the data is done by the KSPKM and SVM methods with the RBF kernel. For each simulation, the parameters are optimized with the grid search method, while the validation and evaluation performance of the model is conducted by the Hold-Out Validation method. According to the simulation that has been conducted using software, the result of the accuracy reached 97% and running time was 0.02 seconds with the KSPKM method. The SVM method achieved an accuracy of 90% with a running time of 0.13 seconds. These results show that the performance of KSPKM is better than SVM with the limitation of the problems provided, which includes the kernel parameters and acute sinusitis data in the form of CT scan results.

Acknowledgments

This research was financially supported by University of Indonesia, with PITTA B 2019 research grant scheme (ID number NKB-0688/UN2.R3.1/HKP.05.00/2019).

References

[1] Alodokter, accessed 4 February 2019, see https://www.alodokter.com/sinusitis.
[2] Mayo Clinic, accessed 4 February 2019, see https://www.mayoclinic.org/diseases-conditions/acute-sinusitis/symptoms-causes/syc-20351671.
[3] C. O. de Lima, K. L. Devita, L. R. B. Vasconcelos, M. do Prado, and C. N. Campos, "Correlation between Endodontic Infection and Periodontal Desease and Their Association with Chronic Sinusitis: A Clinical-tomographic Study", *American Association of Endodontists* (2017).
[4] J. Singh and A. S. Arora, "A Framework for Enchancing the Thermographic Evaluation on Characteristic Areas for Pranasal Sinusitis Detection", *Infrared Physics & Technology*, 85, 457-464 (2017).
[5] V. Velayudhan, Z. A. Chaudhry, W. R. K. Smoker, R. Shinder, and D. Reede, "Imaging of Intracranial and Orbital Complications of Sinusitis and Atypical Sinus Infection: What the Radiologist Needs to Know", *Current Problems in Diagnostic Radiology* (2017).
[6] K. J. T. Lakhan, "Sinus Headaches Sinusitis Versus Migraine", *Physician Assist Clin*. 3, 181–192 (2018).
[7] B. Wyler, W. K. Mallon, "Sinusitis Update", Emerg Med Clin N Am, 37, 41–54 (2019).
[8] A. P. Campbell, R. W. Bergmark, and R. Metson, "Orbital Complications of Acute Sinusitis", *Operative Techniques in Otolaryngology - Head and Neck Surgery* (2017).
[9] R. Duwairi and M. Abu-Rahmeh, "A Novel Approach for Initializing the Spherical K-Means Clustering Algorithm", *Simulation Modelling Practice and Theory*, 54, 49-63 (2015).
[10] S. Zhong, "Efficient Online Spherical K-Means Clustering", *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 18, 790-798 (2005).
[11] T. V. Rampisela and Z. Rustam, "Classification of Schizophrenia data using Support Vector Machine (SVM)", *Journal of Physics: Conference Series* 1108 (1), 012044 (2018).
[12] Z. Rustam, I. Primasari, and D. Widya, "Classification of Cancer Data Based on Support Vectors Machines with Feature Selection using Genetic Algorithm and Laplacian Score", *AIP Conference Proceedings*, 2023 (1), 020234 (2018).
[13] T. Nadira and Z. Rustam, "Classification of Cancer Data using Support Vector Machines with Feature Selection Method Based on Global Artificial Bee Colony", *AIP Conference Proceedings*, 2023 (1), 020205 (2018).
[14] Z. Chunhui, G. Bing, Z. Lejun, and W. Xiaoqing, "Classification of Hyperspectral Imagery Based on Spectral Gradient, SVM and Spatial Random Forest", *Infrared Physics and Technology* 95, 61–69 (2018).

[15] J. Xiao, "SVM and KNN Ensemble Learning for Traffic Incident Detection", *Physica A* 517, 29–35 (2019).
[16] Z. Rustam and D. Zahras, "Comparison Between Support Vector Machine and Fuzzy C-Means as Classifier for Intrusion Detection System", *Journal of Physics: Conference Series* 1028 (1), 012227 (2018).
[17] Z. Rustam and NPAA. Ariantari, "Comparison Between Support Vector Machine and Fuzzy Kernel C-Means as Classifier for Intrusion Detection System using Chi-Square Feature Selection", *AIP Conference Proceedings* 2020 (1), 020214 (2018).
[18] J. Maharani and Z. Rustam, "The Application of Multi-Class Support Vector Machines on Intrusion Detection System with the Feature Selection using Information Gain", *1st Annual International Conference on Mathematics, Science, and Education (ICoMSE)*, 218 (2017).
[19] Z. Rustam and N. Olivera, "Comparison of Fuzzy Robust Kernel C-Means and Support Vector Machines for Intrusion Detection System using Modified Kernel Nearest Neighbor Feature Selection", *AIP Conference Proceedings* 2023 (1), 020215 (2018).
[20] Z. Rustam and R. Faradina, "Face Recognition to Identify Look-Alike Faces using Support Vector Machine", *Journal of Physics: Conference Series* 1108 (1), 012071 (2018).
[21] Z. Rustam and A. A. Ruvita, "Apllication Support Vector Machine on Face Recognition for Gender Classification", *Journal of Physics: Conference Series* 1108 (1), 012067 (2018).
[22] Z. Rustam, F. Nadhifa, and M. Acar, "Comparison of SVM and FSVM for Preditcing Bank Failures using Chi-Square Feature Selection", *Journal of Physics: Conference Series* 1108 (1), 012115 (2018).
[23] Z. Rustam, F. Yaurita, and M. J. Segovia-Vergas, "Application of Support Vector Machines in Evaluating the Internationalization Success of Companies", *Journal of Physics: Conference Series* 1108 (1), 012038 (2018).
[24] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition* (Burlington: Elsevier Inc, 2011).
[25] H. Xue, Q. Yang, and S. Chen, *The Top Ten Algorithms in Data Mining* (New York: Chapman and Hall/CRC, 2009).
[26] N. Cristianini and J. S. Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, (Cambridge University Press, 2000). M.
[27] K. R. Jayadeva and S. Chandra, "Twin Support Vector Machines for Pattern Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5), 905-10 (2007).
[28] A. Zheng, *Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls*, (Sebastopol, CA: O'Reilly Media, Inc, 2015).