Predicting Popularity of Movie UsingSupport Vector Machines

Dwi Rantini, Rosyida Inas, dan Santi Wulan Purnami
Departemen Statistika, Fakultas Matematika, Komputasi dan Sains Data,
Institut Teknologi Sepuluh Nopember
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: *santi_wp@statistika.its.ac.id*

Abstrak— ***There are many movies performed, from low until high rating, which is the movie maybe popular or not popular. If many people watched that movie maybe it is popular, in other hand if a movie is watched by a little person so that movie can called as not popular movie. Popularity of movie can determined by several factors, such as likes, ratings, comments, etc. To determine popular or not popular of movie based on features, will use two classification methods that is logistic regression and Support Vector Machine (SVM). In this research, the data are Conventional and Social Media Movies Dataset 2014 and 2015. To get the best model and without ignoring the principle of parsimony, will do feature selection. The selected features are genre, sentiment, likes, and comments. That features will be used to classify the popularity of movies. This research used two classification methods namely logistic regression and Support Vector Machine (SVM). When used logistic regression, the accuracy is 77.29%, while used SVM the accuracy is 83.78%. Based on the accuracy of both methods, it is found that SVM gives the highest accuracy for CSM dataset. The highest accuracy is obtained from the SVM method using combination kernel between*** $C = 2^5$ ***and*** $\gamma = 2^{-5}$ ***with non-stratified holdout training-testing strategy.***

Keywords— Logistic Regression, Movie, Predicting Popularity, Support Vector Machines

I. INTRODUCTION

There are many users sharing their opinions and experiences via social media, there is aggregation of personal wisdom and different viewpoints. Such aggregation has limitations as viewpoints are subject to change with time. In a sense the social media prediction problem is paralleled by prediction of financial time series based on past history, which has its uses in trading. In general, if extracted and analysed properly, the data on social media can lead to useful predictions of certain human related events. Such prediction has great benefits in many realms, such as finance, product marketing and politics, which has attracted increasing number of researchers to this subject. Study of social media also provides insights on social dynamics and public health. A survey provides us perspective and is helpful for carrying out further research. Prediction of success in business has been of great interest [1]. To the economists and financial experts. With advent of data analytics, the prediction process has been made intelligent by considering the historical data and employing various data analytical techniques to infer the future events. Such studies have been performed in prediction of movies success as well where success and popularity is measured in terms of the Ratings (typically represented by a numeric number from 0-10) and Income. There have been a large number of studies reported in this domain due to reasons such as general interest of public in this popular medium of entertainment, non-requirement of domain experts as required in other domains such as medical and huge number of data freely available on Web resources such as IMDB1. Most of the studies performed for prediction of movies success use conventional attributes, collected from online movies databases. However, with advent of social media, public opinion has been harnessed about various events/entities from forums such as YouTube and Twitter. Similarly, for movies, social media websites have contributed a great amount to the popularity of movies. Now anyone can review, rate, comment or share their opinions about a movie online. Thus social media plays a vital role in predicting the success of a movie. Many researchers believe that one should consider the social factors along with the classical factors for this purpose. Among social media mediums, Twitter has gained remarkable popularity and usage lately. Thus making it a point of focus, for researchers to predict the movie success using sentiments or feedback collected via Twitter. However, most of the studies performed in this domain have shown that sentiments about movies are not determining factor (or among the top factors) in predicting the success of movie while calculating it before release [1]. There are many movies performed, from low rating until high rating, that movie maybe popular or not popular. If many people watched that movie maybe it is popular, in other hand if a movie is watched by a little person so that movie can called as not popular movie. Popularity of movie can determined by several factors, such as likes, ratings, comments, etc. To determine popular or not popular of movie based on features, will use two classification methods that is logistic regression and Support Vector Machine (SVM). Logistic regression is one of the most widely used techniques for classification two categories data purposes today, however more recently, new methodologies based on iterative calculations (algorithms) have emerged, e.g.,neural networks (NN) and machine learning, pure computational approaches have been seen as "black boxes" in which data sets are throw in and solutions are obtained, without knowing exactly what happens inside so that in turn, this limits their interpretation, thats why logistic regression is still being the favourite one among classification method [2]. Beside Binary Logistic Regression, there is a classification method which popular enough that is Support Vector Machine (SVM), SVM according to [3] is a classification and regression method that combines computational algorithms with theoretical results; these two characteristics gave it good reputation and have promoted its use in different areas. Since its appearance, SVM has been compared with other classification methods using real data.

II. LOGISTIC REGRESSION

A. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes) [4]. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, etc.) or 0 (FALSE, failure, etc.). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$\text{logit}(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k \qquad (1)$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \qquad (2)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. For CSM data, classification using logistic regression for original data, imputation with grand mean or mean of each class yield has accuracy equal to 77.0563%.

B. Feature Selection

Feature selection is referred to the process of obtaining a subset from an original feature set according to certain feature selection criterion, which selects the relevant features of the dataset. Feature selection technique can pre-process learning algorithms, and good feature selection results can improve learning accuracy, reduce learning time, and simplify learning results. Notably, feature selection and feature extraction are two ways to dimensionality reduction. Unlike feature selection, feature extraction usually needs to transform the original data to features with strong pattern recognition ability, where the original data can be regarded as features with weak recognition ability. In this research, two feature selection methods are used: filter method and wrapper method (forward and backward). While the method used for feature extraction is principal component analysis. Filter feature selection methods usually use evaluation criteria to enhance the correlation between the feature and the class label and to reduce correlation among features [5]. Wrapper models take the classification error or accuracy rate as the feature evaluation standard. The feature selection result is often produced simultaneously as that of the learning model because the learning method is included in feature selection. In comparison with the filter model, the wrapper model could achieve higher classification accuracy and tend to have a smaller subset size. The highest accuracy for classification in this case is 77.2944% (data with imputation grand mean and feature selection (using forward) with 4 features (genre, sentiment, likes, and comments).

III. SUPPORT VECTOR MACHINES

A. Support Vector Machine

Support Vector Machines (SVM) is a new algorithm of data mining technique, recently received increasing popularity in machine learning community [6]. Support vector machines (SVMs) are a set of new supervised learning methods used for binary classification. SVM utilizes an optimum linear separating hyperplane to separate two data sets in a feature space. This optimum hyperplane is produced by maximizing minimum margin between the two sets [7]. A subset of the data points which determine the location of the hyperplane are known as the support vectors. The support vector machine operates on two mathematical operations: (1) Nonlinear mapping of an input vector into a high-dimensional feature space that is hidden from both the input and output. (2) Construction of an optimal hyperplane for separating the features. For the two-class linearly separable problem in an n-dimensional feature space, the hyperplane can be described by

$$h(\boldsymbol{x}) = \boldsymbol{WTX} + \boldsymbol{b} = \boldsymbol{0} \qquad (3)$$

where $\boldsymbol{W}$ is the normal vector and $\boldsymbol{b}$ is the distance from the hyperplane to the origin. The hyperplane $h(\boldsymbol{x})$ is learned using a training data set $\{x_i, y_i\}, i = 1, ..., l$ where $x_i \in \boldsymbol{R}^n$, and $y_i \in \{+1, -1\}$. Note that the hyperplane $h(\boldsymbol{x})$ can classify the training samples correctly, given the following conditions: if $y_i = +1, h(\boldsymbol{x}) \geq 1$; and if $y_i = -1, h(\boldsymbol{x}) \leq -1$. The points that make $h(\boldsymbol{x}) = +1$ or -1 are known as the support vector. The goal of the SVM is to find a hyperplane in order to maximize the distance between the hyperplane and the training data points which are closest to the hyperplane. The problem can be converted into the following equivalent convex quadratic problem.

$$\begin{gathered} \min_{W,b} \frac{1}{2} \|\boldsymbol{W}\|^2 \\ \text{s.t. } y_i \left(\boldsymbol{W}^T x_i + b\right) \geq 1, \quad i = 1, 2, ..., N \end{gathered} \qquad (4)$$

Using Lagrange multipliers is written as:

$$\begin{gathered} \max_{\alpha} \left\{ \Sigma_{i=1}^N \alpha_i - \frac{1}{2} \Sigma_{i=1}^N \Sigma_{j=1}^N y_i y_j \alpha_i \alpha_j \left\langle x_i, x_j \right\rangle \right\} \\ s.t. \Sigma_{j=1}^N \alpha_i y_i = 0 \\ \alpha_i \geq 0, i = 1, 2, ..., N \end{gathered} \qquad (5)$$

where the original problem is represented by $w = \Sigma_{i=1}^N \alpha_i y_i x_i$ and $0 = \Sigma_{i=1}^N \alpha_i y_i$. Therefore, having obtained Lagrange multipliers α, we can determine both w and b. Among all classification algorithms SVM is strong because of its simple structure and it requires less number of features. SVM is a structural risk minimization classifier algorithm derived from statistical learning theory by Vladimir Vapnik and his colleagues in 1992. Support Vector Machines were first introduced to solve the pattern classification and regression problems.

B. Kernel of Support Vector Machine

The major advantages of the SVM are as follows: first, SVM has only two experimental parameters, namely the upper bound and the kernel parameter. Obtaining an optimal combination of parameters that produce the best prediction performance is an easier task [8] Second, the SVM guarantees the existence of a unique, optimal, and global solution because SVM training is equivalent to solving a linearly constrained QP [8]. Third, the SVM implements the SRM principle that is known to have good generalization performance, Finally, the SVM can be constructed with small training data sets to obtain prediction performance [9]. In a dichotomous classification setting, that is, to predict one or the other class from a combined set of two classes (e.g., popular and not-popular), the development of a support vector machines model, as with other models of prediction, begins with the design of a training sample is the input information for the training object *i* on a set of *m* independent variables and corresponding outcome (dependent variable). Training vectors xi are mapped into a higher (may be infinite) dimensional space by the function φ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space. C > 0 is the penalty parameter of the error term [10]. Furthermore, $\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \varphi^T(\mathrm{x})\varphi(\mathrm{x}_i)$ is called the kernel function. There are many kernel functions in SVM, so how to select a good kernel function is also a research issue. However, for general purposes, there are some popular kernel functions [11-12]:

1. Linear kernel:

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \mathrm{x}^T \mathrm{x}_i \tag{6}$$

2. Polynomial kernel:

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = (\gamma\, \mathrm{x}^T \mathrm{x}_i + \mathrm{r})^d, \gamma > 0 \tag{7}$$

3. RBF kernel:

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \exp(-\gamma \left\| \mathrm{x} - \mathrm{x}_i \right\|^2), \gamma > 0 \tag{8}$$

4. Sigmoid kernel:

$$\mathrm{K}(\mathrm{x}, \mathrm{x}_i) = \tanh(\gamma\, \mathrm{x}^T \mathrm{x}_i + \mathrm{r}) \tag{9}$$

Here, γ, r and d are kernel parameters. In these popular kernel functions, RBF is the main kernel function because of following reasons [13-14]:

1. The RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.
2. The RBF kernel has less hyper parameters than the polynomial kernel.
3. The RBF kernel has less numerical difficulties.

IV. DATASET AND METHODOLOGY

The data used in this study is CSM (conventional and social media movies) dataset 2014 and 2015, which published in UCI Dataset. There are 231 instances. The data is classify by popularity (popular or not) based on 12 features categorizes as conventional and social media features. Both conventional features collected from movie databases on Web as well as social media features (YouTube, Twitter). All of features and their description is shown in Table 1.

Table 1 Data Description

Feature	Description
Ratings	Used to rate a film's suitability for certain audiences based on its content
Gross	Gross box office earnings of a movie in U.S. dollars
Budget	Refers to the process by which a line producer, unit production manager, or production accountant prepares a budget for a film production
Screens	Installation consisting of a surface and a support structure used for displaying a projected image for the view of an audience
Sentiments	Positive or negative audience sentiments via Twitter
Views	Number of audiences that view the movie
Likes	Number of likes to movie in social media
Dislikes	Number of dislikes to movie in social media
Comments	Number of audiences commenting movie in social media

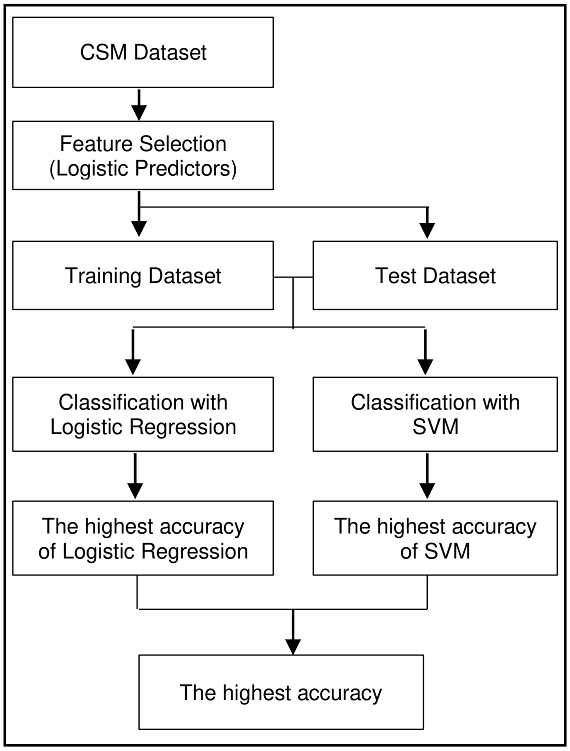
CSM Dataset
Feature Selection (Logistic Predictors)
Training Dataset
Test Dataset
Classification with Logistic Regression
Classification with SVM
The highest accuracy of Logistic Regression
The highest accuracy of SVM
The highest accuracy

Figure 1 Flowchart of the proposed Logistic Regression- SVM framework for movie popularity prediction

A. Data Conventional And Social Media Movies (CSM)

Almost all features have values that are above the upper limit or below the lower limit, in other words there are many data outliers. While the screens is a feature that does not has an outlier. Ratings feature has a nearly normal distribution. Gross, Budget, Views, Likes, Dislikes, Comments, and Aggregate Followers tends to be like an exponential distribution. For missing values, Budget, Screens, and Aggregate Followers are features that have missing values. So, imputation is used in this case with grand mean and mean of each class of popularity. In this study, dataset have been configured using 5 fold cross validation (CV) and holdout method to train and test our designed models respectively, 80% as training data and 20% testing data.

B. Methodology

The first step is preprocessing data. Then proceed with feature selection using forward method. The selected feature will be used to classify the CSM dataset. Classification is done by using logistic regression and SVM. Finally, comparison between the results of the best accuracy of each method to obtain the best classification results for CSM dataset. Following flowchart present the steps of the study.

V. RESULTS AND DISCUSSION

A. Results

In this study used 3 types of Kernel i.e. radial basis function (RBF), sigmoid, and polynomial. Model selection is also an important issue in SVM. Recently, SVM have shown good performance in data classification. Its success depends on the tuning of several parameters which affect the generalization error. We often call this parameter tuning procedure as the model selection. If we use the linear SVM, we only need to tune the cost parameter C. Unfortunately, linear SVM are often applied to linearly separable problems. Many problems are non-linearly separable. For example, Satellite data and Shuttle data are not linearly separable. Therefore, we often apply nonlinear kernel to solve classification problems, so we need to select the cost parameter (C) and kernel parameters (γ, d). Kernel parameters of C and γ exponentially growing sequences is a practical method to identify good parameters (for example, $C = 2^{-5}, 2^{-3}, ..., 2^{15}$, $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$) [14]. In this study, SVM classification used 3 Kernel function with combination between $C = 2^{-5}, 2^{5}, 2^{15}$, $\gamma = 2^{-15}, 2^{-5}, 2^{3}$, and degree = 1,2,3 (for polynomial).

Table 2 Accuracy for Testing Data Used RBF Kernel

gamma (γ)	cost (C)		
	2^{-5}	2^{5}	2^{15}
2^{-15}	70.56%	97.40%	100.00%
2^{-5}	70.56%	100.00%	100.00%
2^{3}	70.56%	100.00%	100.00%

The calculation of the accuracy used RBF Kernel for testing data is shown in Table 2. It can be seen that the maximum value of accuracy is 100%, which is used combination between $C = 2^{5}, 2^{15}$ and $\gamma = 2^{-15}, 2^{-5}, 2^{3}$. While combination between $C = 2^{-5}$ and $\gamma = 2^{-15}, 2^{-5}, 2^{3}$ have the lowest accuracy.

Table 3 Accuracy for Testing Data Used Sigmoid Kernel

gamma (γ)	cost (C)		
	2^{-5}	2^{5}	2^{15}
2^{-15}	68.83%	50.65%	50.65%
2^{-5}	70.56%	66.23%	65.37%
2^{3}	70.56%	70.56%	70.56%

From Table 3, the maximum value of accuracy used Sigmoid Kernel is 66.23% with $C = 2^{5}$ and $\gamma = 2^{-5}$. If it compared to the maximum value of accuracy used RBF Kernel, both achieved the best accuracy used $C = 2^{5}$ and $\gamma = 2^{-5}$. For sigmoid, if γ is large, then the accuracy will increase. As γ decreases, the accuracy will decrease. But if cost (C) is large, the accuracy will decreas, conversely if cost (C) is small. Following Table 4 is shown accuracy for polynomial Kernel.

Table 4 Accuracy for Testing Data Used Polynomial Kernel

degree	gamma (γ)	cost (C)		
		2^{-5}	2^{5}	2^{15}
1	2^{-15}	70.13%	71.86%	32.90%
	2^{-5}	28.57%	32.90%	32.90%
	2^{3}	71.86%	30.74%	69.70%
2	2^{-15}	30.74%	29.87%	29.87%
	2^{-5}	29.87%	29.87%	29.87%
	2^{3}	69.70%	30.74%	29.44%
3	2^{-15}	69.70%	69.70%	69.70%
	2^{-5}	69.70%	69.70%	69.70%
	2^{3}	30.30%	27.71%	70.13%

Overall, accuracy from polynomial kernel is lower than RBF or sigmoid. The best accuracy from polynomial kernel around 70 percent.

Comparison of Classification Result

In order to make comparative study SVM, we compare training-testing strategy used 5-fold cross validation (CV) and repeated holdout 80%. Based on Table 2, it was found that using combination between $C = 2^{5}$ and $\gamma = 2^{-5}$ can be correctly classified as 100%. Thus, this combination is chosen as the best combination kernel parameter for the SVM. This best combination is using to compare training-testing strategy. For each strategy used stratified and non-stratified with experiment results are list in following table.

Table 5 SVM Comparative Study

No	Holdout 80%		5 fold CV	
	Stratified	Non-stratified	Stratified	Non-stratified
1	75.68%	83.78%	70.65%	69.56%
2	67.57%	72.97%	70.27%	71.89%
3	67.57%	75.68%	70.27%	69.19%
4	67.57%	83.78%	70.81%	72.43%
5	67.57%	75.68%	70.81%	69.73%
Mean	69.19	78.38	70.56	70.56
Variance	13.15	25.57	0.08	2.21

From Table 5, the highest accuracy is 83.73% with holdout non-stratified. This result will be compare with regression logistic. The comparison results are list in following data.

Table 6 Comparison of Classification Result

Method	Accuracy of Testing Data
Logistic Regression	77.29%
SVM (RBF Kernel)	83.73%

Based on comparison, it was found that SVM (escpecially RBF Kernel) obtains greater accuracy than logistic regression. Thus, we can simply conclude that SVM is better than logistic regression in analysis for this data. If we look at the advantages of both methods, we can find that if logistic regression is used, in addition to getting a good classification model, we can also know the significant variables that influence the response variable. Thus, although logistic regression is a classical method, this also gives us more information that other methods that can only produce accuracy values.

B. Discussion

This research has presented a process of a design prediction model popularity of movies as feature selection methodology has applied. After applying principal component analysis we can observe on the four components extracted which has eigen value more than 1. If this results compared with CSM original data, its accuracy smaller than original data's accuracy. So, feature extraction not suitable for this data. Then, when we do the normalize, standardize and reduction the data for imbalance, does not change significantly. Actually, the best feature is genre, sentiment, likes, and comments with imputation grand mean feature selection using forward without normalize, standardize and reduction the data.

VI. CONCLUSION

This research aimed to apply and evaluate different statistical and intelligent models to predict popularity of movie in Conventional and Social Media Movie (2014 and 2015) Dataset. Based on the experimental results, we concluded the following: first, for this dataset, the best imputation for missing value is use grand mean not mean of each class. In this dataset, only metric feature which has missing value, so the grand mean used for imputation. If dataset has missing value in categorical feature, modus can used to imputation. Second, The highest accuracy for classification in this case is 77.2944% (data with imputation grand mean and feature selection (using forward) with 4 features (genre, sentiment, likes, and comments). So in this dataset, genre, sentiment, likes, and comments are the best feature to predict popularity of the movies. The last, the best accuracy for classification using SVM in CSM dataset is 100% (using radial basis function or Gaussian). For training-testing strategy, holdout-stratified and 5 folds cross validation-stratified have smaller varians than (holdout/5 folds cross validation) non-stratified. In this study, Gaussian is the best kernel because this kernel given the perfect accuracy, moreover given the smaller variance for stratified training-testing strategy.

REFERENCES

[1] M. Babita and C. K. Jangid, "Survey on Movies Popularity Prediction System Using Social Media Feature," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 4, no. 9, 2016.

[2] D. A. Salazar, J. Veliz and J. C. Salazar, "Comparison between SVM and Logistic Regression: Which one is Better to Discriminate," *Revista Colombiana de Estadistica Numero especial on Biostadistica,* vol. 35, no. 2, pp. 223-237, 2012.

[3] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning,* vol. 20, no. 3, p. 273–297, 1995.

[4] C. Jie, L. Jiawei, W. Shulin and Y. Sheng, "Feature selection in machine learning: a new perspective," *Neurocomputing* , 2018.

[5] https://www.medcalc.org/manual/logistic_regression.php.

[6] S. Purnami, S. Rahayu and A. Embong, "Feature Selection and Classification of Breast Cancer Diagnosis Based on Support Vector Machines," *IEEE,* pp. 1-6, 2008.

[7] Tokan, N. Turker and G. Filiz, "Analysis and Synthesis of the Microstrip Lines Based on Support Vector Regression.," in *Microwave Conference, 2008. EuMC 2008. 38th European. IEEE*, 2008.

[8] Y. Chang and C. Hsieh, "Training and Testing Low-degree Polynomial Data Mappings via Linear SVM," *Journal of Machine Learning Research,* vol. 11, p. 1471–1490, 2010.

[9] B. Ribeiro, C. Silva, N. Chen, A. Vieira and N. Carvalho, "Enhanced default risk models with SVM+," *Expert Systems with Applications,* vol. 39, no. 11, p. 10140–10152, 2012.

[10] S. Haykin, Neural networks and learning machines, Upper Saddle River: Pearson Education, 2009.

[11] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions,* pp. 415-425, 2002.

[12] H. Byun and S.-W. Lee., "A survey on pattern recognition applications of support vector machines," *International Journal of Pattern Recognition and Artificial Intelligence,* pp. 459-486, 2003.

[13] Vanitha A.R. and L. Venmathi, "Classification of Medical Images Using Support Vector Machine," in *Proceedings of International Conference on Information and Network Technology (ICINT 2011).*, 2011.

[14] C.-W. Hsu and L. Chih-Jen, A Practical Guide to Support Vector Classification, National Taiwan University, 2016.