Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Risk Factors of Cervical Cancer using Classification in Data Mining

To cite this article: Nazim Razali *et al* 2020 *J. Phys.: Conf. Ser.* **1529** 022102

View the article online for updates and enhancements.

Recent citations

- Amir Zulhilmi *et al*

Risk Factors of Cervical Cancer using Classification in Data Mining

Nazim Razali[1], Salama A Mostafa[1], Aida Mustapha[1], Mohd Helmy Abd Wahab[2], Nurul Atieqah Ibrahim[1]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400 Batu Pahat, Johor, Malaysia
[2] Faculty Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400 Batu Pahat, Johor, Malaysia

E-mail: salama@uthm.edu.my

Abstract. According to World Health Organization, cervical cancer is the fourth most frequent cancer that have high mortality rate which affected women all around the world especially in low and middle-income countries. As the computer science and information technology field growth, researches on analysing medical datasets such as diabetes, cervical cancer and liver disease and etc also growth. This paper is set to studies classification techniques in data mining on risk factor of cervical cancer datasets. The clssification techniques such as Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) have been used to classify the dataset whether healthy or cancer result for cervical cancer diagnostic. The dataset is needed to be undergoing intense data pre-processing phase due to imbalance and have a lot of missing value. The performance of classification were evaluated using 10-folds cross validation where accuracy, precision and recall as evaluation metric were measured using confusion matrix to determine the performance power for all classification techniques.

1. Introduction

A cross-sectional study was conducted on the knowledge about the cervical cancer. Cervical cancer is amongst cancer often occurs against women who attack the reproductive organs. It occurs when normal cells in the cervix turn into cancerous cells. The main cause for cervical cancer is the human papilloma virus (HPV) transmitted through sexual intercourse. HPV is a group of viruses that commonly infect the reproductive tract of sexually active men and women [1]. The body's immune system that is exposed to HPV can always counteract this virus attack from infecting women, but for a small part of women, the virus has a life of many years until it converts the cells on the surface of the cervix to cancer cells.[2] suggest that every woman in the target age group of 30 to 49 years old must perform screening test at least once in their lifetime. Screening and early detection of cancer contribute to decrease the rate of death among the cancer patients. The cervical cancer is the second most successful killer after breast cancer with approximately 311000 women died from cervical cancer in 2018 [1].

World Health Organization (WHO) has recommends a comprehensive approach on cervical cancer control. This comprehensive approach include three level of prevention based on women age group that consist of HPV vaccination, screening, treatment of pre-cancer lesions and management of invasive cervical cancer and preventive interventions for boys and girls such as health information, sex education, male circumcision and condom promotion and provision for those engaged in sexual activity [1]. Despite the possibility of prevention, cervical cancer can still occur and need a full attention for detection besides using screening test. Clssification in data mining are widely for medical diagnosis where the suitable algorithms can be applied to the dataset which may assist the medical officers to diagnoses and provide the best medications for their patients.

This paper will provide an insight on the best classifier such as Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) for classifying the risk factor for cervical cancer. Section 2 reviews all works related to the cervical cancer in data mining. Section 3 presents the classification methodology used to perform the data mining task along with the dataset and the evaluation metrics. Section 4 presents the results and finally Section 5 concludes with some direction for future work.

2. Related Work

Nowadays, application of data mining which part of Artificial Intelligence (AI) in medical datasets has been widely used. A chronological review on algorithms for screening of cervical cancer conducted by [3] on 15 previous work stated that AI has a lot of potential to assist the expert for cervical cancer early detection and saving unnecessary cost of cervical cancer screening. They discussed about the data sources as well as multiple algorithms that have proposed in previous work and compare it using standard metric for performance measurement. The finding of their review showed that several algorithm such as convolutional neural network (CNN), support vector machine (SVM) and random forest with k-nearest neighbors competitively yields high accuracy for classifying cervical cell image (PAP smear image).

[4] have applied classification algorithms in data mining such as Bagging, k-nearest neighbors (IBK), decision tree J48, JRip, Multilayer perceptron (MP) neural network and Naive Bayes (NB) classifiers to diagnose the selected medical datasets. The dataset consists of Breast Cancer Data, Chronic Kidney Disease, Cryotherapy, Hepatitis, Immunotherapy, Indian Liver Patient Dataset (ILPD), Liver Disorders, and Liver disorders dataset. ILPD and Liver disorders, Pima diabetes, risk factors cervical cancer and Statlog (Heart) dataset were extracted from UCI Machine Learning Repository [5]. The result in term of accuracy showed that no prominent classifier that consistently perform in all medical dataset however Bagging and JRip algorithm have similar accuracy for risks factors cervical cancer dataset compare to other algorithms. Note that, the risks factors cervical cancer dataset used in this work is similar to the dataset used in this paper. Thus, this dataset have been setup as benchmark dataset for cervical cancer diagnostic.

In 2019, [6] has proposed cervical cancer identification with Synthetic Minority Oversampling Technique (SMOTE) and PCA analysis using random forest classifier. The main focuses of this work is to showed that imbalance dataset may affect the overall classification performance. The dataset that been used in their work also is similar with the dataset in this paper. The SMOTE method are proven to increase the quality of metrics including accuracy, sensitivity, specificity, positive predicted accuracy (PPA) and negative predicted accuracy (NPA) compare to dataset without SMOTE which are imbalance.

3. Methodology

This section focus on classification methodology in the paper. The experiment proposed in this paper followed the methodology from work of [7] called knowledge discovery in data mining

(KDD). There are two important stages tha need extra attention and concentration im this methodology which are 1) Data gathering and preparation stages (pre-processing) and 2) Model building and evaluation stages.

Data gathering and preparation stages (pre-processing). This stages is where the existing dataset for cervical cancer contain a lot of missing values and imbalance will be treated. This missing values were caused by unanswer question from several patients due to privacy concerns. As a result, around 12% to 14% of every attributes containing missing values. As solution, the sample mean and mode as data imputation method have been used to treat the missing values in the affected attributes. However, there are 5 attributes such as STDs: Times since first diagnosis, STDs: Times since last diagnosis, Hinselmann, Schiller and Citology were omitted since this 2 from 5 attributes have 92% of missing value while others are target classes since this study focus on Biopsy as main target class. Besides, synthetic minority oversampling technique (SMOTE) [8] have been applied to the dataset after missing data treatment in order to combat the imbalance dataset and generate new instances for undersample target class. Then, the the newly generated data need to be randomised to shuffle the undersample target class.

Model building and evaluation stages. This stages is where the classification algorithms such as Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) are applied to the dataset and evaluation the performance for classification algorithms in term of accuracy, precision and recall are executed using confusion matrix based on 10-folds cross validation. The classification methodology that shown in Figure 1 for implementation.

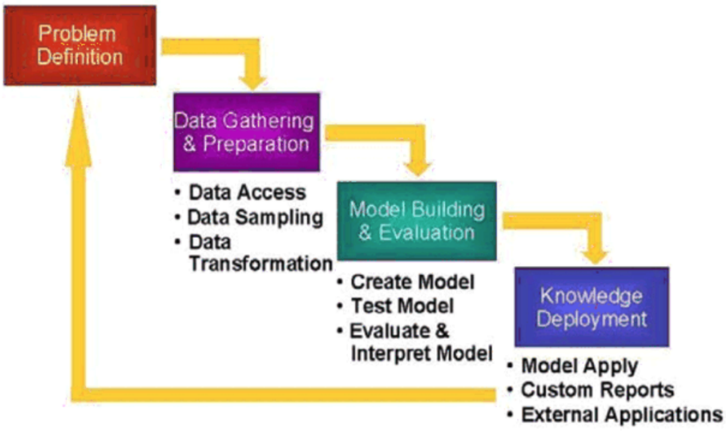
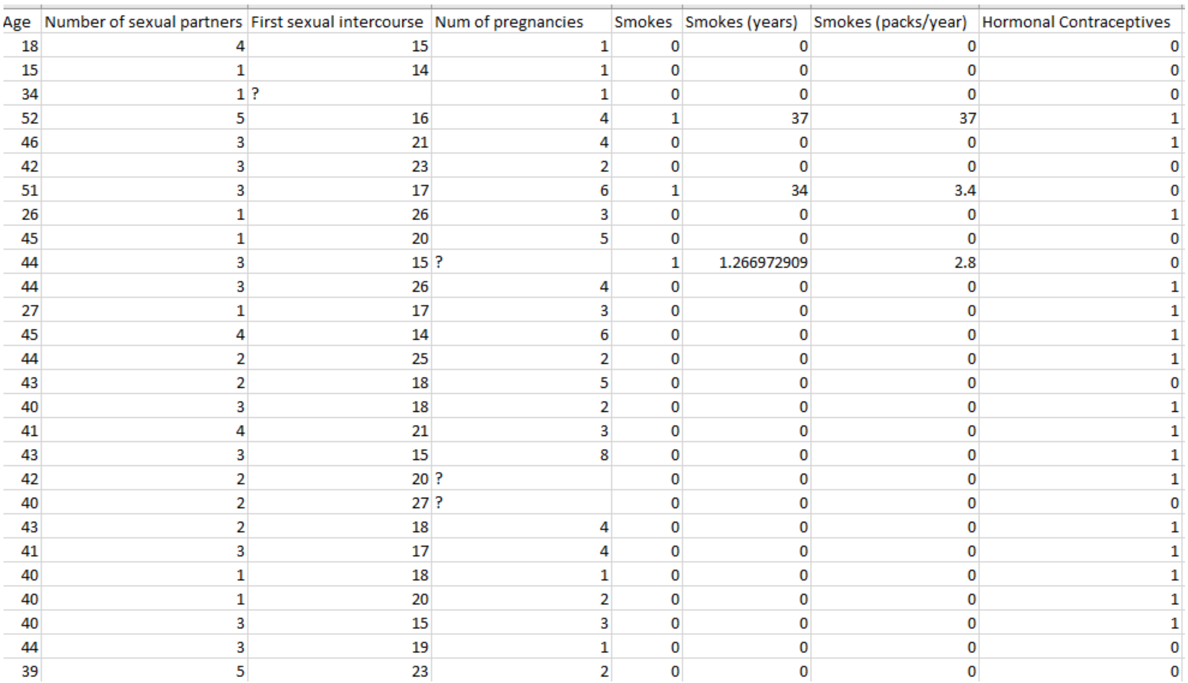


Figure 1. Classification Methodology [7]

3.1. Dataset

The dataset used in this research is the cervical cancer risk factor extracted from UCI Machine Learning Repository [5]. The dataset was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela which consists of demographic information, habits, and historic medical records of 858 patients with 32 attributes and 4 target classes (Hinselmann, Schiller, Cytology and Biopsy)[9]. This paper follow the previous works such as [10] that used Biopsy as target class. However, there are some missing value in the dataset means that there are empty answer for some question due to privacy concerns from several patients. Thus, this research only focus on available attributes that have missing values below 15% for diagnosis of cervical cancer. The excerpt of the dataset is show in Figure 2.



Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives
18	4	15	1	0	0	0	0
15	1	14	1	0	0	0	0
34	1	?	1	0	0	0	0
52	5	16	4	1	37	37	1
46	3	21	4	0	0	0	1
42	3	23	2	0	0	0	0
51	3	17	6	1	34	3.4	0
26	1	26	3	0	0	0	1
45	1	20	5	0	0	0	0
44	3	15	?	1	1.266972909	2.8	0
44	3	26	4	0	0	0	1
27	1	17	3	0	0	0	1
45	4	14	6	0	0	0	1
44	2	25	2	0	0	0	1
43	2	18	5	0	0	0	0
40	3	18	2	0	0	0	1
41	4	21	3	0	0	0	1
43	3	15	8	0	0	0	1
42	2	20	?	0	0	0	1
40	2	27	?	0	0	0	0
43	2	18	4	0	0	0	1
41	3	17	4	0	0	0	1
40	1	18	1	0	0	0	1
40	1	20	2	0	0	0	1
40	3	15	3	0	0	0	1
44	3	19	1	0	0	0	0
39	5	23	2	0	0	0	0

Figure 2. Excerpt of Cervical Cancer Dataset

Table 1 shows list of attributes and type of attributes in dataset cervical cancer risk factor. Several patients decided not to answer some of the questions due to privacy concerns. Hence, the attributes denoted by *boolean* (0 or 1) and *integer* were encoded as data type. Missing values for attribute that have integer data type were filled using the sample mean while boolean were filled using the sample mode. Table 1 shows the features and data type in the Cervical Cancer dataset.

3.2. Algorithms

The previous work on literature review has showed several prominent algorithms for cervical cancer classification. As the result, this research intends to use seven classification algorithms in data mining that are prominent in previous work in order to provide a comparative analysis using the provided benchmark dataset.

- **Naive Bayes** is a most popular and simplest probabilistic algorithm in classification. It has ability to handle missing value and imbalance data. It defines all the attributes are independence or no depency between all the attributes except the attributes that become the target class. Thus, it neglect the effect of correlation of other attributes and solely dependence to the target class.
- **C4.5 Decision Tree** and also known as J48 Decision Tree in WEKA data mining tool is a extension of ID3 decision tree that can incorporate missing values, pruning, continuous data and derivation of rules [11]. J48 also is flexible in tree pruning adjustment in order to avoid over fitting pruning.
- **k-Nearest Neighbors** or alson know as Instance Based Learner (IBk) in WEKA data mining tools uses a distance metric to locate the value of "closseness" for k instances in the training data for each test instance. Prediction will be made based on the selected instances.

Table 1. Attributes and Data Type for Risk Factor Cervical Cancer Dataset

Attribute	Type	Missing Value Yes/No (%)	Attribute	Type	Missing Value Yes/No (%)
Age	Integer	No	STDs: pelvic inflammatory disease	Boolean	Yes (12%)
Number of sexual partners	Integer	Yes (3%)	STDs: genital herpes	Boolean	Yes (12%)
First sexual intercourse (age)	Integer	Yes (1%)	STDs: molluscum contagiosum	Boolean	Yes (12%)
Number of pregnancies	Integer	Yes (7%)	STDs:AIDS	Boolean	Yes (12%)
Smokes	Boolean	Yes (2%)	STDs:HIV	Boolean	Yes (12%)
Smokes (years)	Integer	Yes (2%)	STDs:Hepatitis B	Boolean	Yes (12%)
Smokes (packs/year)	Integer	Yes (2%)	STDs:HPV	Boolean	Yes (12%)
Hormonal Contraceptives	Boolean	Yes (13%)	STDs: Number of diagnosis	Integer	No
Hormonal Contraceptives (years)	Integer	Yes (13%)	STDs: Time since first diagnosis	Integer	Yes (92%)
IUD	Boolean	Yes (14%)	STDs: Time since last diagnosis	Integer	Yes (92%)
IUD (years)	Integer	Yes (14%)	Dx:Cancer	Boolean	No
STDs	Boolean	Yes (12%)	Dx:CIN	Boolean	No
STDs (number)	Integer	Yes (12%)	Dx:HPV	Boolean	No
STDs: condylomatosis	Boolean	Yes (12%)	Dx	Boolean	No
STDs: cervical condylomatosis	Boolean	Yes (12%)	Hinselmann	Boolean	No
STDs: vaginal condylomatosis	Boolean	Yes (12%)	Schiller	Boolean	No
STDs: vulvo-perineal condylomatosis	Boolean	Yes (12%)	Citology	Boolean	No
STDs:syphilis	Boolean	Yes (12%)	Biopsy	Boolean	No

- **Sequential Minimal Optimization (SMO)** proposed by John Platt in 1998 to provide solution for training a Support Vector Machine (SVM) classifier [12]. Training a SVM requires a lot computational power and time consuming to handle quadratic programming (QP) optimization problem. Thus, SMO enhanced SVM by breaks QP as smallest as possible and solved it analytically.
- **Random Forest** is a top ensemble machine learning algorithms in WEKA data mining tools [13]. Actually, random forest is an extension from bagging decision tree that force split points to be selected only from a random subset of input attributes when building the tree. This means Random Forest can be used for both classification and regression tasks.
- **Multilayer Perceptron (MLP)** or also known as Feed Forward Neural Network is the most typical neural network model for generating a set of outputs from a set of inputs. It also known as deep learning method. An MLP consists of several layers of input nodes

inter-connected as a directed graph to the output layers. Training MLP is carried out by using the backpropogation algorithm.

- **Simple Logistic Regression** is one of the most simple and commonly used in data mining for two-class classification or when the dependence attribute is binary which means there are only two possible class such as Yes/No, Positive/Negative or Survived/Death. This algorithm is often used as the baseline comparison in binary classification problem because it only estimates the probability of relations between the dependent and independent variables.

3.3. Evaluation Metrics

The evaluation metrics used in the experiments are accuracy, precision and, recall which represented as a confusion matrix. This evaluation metrics are used to calculate the performance of classification techniques, Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest Decision Tree (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) on risk factor of cervical cancer data between actual and predicted results. This performance result will be observed based on the measurement on how well the prediction were differ from actual output and the evaluation metric were derived from a confusion matrix as shown in Table 2.

Table 2. Confusion Matrix

Confusion Matrix		Class 1	Class 2
		Predicted	Predicted
Class 1	Actual	TP	FN
Class 2	Actual	FP	TN

where class 1 and 2 can be positive and negative, the equations can be defined as:

- Positive (P): Observation is positive (for example: positive cancer diagnosed or cancer).
- Negative (N): Observation is not positive or negative (for example: negative cancer diagnosed or healthy).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

Accuracy can be defined as in Equation 1:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision can be defined as in equation 2 where the total number of correctly classified positive samples are divided by the total number of actual positive samples.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall can be defined as in equation 3 where the ratio of the total number of correctly classified positive samples divided by the predicted total number of positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

4. Experimental Results

The purpose of the experiments is to select the best classification algorithm for cervical cancer diagnostic. The performance result of Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR) based on the risk factor of cervical cancer dataset using classification in data mining are evaluated and compare in term of accuracy, precision and recall using 10-folds cross validation. The results are shown in Table 3.

Table 3. Experimental Results

Algorithm	Accuracy	Precision	Recall
NB	0.686	0.702	0.686
C4.5	0.929	0.929	0.929
kNN	0.926	0.926	0.926
SMO	0.747	0.747	0.746
RF	0.964	0.965	0.964
MLP	0.859	0.860	0.859
SLR	0.769	0.771	0.769

The results showed that RF (96.40%) is slightly better compare to kNN (92.60%) and C4.5 (92.90%) and outperform the remaining classification algorithm such as MLP (85.90%), SMO (74.70%) and SLR (76.90%) in accuracy. Meanwhile, NB gained the lowest value in accuracy by 68.60%. Figure 3 shows the overall result of accuracy in percentage for all proposed classification algorithms.

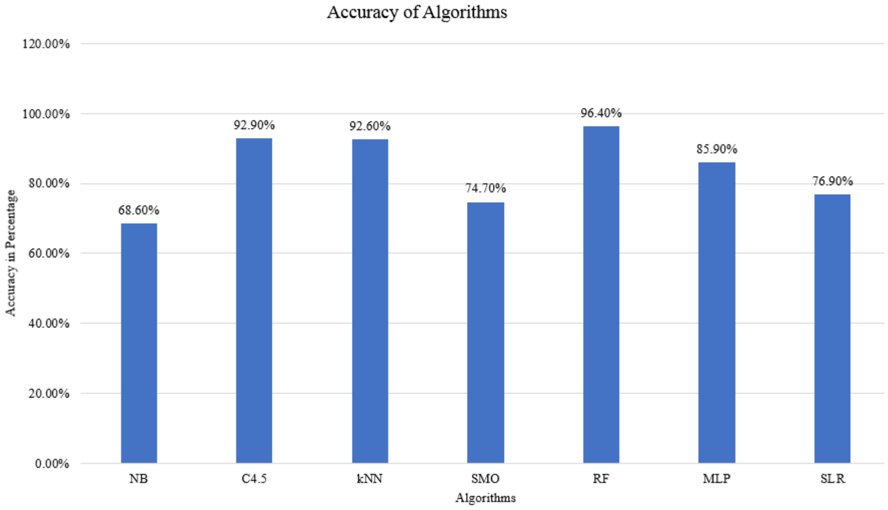


Figure 3. Result on Accuracy for All Algorithms

5. Conclusions

This research presented an analysis of risk factor for cervical cancer data using classification algorithms based on Naive Bayes (NB), C4.5 Decision Tree (C4.5), k-Nearest Neighbors (kNN), Sequential Minimal Optimization (SMO), Random Forest (RF), Multilayer Perceptron (MLP) Neural Network and Simple Logistic Regression (SLR). The data undergoing data imputation to handle missing values and oversampling method called SMOTE to combat imbalance dataset before be randomised to shuffle the newly generated sythentic dataset. The evaluation metric such as accuracy, precision and recall are measured for the given dataset to estimate the performance of each classification techniques. As the result, RF have achieved the highest rate of accuracy, precision and recall compare to other six classification algorithms while Naive Bayes is the lowest in accuracy. It is recommended to use other data imputation method besides using sample mean and mode since this method is the basic method in data imputation. Besides, other undersampling and oversampling method need to be considered for better performance.

Acknowledgements

This research is funded by Universiti Malaysia Perlis.

References

[1] Organization W H 2019 *Retrieved from https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer*

[2] Organization W H 2019 *Retrieved from https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/*

[3] Singh Y, Srivastava D, Chandranand P and Singh S 2018 *Retrieved from https://arxiv.org/abs/1811.00849*

[4] Ramana B and Kumar Boddu R 2017 *9th IEEE Annual Computing and Communication Workshop and Conference, CCWC 2019* 140–145

[5] Repository U M L 2019 *Retrieved from https://archive.ics.uci.edu/ml/index.php*

[6] Geetha R, Sivasubramanian S, Kaliappan M, Vimal S and Annamalai S 2017 *Journal of Medical Systems* **43**

[7] Fayyad U, Piatetsky-Shapiro G and Smyth P 1996 *Advances in knowledge discovery and data mining* 1–34

[8] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 *Journal of Artificial Intelligence Research* **16** 321–357

[9] Fernandes K, Cardoso J S and Fernandes J 2017 *Iberian Conference on Pattern Recognition and Image Analysis 2017* 243–250

[10] Al-Wesabi Y, Choundhury A and Won D 2018 *Proceedings of the 2018 Institute of Industrial and Systems Engineers (IISE) annual conference*

[11] Gaganjot K and Chhabra A 2014 *International Journal of Computer Applications* **98** 13–17

[12] Platt J 1998 *Technical Report MSR-TR-98-14*

[13] Breiman L 2001 *Machine Learning* **45** 5–32