Predicting The Suspect of New Pulmonary Tuberculosis Case using SVM, C5.0 and Modified Moran's I

Rusdah[1,2], Edi Winarko[1], Retantyo Wardoyo[1],
[1]Department of Computer Science and Electronic, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia

[2]Department of Information Systems, Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia

Summary
Indonesia, one of the 22 high-burden countries, has the second largest numbers of tuberculosis (TB) cases in the world. According to WHO's 2015 report, Indonesia was estimated to have one million new TB cases per year. Unfortunately, only one-third of new TB cases are detected. The number shows a serious delay in TB diagnosis and treatment. Delayed treatment of TB is associated with long-term lung damage, which can multiply and spread the bacilli as well. Diagnosis of TB is difficult, especially in the case of pediatric patients, extrapulmonary TB, and smear-negative pulmonary TB, due to various reasons. In addition, some of the tuberculosis symptoms have in common not only with lung cancer but also with other diseases. This study aims to build classification model to predict the suspect of new pulmonary tuberculosis case. The data were taken from the medical record of tuberculosis patients in Jakarta Respiratory Center. A modified Moran's I was proposed in data transformation proses. The training data were classified using Support Vector Machine (SVM). The misclassified data were further used to generate rules using C5.0. The result showed that the proposed method in transforming data used in the proposed model could perform better than comparison model. The proposed model has an accuracy of 84.54%, specificity of 85.24%, and sensitivity of 85.24%.

Keywords:
Modified Moran's I, medical data mining, tuberculosis data, preliminary diagnosis, TB Screening.

1. Introduction

Tuberculosis is a bacterial infection that causes more death in the world than any other infectious disease [1]. Globally, WHO reported that there were 9.6 million new cases of TB in 2014. India, Indonesia, and China had the largest number of cases: 23%, 10% and 10% of the global total, respectively [2]. Indonesia is one of the 22 high-burden countries. High-burden means that there are around 100 cases per 100.000 population or more. WHO estimated one million new TB cases per year in Indonesia. Unfortunately, only one-third of the new TB cases are detected. The number shows a serious delay in TB diagnosis and treatment.

Delayed treatment of TB is associated with long-term lung damage, such as extensive caseous necrosis and the formation of cavity lesions, which can facilitate the multiplication spread of bacilli and can result in serious respiratory dysfunction as well [3]. They reported that a treatment delay of over 12.1 weeks would result in a larger proportion of patients having severe TB, a higher mortality rate, and greater treatment failure.

TB diagnosis is difficult due to some reasons [4], [5]. The first reason is in case of pediatric patients who have a little number of germs [6], [7]. The second is in case of extrapulmonary TB [7]–[10], and the third is in case of smear-negative pulmonary tuberculosis (SNPT) [11]–[14].

The suggestive symptoms of TB are cough, hemoptysis, night sweats, fever, and weight loss [15]. Those symptoms have in common not only with lung cancer [16] but also with other diseases [17], [18]. It leads to delay in the correct diagnosis and exposure to inappropriate medication [16], misdiagnosis and death as well [19]. A longer delay in diagnosis of pulmonary tuberculosis prevents rapid treatment, and the individual remains without being isolated. Additionally, individuals that receive inadequate treatment are more vulnerable to develop multidrug-resistant tuberculosis [1].

Some studies have been done to overcome these problems. The studies related to the diagnosis of TB have been conducted using sound, images, blood miRNA profiles, and variables as input parameters. Some studies using sound as an input are conducted in [20] and [21]. They used coughing sound detection algorithm utilizes lung sound waves to accelerate the process of TB diagnosis with a high degree of accuracy and specificity. Many studies used the image of Mycobacterium tuberculosis in tissue as an input to assist pathologists [22]–[26]. A Recent study was conducted using blood miRNA profiles, and the model was also tested using urine and saliva miRNA [27]. In addition, data mining methods to diagnose tuberculosis using clinical symptoms as input have been widely used in many studies [28], [29].

According to Pedoman Nasional Pengendalian Tuberkulosis (National Guidance for Tuberculosis Control) [30], TB diagnostic process consists of two phases. The first phase is discovering tuberculosis suspect using the data of

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

symptoms and physical examinations. The second one is sputum examination (direct microscopic) to generate a final diagnosis. If the result of the microscopic examination is positive, then the patient is diagnosed with smear-positive pulmonary TB. If it is negative, then additional investigations must be conducted, at least with chest X-ray examination to determine if a patient is diagnosed with smear-negative pulmonary TB [30].

This study aims to build a classification model for predicting the suspect of new pulmonary tuberculosis case. The model is the first phase of the diagnosis of tuberculosis process. Hence, the inputs used in this study are patient's demographic data, anamnesis, and physical examinations [18]. The data transformation needs to be done to obtain the highest accuracy. A new statistical approach for data transformation is proposed to generate a dataset. The dataset is further classified by SVM combined with C5.0. SVM is chosen based on the result of the previous study [31], [32]. The output of this study can be an input for further research in the second phase of diagnosing pulmonary tuberculosis process.

The remaining of this paper is organized as follows. Section two deals with previous studies related to this study. The proposed method is presented in section three. The experiments and results are presented in section four. The findings of this study are discussed in and the last section contains conclusion and recommendation for the future work.

2. Related Work

Previous studies using the data of symptoms and physical examination to find the suspect of new pulmonary TB case have been done by Asha et al. [31], [33]–[35]. They used symptoms, physical examinations, radiographic findings and the result of HIV test as the datasets. Their model has two outputs, namely pulmonary tuberculosis (PTB) and retroviral PTB; which is HIV infected pulmonary tuberculosis. In the preprocessing phase, they replaced missing values with null [31]. They also discretized the numeric attributes using minimum-maximum limits and normalized all attributes using unique integer values [33].

Some classification techniques have been explored to classify PTB and retroviral PTB using the same dataset [31], [33]–[35]. In 2010, they explored ensemble classifiers such as Bagging, AdaBoost, and Random forest trees. The result shows that Bagging obtains the highest with 97% accuracy followed by 96% and 93% for Adaboost and Random forest, respectively. Furthermore, they compared the performance of basic learning classifiers and ensemble of classifiers [34]. They concluded that the best accuracy was obtained by SVM (99.14%) and C4.5 (99%) among other single classifiers and also by Random Forest (99.14%) among other ensemble classifiers. In the same year, they cascaded clustering and classification. Initially, k-means was used to group the dataset into two clusters, namely PTB and retroviral PTB. Subsequently, the dataset was classified using the same classifier as their previous work. The result showed that SVM obtained the best accuracy of 98.7% compared to others. In 2012, they conducted a comparative study using two data mining techniques, namely associative rule mining (ARM) and associative classification (AC). They summarized that AC method has resulted in smaller number rules compared to the ARM. They also compared three important algorithms of AC such as CBA, CMAR, and CPAR and finally concluded that CPAR was better in rule generation compared to others [33].

3. Proposed Method

The main objective of this study is to propose a classification model for predicting the suspect of new pulmonary tuberculosis case with improved accuracy. The proposed approach, to create the proposed model as described in Fig 1, includes the following four steps.

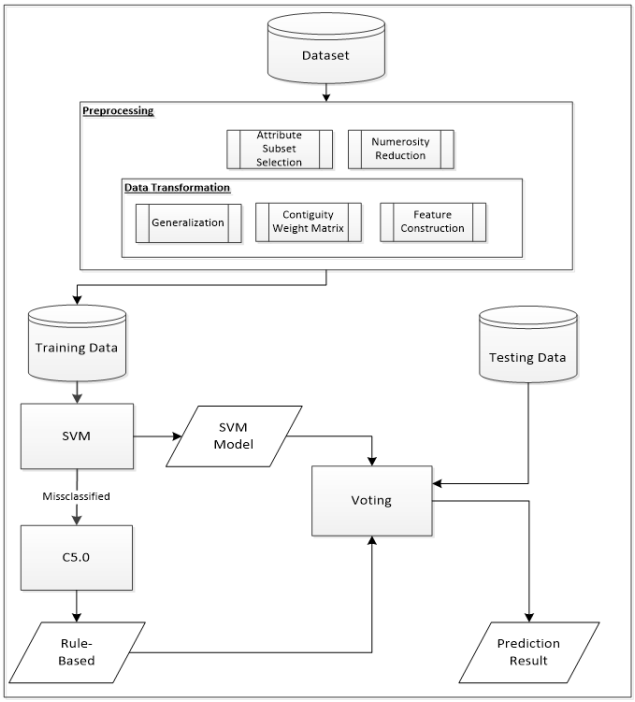

Fig. 1 Proposed model

3.1 Step 1, preprocessing

Preprocessing tasks conducted in this study used the same techniques as the previous study [32]. Those tasks included (1) data reduction by performing attribute subset selection

and numerosity reduction; and (2) data transformation by performing generalization on attribute Preliminary Diagnosis, and Duration of Cough. Data transformation was also done by transforming categorical attributes into weight value using contiguity weight matrix and by feature construction. The dataset consists of 14 attributes and 1122 records. Among them, there are twelve categorical attributes, and two numeric attributes as well, namely Age and Weight. The new statistical approach to transform those categorical attributes into weight value is described as follows.

Initially, those twelve categorical attributes were transformed into weight value (wij) using Eq. (1).

$$w_{ij} = \begin{cases} 1, & if\ the\ j^{th}\ attribute\ has\ connectivity\ with\ \ the\ i^{th}\ class\ attribute \\ 0, & if\ the\ j^{th}\ attribute\ has\ no\ connectivity\ with\ \ the\ i^{th}\ class\ attribute \end{cases} \quad (1)$$

Where w_{ij} is weight value of the j^{th} attribute for the i^{th} record (patient); j is number of attributes in i^{th} record.
Firstly, eight symptom attributes contain "Y", means that a patient has the symptom or "N", means that a patient doesn't have the symptom. If the value of a symptom is "Y", then w_{ij} equals to 1. If the value of a symptom is "N", then w_{ij} equals to 0. w_{ij} equals to 1 means that there is connectivity between the j^{th} attribute and the i^{th} class attribute (Preliminary Diagnosis). Otherwise, w_{ij} equals to 0 means that there is no connectivity between the j^{th} attribute and the i^{th} class attribute.

Secondly, other categorical attributes namely Sex, Duration of Cough and Co-Existing Illness were transformed into weight value (see Table 1). Every attribute was transformed into some new attributes. The number of new attributes depends on the value of the original attribute. For example, attribute Sex consists of "F" for a female and "M" for a male. There are two new attributes namely F and M. If the value of attribute Sex is "F", then the weight value (w_{ij}) of the new attribute F is 1, and 0 for attribute M. The same way was applied to transform the value of Duration of Cough attribute into three new attributes namely, <2W, >=2W, and N. The values of Co-Existing Illness attribute were transformed into three new attributes namely Diabetes Mellitus, Hypertension, and None.

Table 1 : Transformation of categorical attributes into weight value

	New Attributes	
Sex	*M*	*F*
M	1	0
F	0	1

	New Attributes		
Duration of Cough	*<2W*	*>=2W*	*N*
<2W	1	0	0
>=2W	0	1	0
N	0	0	1

	New Attributes		
Co-existing Illness[a]	*DM*	*HT*	*N*
DM	1	0	0
HT	0	1	0
DM&HT	1	1	0
N	0	0	1

[a] DM: Diabetes Mellitus, H: Hypertension, N: None

Feature Construction

The previous step generated sixteen numeric attributes. Next step is feature (attribute) construction using neighbor analysis. Firstly, Weight attribute construction by calculating the weight value of each patient record using Eq. (2):

$$W_i = \frac{\sum_{j=1}^{m} w_{ij}}{m} \quad (2)$$

Where W_i is total weight value for the i^{th} patient record; m is number of attributes indexed by j for the i^{th} patient record; w_{ij} is weight value of the j^{th} attribute for the i^{th} record.

Secondly, attribute construction for ZA and ZW which contain the z-score value of Age and Weight, respectively. *Z-score normalization* is used to normalize the value of Age and Weight attribute, using Eq. (3) [36].

$$Zscore_{x_i} = \frac{x_i - \bar{x}}{\sigma_x} \quad (3)$$

Where $Zscore_{x_i}$ is the normalized value of x attribute for i^{th} patient record; x_i is the value of x attribute for i^{th} patient record that will be normalized; $\bar{x}$ is mean of the overall data for x attribute; σ_x is standard deviation of x attribute.

Thirdly, attribute construction for IA and IW which contain Moran's I value of Age and Weight, respectively. Moran's I is used to perceive if an attribute is correlated with other attributes value in determining a value of the class attribute. Moran's I is notated as Eq. (4) [37].

$$I = \frac{n \sum_i^n \sum_j^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_i^n \sum_j^n w_{ij}\right) \sum_i^n (x_i - \bar{x})^2} \quad (4)$$

Where I is a value of Moran's index; n is the number of spatial units indexed by i and j; x is the variable of interest; $\bar{x}$ is the mean of x; w_{ij} is contiguity weight matrix. w_{ij} equals to 1 if the i^{th} spatial unit is contiguous with the j^{th} spatial unit and w_{ij} equals to 0 if the i^{th} spatial unit is not contiguous with the j^{th} spatial unit.

Moran's I is widely used in the fields of geography to measure of spatial autocorrelation. Spatial autocorrelation is characterized by a correlation among nearby locations in space. It is more complex than one-dimensional autocorrelation because spatial correlation is multi-dimensional (i.e., 2 or 3 dimensions of space) and multi-directional. Values of I usually range from −1 to +1. Values significantly below -1/(n-1) indicate negative spatial autocorrelation and values significantly above -1/(n-1) indicate positive spatial autocorrelation. For statistical hypothesis testing, Moran's I values can be transformed to z-scores.

This study used Moran's I to perceive if Age and Weight attribute of a patient are correlated with other attributes in predicting a result of Preliminary Diagnosis. As the focus is a correlation between attributes in a patient record, so it is unnecessary to analyze the neighbor (or in this case, other patient records). Therefore, modified Moran's I is needed, as notated in Eq. (5):

$$I = \frac{n \sum_i^n \sum_j^m w_{ij} (x_i - \bar{x})}{\left(\sum_i^n \sum_j^m w_{ij}\right) \sum_i^n (x_i - \bar{x})^2} \quad (5)$$

Where I is a value of Moran's index; n is number of records indexed by i; m is number of attributes of the i^{th} record indexed by j; x is the numeric attributes of the i^{th} record; $\bar{x}$ is the mean of x attribute; w_{ij} is binary connectivity between j^{th} attribute in the i^{th} record and the i^{th} class attribute. w_{ij} equals to 1 if the value of j^{th} attribute in i^{th} record is "Y", otherwise w_{ij} equals to 0.

Fourthly, attribute construction for ZIA and ZIW. Those attributes contain the z-score value of Moran's index for Age attribute and Weight attribute respectively. Eq. (3) is used to construct those attributes.

This first step generated a new dataset that consists of 1122 records and eight new numeric attributes. Those attributes are Weight (W), Z-score of Age (ZA), Z-score of Weight (ZW), I value of Age (IA), I value of Weight (IW), z-score for IA (ZIA), z-score for IW (ZIW), and Preliminary Diagnosis.

3.2 Step 2, Classification Process

In this step, the dataset resulted in the previous step were trained by SVM using 10 fold cross-validation. In the training process, some experiments were conducted using the four SVM kernels, namely radial, linear, sigmoid and polynomial. SVM model with the best kernel was saved, and further was validated using all 1122 records of training data.

3.3 Step 3, Decision Tree Induction

The previous step generated 355 records and eight numeric attributes of misclassified data. Some experiments were conducted to analyze those data. The details of those experiments were discussed in the third section. In this step, the misclassified data were transformed back into the original form, which consisted of 349 records, two numeric attributes, and twelve categorical attributes. The decision tree induction was further done using C5.0 to generate some rules from the original misclassified data. The rules are generated to increased the accuracy of SVM model. The output of this step was the SVM-Rule Based model.

3.4 Step 4, Evaluation

This step aims to measure the accuracy, sensitivity, specificity, and precision of the proposed model using confusion matrix. The evaluation used 427 records of testing data.

4. Experiments and Results

4.1 Data Collection

Patients' real data were collected from JRC-PPTI (Jakarta Respiratory Center–*Perkumpulan Pemberantasan Tuberkulosis Indonesia*) from 2010 to 2014. The data were taken from medical records of patients aged over 15 years and had been declared cured for the case of positive Acid-Fast Bacilli (AFB), or completed treatment for the case of negative AFB. Those medical records were in hard copy format and manually recorded into Excel file. The data consist of 1170 records and 17 attributes include Preliminary Diagnosis as a label class. Those 17 attributes were based on patient's demography data, clinical symptoms and physical findings related to the diagnosis of pulmonary tuberculosis. The preprocessing phase conducted by Rusdah et al. [32] provided the best dataset that consists of 1122 records and 14 attributes. Therefore, this study used the dataset. Every record corresponds to the most relevant information on a patient. Those attributes are sex, age, duration of cough, hemoptysis, cough with phlegm, fever, weight loss, loss of appetite, sweating at night, breathless, chest pain, co-existing illness, weight, and preliminary diagnosis. Among those categorical attributes, there are numerical attributes as well, namely age and weight.

4.2 Data Exploration

Among the 1122 patients with laboratory-confirmed TB, there are 535 cases (48%) of pulmonary tuberculosis (PTB) and 587 cases (52%) of non-PTB. Males represent 62% of PTB cases, and 66.1% of them are aged 15-44. The main symptoms of PTB cases are more than two weeks cough (89%), cough with phlegm (76.3%), breathless (64.3%), chest pain (60.2%), loss of appetite (45.8%), weight loss (39.6%), night sweats (39.1%), and fever (25.2%). There is a few complained of hemoptysis (15.9%). PTB patients are more likely to have diabetes (5.4% vs. 2.9%) than non-PTB patients.

The data further were discretized using three methods; manually, using WEKA discretization function and rough set. The data were classified using Support Vector Machine (SVM), C4.5, Naive Bayes (NB), k-Nearest Neighbor (kNN) [34], Bayesian Network (BN) [38], and Backpropagation (BP)[13]. Table 2 shows the result of using those three

discretization methods applied to numeric attributes, namely Age and Weight. The rough set got the highest accuracy, 70.23% if SVM classified the data. Unfortunately, when rough set applied to other classifiers, the accuracy was decreased.

Tabel 2 : Comparison of different classifiers' accuracy with and without discretization

Discretization Methods	Classifiers					
	C4.5	*NB*	*BN*	*kNN*	*SVM*	*BP*
Without	65.11%	67.20%	67.56%	63.81%	63.97%	65.04%
Manual	61.79%	63.68 %	66.36%	65.58%	66.84%	62.14%
Equal width	67.07%	67.75%	67.93%	66.63%	69.18%	63.88%
Equal Frequency	67.07%	67.75%	67.40%	66.54%	69.18%	63.88%
Rough Set	65.40%	66.99%	67.00%	66.40%	70.23%	63.42%

Feature selection, in the next experiment, was conducted to increase accuracy. Wrapper method was done using WEKA. It was applied to two datasets, which were datasets with and without discretization (using Rough Set). Table 3 shows the results of the experiments. It can be seen that feature selection could increase the accuracy of all classifiers if it is applied to the original dataset. On the other hand, feature selection applied to the discretized dataset decreased the accuracy of all classifiers. In addition, none of the attributes resulted in feature selection process consisted of 'cough with phlegm'. Meanwhile, according to National Guidance of Tuberculosis Control [30], the 'cough with phlegm' attribute is one of the important clinical symptoms of pulmonary tuberculosis. Based on those findings, the next experiments will be conducted without using any feature selection and discretization method.

In the first step, the original data (consist of 1122 records and 14 attributes) were preprocessed using the steps discussed in the proposed method section. This process generated new dataset consist of 1122 records and eight new numeric attributes as presented in Table 4.

The second step, the dataset resulted in the previous step were trained by SVM using ten fold cross-validation. In the training process, some experiments were conducted using the four SVM kernels, namely radial, linear, sigmoid and polynomial. Table 5 shows the result of the experiments. The linear kernel obtained the highest accuracy (68,7%) among others. This SVM – Linear model was validated using 1122 records of training data. The training process obtained 68,36% of accuracy. This proses generated 355 records and eight numeric attributes of misclassified data.

Table 3 : Comparison of different classifiers' accuracy with and without feature selection

Classifier	#Attr	Attributes[b]	Accuracy		
			w/out FS	*With FS*	*With FS+RS*
C4.5	7	DC, LA, WL, SN, W, EI, B	66,93 %	67.38%	67.11%
Naive Bayes	9	DC, WL, EI, W, A, LA, B, S, F	65.58 %	68.81%	68.81%
SVM	10	DC, WL, LA, EI, W, A, SN, B, S	70.67 %	69.70%	69.92%
Bagging with C4.5	6	WL, DC, EI, W, LA, SN	67.29 %	68.54%	64.53%
Bagging with NB	9	A, DC, WL, EI, W, LA, S, H, B	68.00 %	69.16%	67.11%
Bagging with SVM	10	A, DC, W, H, WL, LA, S, EI, SN, B	69.88 %	70.41%	69.61%

A=Age; B=Breathless; DC=Duration of a Cough; EI=co-existing illness; F=Fever; H=Hemoptysis; LA=Loss of Appetite; S=Sex; SN=Sweating at Night; W=Weight; WL=Weight Loss.

Table 4 : New attributes generated from data transformation process

W	ZA	ZW	IA	ZIA	IW	ZIW	Preliminary Diagnosis
0.50	55.49	44.19	0.25	0.42	-0.25	0.60	NON TB
0.38	41.49	47.19	-0.08	0.10	-0.75	0.10	PTB
0.56	43.49	40.19	-0.14	0.03	-0.14	0.71	PTB
0.38	71.49	45.19	0.04	0.21	-0.25	0.60	PTB
0.69	45.49	64.19	-0.23	-0.06	0.09	0.94	PTB

In the third step, C5.0 is used to generate some rules from the 355 records misclassified data. Initially, the misclassified data were transformed back into the original dataset that contains 14 attributes. Subsequently, the dataset was classified by C5.0 to generate some rules. In the training phase, the rules were further used to classify the 355 records misclassified data. This process obtained 99,38% of accuracy. It proofed that the rule-based could increase the accuracy of SVM by 31.02%.

Table 5 : Experiments of using different SVM kernels

SVM Kernels	Training Accuracy
SVM – Radial	67,4%
SVM – Linear	68,7%
SVM – Sigmoid	67,3%
SVM – Polynomial	65,45%

The fourth step is an evaluation to measure the accuracy, sensitivity, and specificity of the proposed model using confusion matrix. The evaluation used 301 records of new testing data. The proposed model has the accuracy of 51.5%, the sensitivity of 53.5%, and specificity of 47.4%.

As the accuracy of the proposed model in classifying the testing data is not satisfactory, the next experiment is needed.

The given training data (1122 records and eight attributes) and testing data (301 records and eight attributes) were merged. Those merged data were then divided into 70% new training data and 30% new testing data using stratified random sampling method. The new 996 records of training data and 427 records of testing data were used to conduct next experiments.

Initially, the 996 records were trained by SVM using 10 fold cross-validation. It resulted 64.36% accuracy in the training phase. Subsequently, all misclassified data were used to generate rule using C5.0. The rule generated then used to classify those misclassified data using 10 fold cross-validation. Accuracy in the training phase was calculated using the voting technique as illustrated in Table 6.

Table 6 : Illustration of voting to make final decision

Model Prediction			Final Decision
SVM	$C5.0_{misData}$	$C5.0_{training}$	
PTB	PTB		PTB
NON-TB	NON-TB		NON-TB
PTB	NON TB	PTB	PTB
NON TB	PTB	NON TB	NON TB

The C5.0 model could increase 32.63% of SVM accuracy. The training phase using a combination of SVM – C5.0 resulted in 97.59% accuracy. The testing phase used dataset, namely Dataset B (for the proposed method) and DatasetA (for the comparative model). DatasetB consists of 996 training records and 427 testing records and eight attributes. DatasetA consists of 996 training records and 427 testing records and 14 attributes. Table 7 shows the result of the testing phase for those models. It can be seen that the proposed model could perform better. It shows that the proposed method in transforming data could increase the accuracy of the model.

Table 7 : Performance of the proposed and comparative model

Model	Dataset	Accuracy	Specificity	Sensitivity
Proposed	DatasetB	84.54%	85.24%	85.24%
Comparative	DatasetA	83.31%	84.46%	84.46%

5. Discussion

According to the finding of this study, all discretization methods used in Table 2 could increase the accuracy of SVM, while Dag et.al [39] concluded that for SVM the discretization, using information gain and gain ratio, did not have a clear advantage. Discretization seemed to have significant effect on most of the classification algorithms such as Naïve Bayes and C4.5, both of which rely more on the categorical attributes. SVM and Artificial Neural Network algorithms increases when working with the continuous features [39].

It can be seen from Table 3 that none of the attributes resulted in feature selection process consisted of 'cough with phlegm'. Meanwhile, according to National Guidance of Tuberculosis Control [30], the 'cough with phlegm' attribute is one of the important clinical symptoms of pulmonary tuberculosis. Based on those findings, the modified Moran's I is proposed in the data transformation process. So that all of fourteen attributes are used to build the proposed model. Table 7 shows that the proposed model can perform better than the comparative model.

6. Conclusion

In this paper, a new method for data transformation in predicting the suspect of new pulmonary tuberculosis case based on patient demographic data, anamnesis, and physical examination has been presented. A modified Moran's I was proposed. This model is in accordance with National Guidance of Tuberculosis Control [30] and International Standard for Tuberculosis Care [18]. The main advantage of the proposed model is to improve adherence to standard diagnostic path [40]. Early detection of the TB suspect is the first step in TB management and also the most effective prevention activity of TB disease transmission. In addition, the proposed model has an accuracy of 84.54%, specificity of 85.24%, and sensitivity of 84.24%. As the performance of the proposed model is competitive to the previous study, it can help to ensure which patient is TB suspected. Thus, the model helps to decrease the delay in TB diagnosis and to reduce the number of TB-caused immortality rate. The result of this study will be used in our future research, which is a diagnosis of smear-negative pulmonary tuberculosis.

References

[1] M. A. Sanchez, S. Uremovich, and P. Acrogliano, "Mining Tuberculosis Data," in Data Mining and Medical Knowledge Management : Cases and Applications, R. Bellazzi, R. Jirousek, K. Mourik, J. Paralik, L. Torgo, and B. Zupan, Eds. Hersey, New York: Medical Information Science Reference, 2009, pp. 332–349.

[2] WHO, WHO Global Tuberculosis Report 2015, 20th ed. WHO Library Cataloguing-in-Publication Data, 2015.

[3] Z. X. Zhang, L.-H. Sng, Y. Yong, L. M. Lin, T. W. Cheng, N. H. Seong, and F. K. Yong, "Delays in diagnosis and treatment of pulmonary tuberculosis in AFB smear-negative patients with pneumonia," Int. J. Tuberc. Lung Dis., vol. 21, no. 5, pp. 544–549, 2017.

[4] G. Alvarez-Uria, J. M. Azcona, M. Midde, P. K. Naik, S. Reddy, and R. Reddy, "Rapid Diagnosis of Pulmonary and Extrapulmonary Tuberculosis in HIV-Infected Patients. Comparison of LED Fluorescent Microscopy and the GeneXpert MTB/RIF Assay in a District Hospital in India.," Tuberc. Res. Treat., pp. 1–4, 2012.

[5] A. A. Imianvan and J. C. Obi, "Decision Support System for the Identification of Tuberculosis using Neuro Fuzzy logic," Niger. Ann. Nat. Sci., vol. 12, no. 1, pp. 12–20, 2012.

[6] H. M. S. C. Kusuma, "Diagnostik Tuberkulosis Baru," Sari Pediatr., vol. 8, no. 4, pp. 143–151, 2007.
[7] A. Nesredin, "Mining Patients' Data for Effective Tuberculosis Diagnosis: The Case of Menelik II Hospital," Addis Ababa University, 2012.
[8] M. Bahadori and M. H. Azizi, "Common Challenges in Laboratory Diagnosis and Management of Tuberculosis," Iran. Red Crescent Med. J., vol. 14, no. 1, pp. 3–9, 2012.
[9] Rusdah and E. Winarko, "Review on Data Mining Methods for Tuberculosis Diagnosis," in Information Systems International Conference (ISICO), 2013, pp. 563–568.
[10] A. Jain, "Extra Pulmonary Tuberculosis : a Diagnostic Dilemma," Indian J. Clin. Biochem., vol. 26, no. 3, pp. 269–73, Jul. 2011.
[11] F. C. D. Q. Mello, L. G. do V. Bastos, S. L. M. Soares, V. M. Rezende, M. B. Conde, R. E. Chaisson, A. L. Kritski, A. Ruffino-netto, and G. L. Werneck, "Predicting Smear Negative Pulmonary Tuberculosis with Classification Trees and Logistic Regression : A Cross-sectional Study," BMC Public Health, vol. 6, no. 43, pp. 1–8, 2006.
[12] C. M. Muvunyi and F. Masaisa, "Diagnosis of Smear-Negative Pulmonary Tuberculosis in Low-Income Countries : Current Evidence in Sub-Saharan Africa with Special Focus on HIV Infection or AIDS," in Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis, P.-J. Cardona, Ed. InTech, 2006, pp. 127–146.
[13] Y. Benfu, S. Hongmei, S. Ye, L. Xiuhui, and Z. Bin, "Study on the Artificial Neural Network in the Diagnosis of Smear Negative Pulmonary Tuberculosis," in 2009 World Congress on Computer Science and Information Engineering, 2009, pp. 584–588.
[14] R. Santiago-Mozos, F. Perez-Cruz, M. Madden, and A. Artes-Rodriguez, "An Automated Screening System for Tuberculosis," IEEE J. Biomed. Heal. Informatics, no. 99, pp. 1–8, Oct. 2013.
[15] A. H. van't Hoog, M. W. Langendam, E. Mitchell, F. G. Cobelens, D. Sinclair, M. M. G. Leeflang, and K. Lonnroth, "A Systematic Review of the Sensitivity and Specificity of Symptom and Chest Radiography Screening for Active Pulmonary Tuberculosis in HIV-Negative Persons and Persons with Unknown HIV Status," 2013.
[16] M. Bhatt, R. Bhaskar, and S. Kant, "Pulmonary Tuberculosis as Differential Diagnosis of Lung Cancer," South Asian J. Cancer, vol. 1, no. 1, p. 36, 2012.
[17] F. M. E. Uzoka, J. Osuji, and O. Obot, "Clinical decision support system (DSS) in the diagnosis of malaria: A case comparison of two soft computing methodologies," Expert Syst. Appl., vol. 38, no. 3, pp. 1537–1553, Mar. 2011.
[18] TB CARE I, International Standards for Tuberculosis Care, 3rd ed. The Hague: TB CARE I, 2014.
[19] A. Kusiak, K. H. Kernstine, J. A. Kern, K. A. Mclaughlin, and T. L. Tseng, "Data Mining : Medical and Engineering Case Studies," in Proceeding of the Industrial Engineering Research 2000 Conference, 2000, pp. 1–7.
[20] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. López, and R. H. Gilman, "Cough Detection Algorithm for Monitoring Patient Recovery from Pulmonary Tuberculosis," in 33rd Annual International Conference of the IEEE EMBS, 2011, pp. 6017–6020.
[21] R. Lestari, M. Ahmad, B. Alisjahbana, and T. Djatmiko, "The Lung Diseases Diagnosis Software : Influenza and Tuberculosis Case Studies in The Cloud Computing environment," in International Conference on Cloud Computing and Social Networking, 2012, pp. 1–7.
[22] M. K. Osman, M. Y. Mashor, H. Jaafar, R. A. A. Raof, and N. H. Harun, "Performance Comparison between RGB and HSI Linear Stretching for Tuberculosis Bacilli Detection in Ziehl-Neelsen Tissue Slide Image," 2009 IEEE Int. Conf. Signal Image Process. Appl., pp. 357–362, 2009.
[23] M. K. Osman, F. Ahmad, Z. Saad, M. Y. Mashor, and H. Jaafar, "A Genetic Algorithm-Neural Network Approach for Mycobacterium Tuberculosis Detection in Ziehl-Neelsen Stained Tissue Slide Images," in 2010 10th International Conference on Intelligent Systems Design and Applications, 2010, pp. 1229–1234.
[24] M. K. Osman, M. Y. Mashor, and H. Jaafar, "Detection of Mycobacterium Tuberculosis in Ziehl - Neelsen Stained Tissue Images using Zemike Moments and Hybrid Multilayered Perceptron Network," 2010, pp. 4049–4055.
[25] M. K. Osman, M. H. M. Noor, M. Y. Mashor, and H. Jaafar, "Compact Single Hidden Layer Feedforward Network for Mycobacterium Tuberculosis Detection," in 2011 IEEE International Conference on Control System, Computing and Engineering, 2011, pp. 432–436.
[26] H. Wang, C. Zhao, and F. Li, "Identification of M. tuberculosis complex by a novel hybridization signal amplification method," Proc. 2011 Int. Conf. Hum. Heal. Biomed. Eng., pp. 1085–1088, Aug. 2011.
[27] M. Lauria, "Rank-based miRNA Signatures for Blood-based Diagnosis of Tuberculosis," in Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, 2015, pp. 4462–4465.
[28] A. A. Bakar and F. Febriyani, "Rough Neural Network Model for Tuberculosis Patient Categorization," in Proceedings of the International Conference on Electrical Engineering and Informatics, 2007, no. 1, pp. 765–768.
[29] T. Uçar, A. Karahoca, and D. Karahoca, "Tuberculosis Disease Diagnosis by using Adaptive Neuro Fuzzy Inference System and Rough Sets," Neural Comput. Appl., vol. 23, no. 2, pp. 471–483, Apr. 2013.
[30] K. R. D. P2PL, Pedoman Nasional Pengendalian Tuberkulosis. Kemenkes RI, 2014.
[31] T. Asha, S. Natarajan, and K. N. B. Murthy, "A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification," J. Comput., vol. 3, no. 4, 2011.
[32] Rusdah, E. Winarko, and R. Wardoyo, "Preliminary Diagnosis of Pulmonary Tuberculosis Using Ensemble Method," in The 2nd International Conference on Data and Software Engineering (ICoDSE) 2015, 2015.
[33] T. Asha, S. Natarajan, and K. N. B. Murthy, "Data Mining Techniques in the Diagnosis of Tuberculosis," in Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis, P.-J. Cardona, Ed. InTech, 2012, pp. 333–352.
[34] T. Asha, S. Natarajan, and K. N. B. Murthy, "Effective Classification Algorithms to Predict the Accuracy of Tuberculosis-A Machine Learning Approach," Int. J. Comput. Sci. Inf. Secur., vol. 9, no. 7, pp. 89–94, 2011.
[35] T. Asha, S. Natarajan, and K. N. B. Murthy, "Diagnosis of Tuberculosis using Ensemble methods," in 3rd IEEE

International Conference on Computer Science and Information Technology (ICCSIT), 2010, pp. 409–412.
[36] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, 3rd ed., vol. 1. Morgan Kaufmann Publishers, 2012.
[37] P. A. Rogerson, Statistical Methods for Geography: A Student's Guide, 4th ed. SAGE, 2014.
[38] A. Ali, M. Elfaki, and D. N. A. Jawawi, "Using Naïve Bayes and Bayesian Network for Prediction of Potential Problematic Cases in Tuberculosis," Int. J. Informatics Commun. Technol., vol. 1, no. 2, pp. 63–71, 2012.
[39] H. Dağ, K. E. Sayın, I. Yenidoğan, S. Albayrak, and C. Acar, "Comparison of Feature Selection Algorithms for Medical Data," in 2012 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2012, pp. 1–5.
[40] T. Y. Aditama and M. Subuh, Strategi Nasional Pengendalian TB di Indonesia 2010-2014. Dirjen P2PL Kemenkes RI, 2011.

Author's Profile

Rusdah. She received her S1 degree (S.Kom) in Information System and M.Kom in Software Engineering from Universitas Budi Luhur. She is a Ph.D. student at Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. She is a lecturer at Department of Information System, Faculty of Information Technology, Universitas Budi Luhur. Her research interests are decision support, data warehousing, and data mining.

Edi Winarko. He received his Undergraduate degree (Drs.) in Statistics from Universitas Gadjah Mada, Indonesia; Master degree in Computer Sciences (M.Sc.) from Queen's University, Canada, and Ph.D. in Computer Science from Flinders University, Australia. He is a lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada. His research interests are Data Warehousing, Data Mining, and Information Retrieval.

Retantyo Wardoyo. He received his undergraduate degree (Drs.) in Mathematics from Universitas Gadjah Mada, Indonesia; MSc. in Computer Science from The University of Manchester, UK, and Ph.D. in Computation from University of Manchester Institute of Science and Technology, UK. He is a lecturer at the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. His research interests are Intelligent System and Computation.