Identifying Irregularity Electricity Usage of Customer Behaviors using Logistic Regression and Linear Discriminant Analysis

Armin Lawi, Supriyadi La Wungo, Salama Manjang

Department of Computer Science, Post-graduate Program of Electrical Engineering, Dept. Electrical Engineering
Universitas Hasanuddin, Makassar, Indonesia
Email: armin@unhas.ac.id, supriyadi.la.wungo@gmail.com. firmanaziz88@gmail.com, arvasit@gmail.com

***Abstract*—This study aims to implement a machine learning technique in identifying the irregularities of customer behavior on the use of prepaid electricity pulses. The methods used are Linear Discriminant Analysis and Logistic Regression. The performance of the classification system will be evaluated using the 10-fold cross-validation technique. Validation results are measured using accuracy, precision and recall values. In this research shows that the use of machine learning technique has a good performance in classification of electrical consumption behavior. Experimental results with the different amount of data testing indicate that Logistic Regression method has high accuracy, precision, and recall value when compared with Linear Discriminant Analysis that is 100%. This is due to Logistic Regression method can predict irregularities accurately because the addition of the amount of data does not affect the performance of the method.**

Keywords*—*abnormal electricity identification, linear discriminant analysis, logistic regression

I. Introduction

Prepaid Electricity is the latest product issued by electric power supply company that has some advantages and ease of use because the electricity consumption is fully controlled by the customer so that more comfortable and controlled [1]. However, with the ease that has been provided by energy provider companies are still found by many customers who indicated theft electricity in the field such as the customer to connect the power cable without going through kWh meter and change the pattern of wiring on kWh meter [2, 3] so the counter on kWh does not work and the kWh meter keeps working even though the remaining credit token/voucher has run out. Based on statistical data of PLN and Electricity of the Ministry of Energy and Mineral Resources, the depreciation of electric energy averages about 7% per year and the estimated state losses due to the depreciation of Rp 1.5 trillion per year[4, 5].

Control of energy provider companies to customers is minimal. This happens because the company no longer uses the service of recording and manual checking each month. To monitor the misuse of electricity usage transactions that are considered unreasonable then the provider of electrical energy to filter the purchase transaction period of token/voucher. But this way still has the weakness because if the customer only does token purchase transaction/voucher with nominal value 20.000 and transaction amount only once in 1 month even though the target of 100.000 / month usage is not fulfilled hence the customer is considered to have reasonable transaction by the company. Therefore, to overcome the problem the authors propose to do also filter by the number of transactions and nominal purchase of tokens/vouchers of customers.

Several related works on the abuse of electrical energy usage has been proposed by previous researchers as follows. Depuru, et al., in [6] proposes the detection of electricity theft through energy consumption patterns of some customers involved in the theft. They used classification method using Support Vector Machine (SVM) with kernel rbf, value C = 1, and probability estimation = 0, and the resulted accuracy was 98.4%. Another computational technique for fraudulent classification in electricity consumption through a power consumption profile has been posed in [7]. The method used Fuzzy C-Means (FCM) to group the customers who have the same pattern from within database and then process the customer classification which is not normal. Classification based on Fuzzy C-Means was classified using 2 metrics namely Assertiveness and Sensitivity. The accuracy of both metrics is 0.745 and 0.100, respectively. Babu, et al., in [8] also used FCM to investigate non-technical loss detection by monitoring the profile of irregular customer consumption in power distribution systems. Their Fuzzy based classification method detected non-technical losses with accuracy of 80%. The further research in [9] proposed a comprehensive top-down scheme based on the Decision Tree (DT) and SVM to detect and locate real-time power theft at any power level in transmission and distribution. DT is used to calculate the energy consumption that has been used by the customer based on the value of the attribute. Then the results of this calculation will be used as input on SVM to classify normal or abnormal customers in the use of electrical energy. The

Corresponding Author: supriyadi.la.wungo@gmail.com

combination of both methods has an accuracy rate of 92.5% and a very low false positive of 5.12%.

II. PRELIMINARIES

A. *Customer Electricity Usage Behavior Problems*

Currently, the control system of electric power supply companies to customers is very minimal. This is because they no longer use the service of recording and manual checking every month. To conduct monitoring of customers indicated to abuse in the transaction of electrical energy usage that is considered unreasonable, the provider of electrical energy conducting filter mechanism against the purchase transaction period of the token or voucher. However, the filter made by the company still has weaknesses and manual checking takes a long time and costs a lot [9].

Some of the causes of unreasonable customer behavior are the customer connecting the power cord without going through the kWh indicator, putting the magnet above the indicator to reduce disc rotation and changing the wiring pattern at the kWh indicator [2, 3, 6], therefore the counter on the indicator does not work and it keeps working even though the remaining credit token/voucher has run out.

The impact of the misuse of electrical energy for the company is a material loss that reaches millions and even billions of dollars due to power loss, whereas for consumers the power shared by other consumers is reduced, there is often a power outage, fire may occur due to short circuit [6].

B. *Identification variable*

In the case of the transaction history of purchase of pulses there are several variables used to identify irregularities of electricity usage of each customer as follows.

- *Customer id*: used to identify how many customers make credit purchase transactions in a month.
- *Installed Power* (kWh): to identify the power consumption that customers use based on their power.
- *Total Credit in Voucher*: Number of kWh pulse in the voucher when transaction by the customer.
- *Total Power Consumption*: The difference between max usage and min usage to know the amount of usage in a month.
- *Duration of Voucher Usage*: Long duration of voucher usage by the customer during a certain period.
- *Status*: The power consumption usage of customers utilizes vouchers.

The selection of these five attributes is considered to represent each customer in identifying usage patterns, and to support the analysis of electrical data as well as to detect the morbidity of existing systems. The reason for the one-month span is to minimize the losses in the company, the electricity payment by the customer is done monthly, and the transaction of the customer's credit purchase by the company is done per month. Therefore the determination of this time span is very relevant for evaluation [7, 8].

C. *Identification of irregularities*

Identification of irregularities is a technique of searching data to find an object that does not meet certain criteria when compared with other objects. These criteria can be known through the attribute value of an object. An object is said to be unnatural if the value of an attribute possessed by that object does not meet certain criteria compared to the attribute values of another object that meets the criteria. Therefore, to determine the imperfection of an object must select the appropriate attributes [3].

Identification of irregularities can be done by detecting a fraud or theft. In the case of electric theft, irregularities can be identified by looking at the usage patterns of electricity consumption for customers who have above average usage of actual usage targets. Customer behavior patterns of electricity consumption can be seen in the historical data of energy consumption. In large-scale companies, historical data is used to identify customer usage patterns that are considered to have an unnatural transaction in electricity consumption. From the morbidity is then indicated the theft / electric fraud committed by the customer [10, 11].

D. *Classification techniques*

1) *Linear Discriminant Analysis*

Fisher Linear Discriminant Analysis (LDA) is a classification method used in statistics, pattern recognition and machines learning to find linear combinations of features that characterize or separate two or more object classes. For the LDA case, the transformation is based on maximizing the ratio of "between class variance" to "within the variance class" with the aim of reducing a variation of data in the same class and increasing the separation between classes. LDA is to find the discriminant vector so that the ratio of the distance between the classes to the distance in the class is maximized [12]. For that purpose Fisher LDA is used to maximize the following equation (1) by obtaining scalar y by projecting sample *X*.

$$y = \theta^T X \tag{1}$$

The following is the LDA implementation for the case of two classes; defined dataset D as $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, where N_1 in class C_1 and N_2 in C_2, and the average vector of two classes in space *X* is given by equation (2).

$$u_k = \frac{1}{N_k} \sum_{i \in C_k} x^{(i)} \text{ where k = 1.2} \tag{2}$$

and in *y* space,

$$\hat{u}_k = \frac{1}{N_k} \sum_{i \in C_k} y^{(i)} = \frac{1}{N_k} \sum_{i \in C_k} \theta^T = \theta^T u_k,$$

where k =1.2 (3)

Then, we determine the size for the separation between the two classes by choosing the distance between the existing projection means in *y* space, therefore the variance between classes can be given in the following equation (4)

$$\hat{u}_2 - \hat{u}_1 = \theta^T (u_2 - u_1), \tag{4}$$

and to determine the variance within the class for each class $\sum_{i=1}^{m} a_i$ is given in the following equation (5)

$$\hat{s}_k^2 = \sum_{i \in C_k} \left(y^{(i)} - \hat{u}_k\right)^2 \text{where } k = 1.2. \qquad (5)$$

After getting the variance between the class and the variance within the class, then we determine the objective function of *J(θ)* as follows,

$$\max \; J(\theta) = \frac{(\hat{u}_2 - \hat{u}_1)^2}{\hat{s}_1^2 - \hat{s}_2^2} \qquad (6)$$

To maximize the objective function *J*, we look for which projections of the projected classes are very close to each other and at the same time the means are projected as far as possible.

To find the optimal point of θ^*, the function $J(\theta)$ is presented as a function θ. The following are some scatter of the variance.

- Scatter in feature space X:

$$S_k = \sum_{i \in C_k} \left(x^{(i)} - u_k\right)\left(x^{(i)} - u_k\right)^T$$

- Scatter matrix within the class: $S_W = S_1 - S_2$
- Scatter matrix between classes:

$$S_B = (u_2 - u_1)(u_2 - u_1)^T$$

On the equation (6), the scatter of the projection y can be exposed as a function of the matrix scatter in the feature space x so as to obtain Fisher criteria for S_w and S_B as follows:

$$J(\theta) = \frac{\theta^T S_B \theta}{\theta^T S_W \theta} \qquad (7)$$

To maximize the function of object J is meant to decrease and set it to 0, and then divided by $\theta^T S_W \theta$ to get direction θ which is optimum θ^*,

$$\theta^* \propto S_W^{-1}(u_2 - u_1) \qquad (8)$$

This is called Fisher Discriminant Linear, which is to project data to one dimension with function $y = \theta^{*T} X$ [13].

Non-linear classification techniques based on Fisher discriminant using kernel tricks enable efficient calculations within the feature space. So the classification in the feature space corresponds to the non-linear decision in the input space. Determined Φ is a non-linear mapping for some feature space F. To determine linear discriminant in feature F feature then we need to maximize

$$J(w) = \frac{w^T S_B^{\Phi} w}{w^T S_W^{\Phi} w} \qquad (9)$$

where $w \in F$ and S_B^{Φ} and S_w^{Φ} is Matrix that exists in space feature F, that is:

$$S_B^{\Phi} := (m_1^{\Phi} - m_2^{\Phi})(m_1^{\Phi} - m_2^{\Phi})^T$$

$$S_w^{\Phi} := \sum_{i=1.2} \sum_{x \in X_i} (\Phi(x) - m_i^{\Phi})\,(\Phi(x) - m_i^{\Phi})^T$$

with $m_i^{\Phi} := \frac{1}{l_i} \sum_{j=1}^{l_i} \Phi\left(x_j^i\right)$

To find Fisher discriminant in feature space F, formula (9) will be converted to product dot function from input pattern and then will be changed to the kernel function. To build the kernel any w∈F solution should be within the range of all the training samples in space F feature. So that it can find the deployment for w in on the following functions:

$$w = \sum_{i=1}^{l} \alpha_i \, \Phi(x_i) \qquad (10)$$

by using the spread (10) and the definition of m_i^{Φ} it will get the following equation:

$$w^T m_i^{\Phi} = \frac{1}{l_i} \sum_{j=1}^{l} \sum_{k=1}^{l_i} \alpha_j \, k\left(x_j, x_k^i\right)$$

$$= \alpha^T M_i \qquad (11)$$

which is defined $m_i^{\Phi} = \frac{1}{l_i} \sum_{k=1}^{l_i} k\left(x_j, x_k^i\right)$ and replace the dot product functions with the kernel functions. Furthermore, considering the numerator in the function (9) can use the definition of S_B^{Φ} and (11) can be rewritten as follows:

$$w^T S_B^{\Phi} w = \alpha^T M \alpha \qquad (12)$$

where $M := (M_1 - M_2)(M_1 - M_2)^T$. Given the denominator using (10) with the definition of m_i^{Φ} and the same transformation on (12) it will be obtained

$$w^T S_W^{\Phi} w = \alpha^T N \alpha \qquad (13)$$

where $N := \sum_{j=1.2} K_j \left(I - 1_{l_j}\right) K_j^T, K_j$ is a matrix $(l \times l_j)$ with $(K_j)_{nm} := k(x_n, x_m^j)$ This is the kernel matrix for class j, I is identity and 1_{l_j} matrix with all entries $1/l_j$. The combination of equations (12) and (13) can be found Fisher linear discriminant in space F by maximizing:

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha} \qquad (14)$$

this problem can be solved by finding the main eigenvector of $N^{-1}M$. This approach is called the *Fisher Discriminant Kernel (KFD)*. The new projection pattern from x to w is determined by the following equation:

$$\left(w.\Phi(x)\right) = \sum_{i=1}^{l} \alpha_i \, k(x_i, x) \qquad (15)$$

However, the numerical problem that causes the N matrix is not positive is required a way to control the capacity of F space. For this purpose, it will add an identity matrix to N that is replacing N with N_μ where:

$$N_\mu := N + \mu I \qquad (16)$$

Thus, the numerical problem can be solved because if N is of great value then N_μ will be positive, decreasing bias in the sample by eigenvalue, enacting regulation at $\|\alpha\|^2$ (by maximizing equation (14)) [14].

2) *Logistic Regression*

Regression analysis in statistics is one method to determine a cause-effect relationship between one variable with other variables. The variable is divided into two variables *X* and *Y*. Variable *X* (often described in graphics as a base, or *X* axis) is called by various terms, such as explanatory variables, explanatory variables, and independent variables. While the variable *Y* "affected" is known as influenced variable, dependent variable and dependent variable. Both of these variables can be random variables, but the variables that are affected must always be random variables.

In logistic regression, if the response variable consists of two categories, e.g., *Y* = 1 represents the result obtained "success" and *Y* = 0 represents the result obtained "failed" then logistic regression using binary logistic regression. According to [15] the variable y is more accurately said to be an indicator variable and satisfies the Bernoulli distribution. The opportunity distribution function for y with the π_i parameter is

$$f(y_i;\pi_i) = \begin{cases} \pi_i(1-\pi_i)^{1-y_i} for\ y_i = 0,1 \\ 0 \qquad for\ the\ other\ y\_1 \end{cases}$$

with $\pi_i = P(Y_i = 1)$. From the distribution function can be obtained the average

$$E(Y) = 1.P(Y=1) + 0.P(Y=0) \\ = P(Y=1)$$

Suppose this probability is denoted as $\pi(x)$ which depends on the explanatory variable $X = (X_1, \dots, X_k)$ with $E(y) = \pi$ and $0 \le \pi \le 1$, thus we obtain

$$E(Y^2) = 1^2\pi(x) + 0^2[1-\pi(x)] = \pi(x)$$

and the variance of *Y* is

$$V(Y) = E(Y^2) - [E(Y)]^2 = \pi(x)[1-\pi(x)]$$

In general, the logistic regression probability model involving several predictor variables (*x*) can be formulated as follows:

$$E(y|x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}} \qquad (17)$$

where $E(y|x)$ is a sum of $\pi(x)$. The function $\pi(x)$ is a non-linear function, therefore it is necessary to perform a logit transformation to obtain a linear function in order to see the relationship between the response variable (y) and the predictor variable (x). The logit form of $\pi(x)$ is expressed as $g(x)$ as follows,

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (18)$$

Equations (1) and equation (2) are substituted, and we obtain

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k. \quad (19)$$

In order to obtain an estimate of logistic regression parameters, we use *Maximum Likelihood Estimation* (MLE). Basically, the MLE method gives the estimated value of β by maximizing its likelihood function [15] as expressed in equation (20)

$$f(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad (20)$$

Since each observation is assumed to be independent, then the likelihood function is the multiplication of each likelihood function,

$$l(\beta) = \prod_{i=1}^{n} f(x_i) \qquad (5)$$

and its likelihood logarithm is expressed as

$$l(\beta) = ln[l(\beta)] \\ = \sum_{i=1}^{n}\{y_i.ln(\pi(x_i)) + (1-y_i)ln(1-\pi(x_i))\} \quad (6)$$

To obtain the value β, we maximizing the value of $l(\beta)$ and differentiating $l(\beta)$ with respect to β and equating it with 0.

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \qquad (7)$$

and thus, likelihood equation is

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \qquad (8)$$

E. *Performance Evaluation*

In general, the confusion matrix is one of the means used to measure the performance of a classification. The confusion matrix is a table used to display performance results from the classification of normal and abnormal data. Confusion matrix also called contingency table can be seen in Table I [16].

TABLE I. CONFUSION MATRIX

Actual × Prediction	**Normal**	**Abnormal**
Normal	TP	FN
Abnormal	FP	TN

Using the confusion matrix in Table 1, we can evaluate the accuracy, precision and recall as follows.

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

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

TP is the number of normal data that is correctly predicted as a normal class. FN is the number of normal data that is predicted as an abnormal class. TN is the number of abnormal data that is correctly predicted as an abnormal class, and FP is the number of abnormal data that is predicted as a norma.

Accuracy is a comparison of the correct classification results of the total classification. Precision is the correct comparison of normal data classification of the total classification of mismatch. The recall is a classification comparison detecting an irregularity to total impropriety. If the model shows a high recall value, then this model is reliable. This is because the model rarely makes a mistake in diagnosing normal data to an abnormal class. If the recall value shows a high value then the model's performance is very good since the model has a high accuracy in diagnosing normal data of morbidity.

III. Results

In this study, using the dataset history of purchase transaction pulses obtained from the company PT. PLN Mamuju, West Sulawesi, Indonesia. The transaction history data set of the pulse has a real-type attribute with a total of 10000 lines of data consisting of 7783 fair transactions and 2217 unusual transactions and has 4 attributes and 2 classes. The dataset will be divided into 10 sections using k-fold cross validation as data validation, consisting of 9 sections will be used for training data and 1 part will be used for data testing. Experimental research results are processed using an ASUS laptop with an Intel® Core i5 processor CPU M 460 @ 2.53 GHz, 4.00 GB of RAM, and the Linux operating system 64-bit Ubuntu version 6.10. While the software to develop applications is Python version 2.7.

TABLE II. Accuracy, Precision, Recall Using LDA

Performance	**Percentage of data testing**					
	20	33	45	55	67	80
Accuracy	0.99	0.99	0.98	0.93	0.93	0.93
Precision	0.92	1.00	0.99	0.73	0.72	0.72
Recall	1.00	1.00	1.00	0.99	0.99	0.99

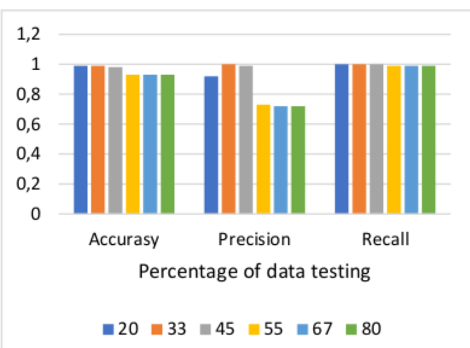

Fig. 1. Result of accuracy, precision, and recall using LDA.

Table II, and Fig. 1, shows that the amount of test data used can affect the accuracy of the Linear Discriminant Analysis method using Singular value decomposition in predicting the irregularities of electricity purchase transactions. The more test data used, the lower the accuracy result. Likewise, the precision and recall value decreases if the test data increases. This means that the addition of test data causes the results of FP and FN also increased. The addition of FP and FN will reduce the value of TP and TN. Therefore, if precision and recall decreases, then accuracy will also decrease.

TABLE III. Accuracy, Precision, Recall Using LR

Performance	**Percentage of data testing**					
	20	33	45	55	67	80
Accuracy	1.00	1.00	1.00	1.00	1.00	1.00
Precision	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00

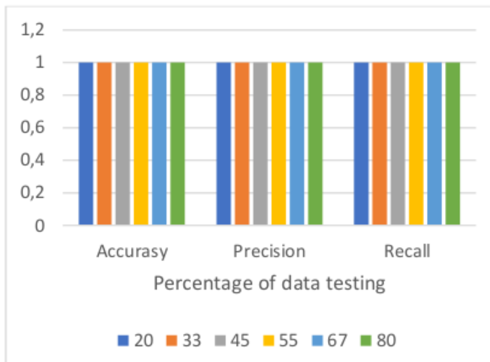

Fig. 2. Result of accuracy, precision, and recall using LR

Table III, and Fig 2, shows that the addition of test data does not affect the accuracy of classification of LR method by using Maximum Likelihood Estimation (MLE) kernel to predict the transaction of irregularities purchase of electricity with an accuracy level of 1.00%. Likewise, the precision and recall value did not change at 1.00%. This shows that with the addition of test data, the value of False Positive and False Negative remain 0. Because in each addition of precision value test data, accuracy and recall remain, it can be said that LR method can predict irregularities accurately.

IV. Conclusion

In this study, we proposed classification methods that are used Linear Discriminant Analysis (LDA) and Logistic Regression (LR). The results showed that the LDA method has the accuracy rate of 99%, precision value of 92% and recall value of 100%. Whereas the LR method reaches a high level of accuracy, precision and recalls with a value of 100%. However, deep investigation is still needed to know which method with the best accuracy, precision and recall values by the some data simulation characteristics. Based on our result, we found that accuracy using LR in classifying the behavior of prepaid credit customers through purchase transaction history has a good accuracy level when compared with LDA.

REFERENCES

[1] I.S. Jhra, S. Subir , and V. Agarwal. "Advanced metering infrastructure analytics—A Case Study." In *Power Systems Conference (NPSC), 2014 Eighteenth National*, pp. 1-6. IEEE, 2014.

[2] Dangar, Bharat, and S.K. Joshi. "Electricity theft detection techniques for metered power consumer in GUVNL, GUJARAT, INDIA." In *Power Systems Conference (PSC), 2015 Clemson University*, pp. 1-6. IEEE, 2015.

[3] Ford, Vitaly, Ambareen Siraj, and W. Eberle. "Smart grid energy fraud detection using artificial neural networks." In *Computational Intelligence Applications in Smart Grid (CIASG), 2014 IEEE Symposium on*, pp. 1-6. IEEE, 2014.

[4] Sekretariat Perusahaan PT. PLN, "Statistik PLN 2014," no. 2701, p. iv, 2015.

[5] K. ESDM, "Statistik ketenagalistrikan 2014," *Stat. Ketenagalistrikan*, p. 5, 2014.

[6] Depuru, S.S.S. Reddy, L. Wang, and V. Devabhaktuni. "Enhanced encoding technique for identifying abnormal energy usage pattern." In North American Power Symposium (NAPS), 2012, pp. 1-6. IEEE, 2012.

[7] E.W.S. Angelos, O.R. Saavedra, O.A.C. Cortés, and A.N. de Souza. "Detection and identification of abnormalities in customer consumptions in power distribution systems."IEEE Transactions on Power Delivery 26, no. 4 (2011): 2436-2442.

[8] Babu, T. Vandana, T.S. Murthy, and B. Sivaiah. "Detecting unusual customer consumption profiles in power distribution systems—APSPDCL." In Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, pp. 1-5. IEEE, 2013.

[9] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra. "Decision tree and SVM-based data analytics for theft detection in smart grid." IEEE Transactions on Industrial Informatics 12, no. 3 (2016): 1005-1016.

[10] Ch.Cody, V. Ford, and A. Siraj. "Decision Tree Learning for Fraud Detection in Consumer Energy Consumption." In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on, pp. 1175-1179. IEEE, 2015.

[11] S.Y. Han, J.G. No, J.-H. Shin, and Y.J. Joo. "Conditional abnormality detection based on AMI data mining." IET Generation, Transmission & Distribution 10, no. 12 (2016): 3010-3016.

[12] Zh. Liang, Y. Li, and P. Shi. "A note on two-dimensional linear discriminant analysis." Pattern Recognition Letters 29, no. 16 (2008): 2122-2128.

[13] C. Li and B. Wang, "Fisher Linear Discriminant Analysis," pp. 1–6, 2014.

[14] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers. "Fisher discriminant analysis with kernels." In Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop., pp. 41-48. IEEE, 1999.

[15] S.R. Hosmer DW, S. Lemeshow, *Applied Logistic Regression, 3rd Edition*. 2013.

[16] T. Fawcett, "An introduction to ROC analysis." Pattern recognition letters27, no. 8 (2006): 861-874.

C-39

ORIGINALITY REPORT

42%	38%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eudl.eu Internet Source	38%
2	Submitted to Kaplan University Student Paper	1%
3	Budi Santosa, Andiek Sunarto. "Development of Kernel Fisher Discriminant Model Using the Cross-Entropy Method", 2009 International Conference of Soft Computing and Pattern Recognition, 2009 Publication	1%
4	Submitted to University of Surrey Student Paper	1%
5	Tong Wang, Yan Xia Pang, Yue Ping Wu, Yi Du. "Research on the Prediction of Protein Functional Sites Based on the Dimension Reduction Algorithm", Applied Mechanics and Materials, 2013 Publication	1%
6	Submitted to Vels University Student Paper	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		