Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization

by Moses De Rosal

Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization

Muhammad Misdram
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
misdram@mhs.dinus.ac.id

Edi Noersasongko
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
edi.noersasongko@dsn.dinus.ac.id

Abdul Syukur
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
abah.syukur01@dsn.dinus.ac.id

Purwanto
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
purwanto@dsn.dinus.ac.id

De Rosal Ignatius Moses Setiadi
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
moses@dsn.dinus.ac.id

***Abstract* –The classification method in data mining requires a good learning process to get optimal accuracy. This can be done if the dataset used is ideal, balanced, and has a lot of records, but in reality, it is difficult to get such a dataset. The imputation method is one way to fill in missing values, in a dataset that is not ideal. A large number of missing values can reduce the number of records in the learning process and affect accuracy. This research aims to analyze the effects of zero and mean imputation methods on classification accuracy in small datasets using the Naïve Bayes classifier (NBC) and NBC which have been optimized with Particle Swarm Optimization (PSO). Tests were carried out on five types of datasets originating from the UCI database, where one of the datasets did not require an imputation method because it did not have a missing value. Based on the results of the PSO testing proven to be able to improve the accuracy of the NBC classification on all datasets. While the imputation method can improve classification accuracy up to 4.33% in Biomarker datasets.**

Keywords— Imputation Methods, Unbalanced Datasets, Small Dataset, Classification, Naïve Bayes, Particle Swarm Optimization

I. INTRODUCTION

A classification is a form of data mining that requires the learning process of a dataset. To produce good learning an ideal dataset is needed, where the amount of data in the data used must be ideal, balanced, and without missing values in the part of the record attribute[1]. To solve with missing values, the simplest approach is to delete records whose values for each attribute are incomplete and to learn based only on records with complete values. However, this method can lead to missing lots of records and data deviations that can lead to incorrect or biased conclusions[2]. Classification problems become more severe when faced with a dataset with a small amount, not balanced, and has a value that is empty on some attributes, outliers, or data formats that are not compatible with the system[3]. Missing values in a small dataset will result in less data that can be used for learning so that the prediction of the classification results from that data will decrease [4].

The imputation method is a method that is applied at the preprocessing stage to overcome the problem of missing values[5], which aims to improve the performance of classifications in datasets that are not ideal especially for small numbers of datasets [6]. Several imputation methods have been published and many are proposed, one of which is to use the most frequently occurring value or zero value as a constant replacement for the missing value. [7]. Besides, there are also imputation methods with mean values [8], so the mean value resulting from the calculation of the imputation method is used as the imputation value on the attribute whose value is missing [5].

In the mean imputation method, the process is to use the mean to fill in some algorithms to estimate the right data and compare the predicted values on different algorithm estimates, the goal is to fill in the data that has the missing value and get the complete data[5], [9]. While the imputation method uses the most frequent or zero values, it is done by replacing the value of the empty attribute with the zero value during the preprocessing process [5].

In the next stage is the classification process. One of the most widely used classification methods is Naïve Bayes (NB). NB has the advantage of being an independent analysis model and can learn well with small amounts of datasets. However, if the dataset used has a lot of missing values, it will affect the results of predictions and the reliability of any classification model[10]. Related research conducted by Li et al [9] proposed the NB method combined with Particle Swarm Optimization (PSO) to improve classification predictions. The test results generated with an accuracy of 85.08%, and after optimization with PSO predictions can increase to 89.31%. Other research conducted by Dulhare et al [11] NB also proposes to classify heart disease, which is relatively small with 14 features of the UCI repository. The results of classification using NB which is optimized with PSO produces an accuracy of 87.91%, whereas if using a genetic algorithm (GA) as an optimizer an accuracy of 86.29% results

Based on some literature related above, this research was conducted to analyze the zero and play imputation method in the classification using Naive Bayes and PSO on data sets in small and unbalanced numbers, in which there are also attributes whose values are missing. The data set used uses five public datasets from the UCI repository[12] which has been widely used in previous researches. In this way, the logic of classification performance can be maintained with the imputation method, because with this method it can avoid the less number of datasets used in the learning process can also reduce the possibility of increasing the imbalance of the dataset.

II. THEORIES

A. *Unbalanced and Small Dataset*

Unbalanced data consists of several conditions because the distribution of data is not balanced, the number of data classes is more or more than the number of other data classes. Data class groups that are less known as minority groups, other data class groups are called majority groups. Small dataset problems still arise in certain areas that make analysis and decisions difficult to make [6]. There may be two types of imbalance in a data set[8]. One of them is the imbalance between classes, in this case, some classes have more examples than others [3], [4]. The other is an imbalance in class, in this case, several subsets of one class have far fewer examples than other subsets of the same class [5].

B. *Imputation method*

Many real-world datasets may contain missing values due to variety. They are often coded as NaN, blank or other placeholders. Model training with datasets that have a lot of missing values can drastically impact the quality of machine learning models. Some algorithms like Sci-kit learn estimators assume that all values are numeric and have and have meaningful values. One way to overcome this problem is to get rid of observations that have missing data. However, you risk losing data points with valuable information. A better strategy is to blame the missing value. In other words, we need to infer the values that are missing from existing pieces of data. There are three main types of missing data: missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR))[5].

In this study to overcome the loss of data on each attribute by not eliminating blank data since it is part of a small size data, then imputation is performed. In this study only uses two ways of imputation out of six kinds of imputation, among others

1) *Zero Imputation*

Zero or constant imputation is to replace the lost value directly with zero or the constant value specified. Example Fig. 1 way to determine the imputation with a zero / constant value [5], where the empty attribute value is marked with the NaN symbol, then the NaN sign is replaced by the constant 0 value.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0) →

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Fig. 1. Sample of zero imputation

2) *Mean Imputation*

In this way, it works by calculating the mean of the values that are not lost in an attribute of some data and then replacing the missing values in each attribute separately and separately from the others. This can only be used with numerical data. Example in Fig. 2 is a way to determine the imputation with the Mean value[5], where the empty attribute value is marked with the NaN symbol, then the NaN sign is replaced by the average value of the number of values in one column divided by the number of values in that column.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean() →

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Fig. 2. Sample of the Mean imputation process

C. *Classification*

Classification is a very important sequence in the data mining community [13]. Classification is one of the predictive data mining techniques that makes predictions about data values using known results found from different data sets. The problem of accuracy of many classification algorithms is known to experience a decrease in information when faced with unbalanced data, for example, when the distribution of samples across classes is very skewed[14].

Naive Bayes (NB) classifier with Particle Swarm Optimization (PSO) algorithm is used which will classify the data set and then conduct an independent analysis [10]. The NB classifier combined with the PSO hybrid feature selection method proved to be the best feature of selection capability without lowering the classification prediction. The architecture NB classifier combined with hybrid PSO feature selection proves to be the best feature selection capability without reducing classification prediction. This method also proved to be the most suitable method for mining large structural data in much less computational time[10].

Naïve Bayes classifiers are simple probabilistic classifiers based on using the Bayes theorem with strong (Naïve) independence assumptions between features [15], [16]. Bayes' theorem can be calculated by Eq. 1[16]

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right) \times P(H)}{P(X)} \quad (1)$$

$$P\left(\frac{H}{X}\right) = P\left(\frac{X}{H}\right) \times H \quad (2)$$

Where X is data with unknown classes, H is hypothesis X data for a particular class, $(\frac{H}{X})$ is H hypothesis probability based on condition X or posterior probability, $P(H)$ is the probability of hypothesis H (previously probable), $P\left(\frac{X}{H}\right)$ probability of X based on conditions that are hypothesis H, $P(X)$ is the probability of X

While the Particle Swarm Optimization (PSO) method is an optimization technique developed by Kennedy and Eberhart which is explained in [17]. The PSO algorithm is a very simple, effective way to optimize various functions, where PSO can be calculated with Eq 3, 4, and 5.

$$V_i^{t+1} = WV_i^t + C_1 \times R_1 \times (P_i^t - X_i^t) + C_2 \times R_2 \times (-X_i^t) \quad (3)$$

$$W = W_{max} - \frac{W_{max} - W_{min}}{itermax_x} \times iter \quad (4)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (5)$$

Where V_i^{t+1} is the velocity of the i dimension in the t iteration + 1, W is inertia weight, $C1$ is 1st acceleration coefficient value, C2 is 2nd acceleration coefficient value, R is the random value of [0, 1] X_i, t is the position of the i^{th} dimension in the t-iteration, P_i^t is the value of the t-dimension.

III. Research Method

In this study, several steps are carried out, which are illustrated in Fig. 3.

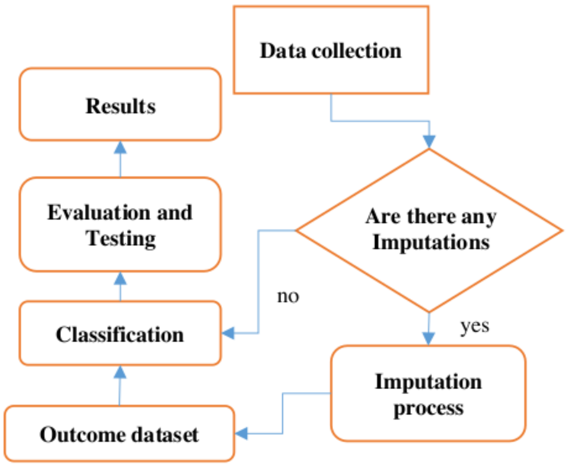

Fig. 3. Research method

A. Data collection

The data set used in this study is five medical data sets, namely hepatitis, immunotherapy, echocardiogram, biomarkers and chronic kidney, which are taken from the UCI repository [12], where all of them include data sets with relatively small and unbalanced amounts in each class. But specifically on the immunotherapy dataset, there is no missing value so that the imputation process is not carried out. This dataset is only used as a comparison of the NB method performance and NB method which is optimized with PSO.

B. Imputation Process

The imputation method is part of preprocessing data, wherein this research two methods will be used, namely zero / constant imputation and mean imputation. The goal is to provide values from data sets that are empty to valuable to optimize the results of predictions. For example in Table 1 is a sample Hepatitis dataset that contains an empty value marked with a "?", where the results of the zero value imputation are produced in Table 2 and the results of the mean value imputation method are presented in Table 3, where the media value can be calculated by Eq. 6[16].

$$Y = \frac{1}{n}\sum_{J=1}^{J} yj \quad (6)$$

Where Y is the mean value of the sum calculation yi of the number of values of variable J divided by n, J is some attributes that are valued in a group of fields or records, n are many attributes in one field or record, yj is a sum of all values from the number of attribute J.

TABLE I. Original Sample Record of Hepatitis Dataset

Class	Age	Sex	Liver Big	Liver Firm	Spleen Palpable	Varices	Bilirubin	Sgot	Albumin	Protime
1	51	1	2	2	1	2	?	0	00:00:00	0
1	39	1	2	1	2	2	02:30	98	03:08:00	40
1	62	1	?	2	2	2	01:00	60	00:00:00	0
1	37	1	2	2	2	2	0,041667	28	04:02:00	0
1	57	1	2	?	2	2	04:10	48	02:06:00	73
1	34	1	1	1	2	2	0,138889	182	00:00:00	0
1	58	1	1	1	1	2	02:00	242	03:03:00	0

TABLE II. Sample Record of Hepatitis Dataset After Zero Imputations

Class	Age	Sex	Liver Big	Liver Firm	Spleen Palpable	Varices	Bilirubin	Sgot	Albumin	Protime
1	51	1	2	2	1	2	0	0	00:00:00	0
1	39	1	2	1	2	2	02:30	98	03:08:00	40
1	62	1	0	2	2	2	01:00	60	00:00:00	0
1	37	1	2	2	2	2	0,041667	28	04:02:00	0
1	57	1	2	0	2	2	04:10	48	02:06:00	73
1	34	1	1	1	2	2	0,138889	182	00:00:00	0
1	58	1	1	1	1	2	02:00	242	03:03:00	0

TABLE III. SAMPLE RECORD OF HEPATITIS DATASET AFTER MEAN IMPUTATIONS

Class	Age	Sex	Liver Big	Liver Firm	Spleen Palpable	Varices	Bilirubin	Sgot	Albumin	Protime
1	51	1	2	2	1	2	0,083333	0	00:00:00	0
1	39	1	2	1	2	2	02:30	98	03:08:00	40
1	62	1	1	2	2	2	01:00	60	00:00:00	0
1	37	1	2	2	2	2	0,041667	28	04:02:00	0
1	57	1	2	1	2	2	04:10	48	02:06:00	73
1	34	1	1	1	2	2	0,138889	182	00:00:00	0
1	58	1	1	1	1	2	02:00	242	03:03:00	0

C. Classification

There are two proposed classification methods, the first method is a classification with NB that is optimized with PSO and the second method is a classification with NB only. Both of these methods will be performed on all datasets of hepatitis, immunotherapy, echocardiogram, biomarkers, and chronic kidney as input and will produce an outcome that is the classification results on that dataset. In detail the classification process proposed in method 1 is as follows 1) Read the dataset, 2) Apply particle cluster optimization for selected features, 3) Remove features with low PSO values, 4) Apply NB classification on relevant features. Whereas the second method is not optimized by PSO.

D. Evaluation and Testing

In this study the variables used in both methods, PSO and NB are all the attributes that exist in each data set other than the target attribute. Data mining classification, PSO is used as a feature selection, which is used to select attributes that are considered dominant to determine the target attributes. But in this study, all attributes are considered as variables that together determine the classification results that produce accuracy values through the Rapidminer as a machine learning application.

IV. RESULTS AND DISCUSSION

A. Description of the dataset used

Data used in this trial are medical dataset of hepatitis, immunotherapy, echocardiogram, biomarkers, and chronic kidney taken from the UCI Machine Learning Repository [12]. Of the five datasets, four datasets have missing values so that the imputation process is needed, the four datasets are hepatitis, echocardiogram, biomarker, and chronic kidney while the other dataset, Immunotiency, has no missing value, which can also be used as an accurate comparison. The description of the dataset is presented in Table 4.

TABLE IV. ORIGINAL DATASET DESCRIPTIONS

Dataset Name	Num. of records	Num. of feature	Unbalanced ratio	Minority class	Majority class
Hepatitis	155	20	60 : 40	Yes	NO
Immuno-terapy	90	8	80 : 20	NO	YES
Echordio-gram	131	13	70 : 30	Alive	Dead
Biomarker	130	60	70 : 30	Yes	No
Chronic Kidney	382	25	60 : 40	NotCKD	CKD

Based on the data presented in Table 4, the number of records and features will not change due to the imputation method. Nevertheless, the imputation method in this study is also compared to the non-imputation method. The dataset without imputation will change the description of the dataset is presented in Table 5.

TABLE V. DATASET DESCRIPTIONS WITHOUT IMPUTATIONS

Dataset Name	Num. of records	Num. of feature	Unbalanced ratio	Minority class	Majority class
Hepatitis	81	20	60 : 40	Yes	NO
Immuno-terapy	90	8	80 : 20	NO	YES
Echordio-gram	63	13	70 : 30	Alive	Dead
Biomarker	30	60	70 : 30	Yes	No
Chronic Kidney	156	25	60 : 40	NotCKD	CKD

B. Results

In this research, all of the testing processes were carried out with Rapid Miner 5.3. Two classification methods i.e. NB + PSO and NB only were tested on three preprocessing models, namely without imputation (WI), with Zero imputation (ZI) and with Mean imputation (MI). Where the results of the calculation of the accuracy of the existing classification are presented in Table 6.

TABLE VI. ACCURACY OF CLASSIFICATION RESULTS

Dataset Names	Naïve Bayes (%)			Naïve Bayes + PSO (%)		
	WI	***ZI***	***MI***	***WI***	***ZI***	***MI***
Hepatitis	83.95	84.94	**85.00**	**95.06**	86.25	88,12
Immuno-terapy	77.78	77.78	77.78	87.78	87.78	87.78
Echordio-gram	87.30	**89.31**	83.21	90.08	**93.13**	90.08
Biomarker	93.33	**96.92**	**96.92**	93.33	**97.69**	**97.69**
Chronic Kidney	**100**	95	95	**100**	98	97

C. Discussion

Based on the experiments presented in Table 6, it appears that it appears that the use of the NB method optimized with PSO results in increased accuracy in all datasets, both methods with imputation and without imputation. But the use of optimization with PSO does not provide increased accuracy in the Biomarker dataset without the Mean imputation method. The zero imputation method can work well on two datasets, namely Echordiogram and Biomarker where the largest accuracy increase is 4.33%. While the Mean imputation method can also work well on the Biomarker dataset with the greatest accuracy improvement with the Zero imputation method, which is 4.33%. The mean imputation method can also improve accuracy in the Hepatitis dataset but only in the classification method using NB. Whereas in the Chronic Kidney dataset the best accuracy results even without using the imputation method. So it can be concluded that the imputation method can give effect to the accuracy results, even with simple imputation methods such as the Zero and Mean methods. Of the four datasets undergoing an imputation process, three datasets produce a trend of increasing accuracy with the imputation method tested.

V. Conclusions

Data mining requires the learning process of a dataset, to produce good learning the ideal dataset is needed. Data mining sometimes fails to produce abnormal data sets that are missing attribute values, such dataset cases are usually found in many medical health data sets as used in the experimental section above. In this study using medical datasets taken from the UCI repository, five datasets have been tested by applying the classification method, namely the Naeye Bayes (NB) method, and the Particle Swarm Optimization (PSO) method. There are two dataset pre-processing models, first, there is no imputation, where the empty attribute is deleted and second the data is imputed before the classification process is carried out. The test results based on the Naive Bayes method which was optimized by PSO proved to be better than the Naïve Bayes method without PSO. While the imputation method can tend to result in increased accuracy in three of the four datasets used. Based on this conclusion, in the next research, it is necessary to study and analyze more deeply the method of imputation on more datasets, it is also necessary to test other imputation methods which are considered to be more recent and adaptive

References

[1] J. D. A. Silva and E. R. Hruschka, "An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks," *Data Knowl. Eng.*, vol. 84, pp. 47–58, Mar. 2013.

[2] S. Zhang, "Shell-neighbor method and its application in missing data imputation," *Appl. Intell.*, vol. 35, no. 1, pp. 123–133, Aug. 2011.

[3] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 77–108, Jul. 2012.

[4] C. Beyan and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognit.*, vol. 48, no. 5, pp. 1653–1672, May 2015.

[5] W. Badr, "6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)," *Toward Data Science*, Jan-2019. [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. [Accessed: 24-May-2020].

[6] S. Chandrasekaran *et al.*, "Data Preprocessing: A New Algorithm for Univariate Imputation Designed Specifically for Industrial Needs," in *Workshop Computational Intelligence*, 2016.

[7] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification," *Knowledge-Based Syst.*, vol. 182, p. 104838, Oct. 2019.

[8] M. Mohamed, S. A. Sedory, and S. Singh, "Improved Mean Methods of Imputation," *Stat. Optim. Inf. Comput.*, vol. 6, no. 4, pp. 526–535, Nov. 2018.

[9] J. Li, L. Ding, and B. Li, "A novel naive bayes classification algorithm based on particle swarm optimization," *Open Autom. Control Syst. J.*, vol. 6, no. 1, pp. 747–753, 2014.

[10] T. GopalaKrishnan and P. Sengottuvelan, "A hybrid PSO with Naïve Bayes classifier for disengagement detection in online learning," *Progr. Electron. Libr. Inf. Syst.*, vol. 50, no. 2, pp. 215–224, Apr. 2016.

[11] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018.

[12] D. Dua and C. Graff, "UCI Machine Learning Repository," *University of California, School of Information and Computer Science*., 2019. [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: 24-May-2020].

[13] Z. Gong and H. Chen, "Model-based oversampling for imbalanced sequence classification," in *International Conference on Information and Knowledge Management, Proceedings*, 2016, vol. 24-28-October-2016, pp. 1009–1018.

[14] A. Moreo, A. Esuli, and F. Sebastiani, "Distributional random oversampling for imbalanced text classification," in *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 805–808.

[15] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, p. 799, Apr. 2020.

[16] M. Bramer, *Principles of Data Mining*, Third Edition. Springer, 2007.

[17] J. Kennedy, *Encyclopedia of Machine Learning*. Springer US, 2010.

Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization

ORIGINALITY REPORT

17%	8%	11%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	3%
2	GopalaKrishnan T, P Sengottuvelan. "A hybrid PSO with Naïve Bayes classifier for disengagement detection in online learning", Program, 2016 Publication	2%
3	Kristiawan Nugroho, Edy Noersasongko, Purwanto, Muljono, Ahmad Zainul Fanani, Affandy, Ruri Suko Basuki. "Improving Random Forest Method to Detect Hatespeech and Offensive Word", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Publication	1%
4	www.alliedacademies.org Internet Source	1%
5	Rianto Sitanggang, Tulus, Zakarias Situmorang.	

"THE ANALYSIS PERFORMANCE METHOD NAIVE BAYES ANDSSVM DETERMINE PATTERN GROUPS OF DISEASE", Journal of Physics: Conference Series, 2017 | 1%
Publication

6	Sfenrianto, Indah Purnamasari, Rizal Broer Bahaweres. "Naive Bayes classifier algorithm and Particle Swarm Optimization for classification of cross selling (Case study: PT TELKOM Jakarta)", 2016 4th International Conference on Cyber and IT Service Management, 2016 Publication	1%
7	Submitted to Kyungpook National University Student Paper	1%
8	Edi Jaya Kusuma, Oktaviana Rena Indriani, Christy Atika Sari, Eko Hari Rachmawanto, De Rosal Ignatius Moses Setiadi. "An imperceptible LSB image hiding on edge region using DES encryption", 2017 International Conference on Innovative and Creative Information Technology (ICITech), 2017 Publication	1%
9	Submitted to SASTRA University Student Paper	1%
10	Submitted to University of Lancaster Student Paper	<1%

11	Submitted to Technische Universiteit Delft Student Paper	<1%
12	Eko Hari Rachmawanto, Ignatius Moses Setiadi De Rosal, Christy Atika Sari, Heru Agus Santoso et al. "Block-Based Arnold Chaotic Map for Image Encryption", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Publication	<1%
13	Submitted to Academic Library Consortium Student Paper	<1%
14	arxiv.org Internet Source	<1%
15	Submitted to RDI Distance Learning Student Paper	<1%
16	Oktaviana Rena Indriani, Edi Jaya Kusuma, Christy Atika Sari, Eko Hari Rachmawanto, De Rosal Ignatius Moses Setiadi. "Tomatoes classification using K-NN based on GLCM and HSV color space", 2017 International Conference on Innovative and Creative Information Technology (ICITech), 2017 Publication	<1%
17	Zhang, S.. "Missing data imputation by utilizing information within incomplete instances", The Journal of Systems & Software, 201103	<1%

Publication

18	www.mdpi.com Internet Source	<1 %
19	ijmp.mums.ac.ir Internet Source	<1 %
20	Ying Sun, Fariborz Haghighat, Benjamin C.M. Fung. "A Review of the-State-of-the-Art in Data-driven Approaches for Building Energy Prediction", Energy and Buildings, 2020 Publication	<1 %
21	ijimai.org Internet Source	<1 %
22	scholar.uwindsor.ca Internet Source	<1 %
23	"Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making", Springer Science and Business Media LLC, 2020 Publication	<1 %
24	Eko Hari Rachmawanto, De Rosal Ignatius Moses Setiadi, Christy Atika Sari, Pulung Nurtantio Andono, Omar Farooq, Nungki Pradita. "Spread Embedding Technique in LSB Image Steganography based on Chaos Theory", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019	<1 %

Publication

25	lib.um.ac.id Internet Source		<1%
26	Submitted to Savitribai Phule Pune University Student Paper		<1%
27	Submitted to Sriwijaya University Student Paper		<1%
28	P. Amudha, H. Abdul Rauf. "Performance Analysis of Data Mining Approaches in Intrusion Detection", 2011 International Conference on Process Automation, Control and Computing, 2011 Publication		<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		