Timor-Leste Journal of Engineering and Science

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

Pest and Disease Identification with Incomplete Data

Edio da Costa
Department of Computer Science, School of Engineering and Science, Dili Institute of Technology, Timor-Leste
Email: ediocosta73@gmail.com

ABSTRACT

The problem of similarity of the symptoms causes a high degree of ambiguity in identifying pests and diseases, another problem in identifying pests and diseases is the incomplete symptoms (missing data) are told by the farmers because the symptoms conveyed have similarities with pests and other diseases making it difficult to identify. The objective of this study is to identify pests and diseases based on incomplete data. The similarity method with Jaccard Similarity (JS), Cosine Similarity (CS), and Dice Similarity (DS) is used to solve the problem of incomplete data. The purpose of the three methods is to find the best accuracy to solve the problem of incomplete data of symptoms to identify the pests and diseases of rice plants. The result of the experiment shows DS obtained the highest performance of accuracy compared to JS and CS.

Keywords: Missing data, ambiguity, similarity, pest and diseases.

1. Introduction

The research of the missing data problems has been carried out in the last 22 years, several methods have been applied to deal with missing data problems such as smoothing methods. In the modern era, the domain of this research has only been carried out by (Rubin, 1976) by building a theoretical framework to deal with the problem of missing data. Many techniques for missing data imputation have been suggested by (García-Laencina, 2015) Since 1980.

The problem of missing data occurs in a variety of domains, for several different reasons, and regardless of whatever they might be, has serious implications for knowledge extraction and classification performance (Santos et.al., 2019). When datasets are incomplete, pattern classification turns into a more complex task (Little et.al, 2015). Missing data diminishes the effectivity of statistical results, and may cause bias estimates, which in turn leads to unsound judgment (Capariño, 2019). Several classical approaches have been used to solve the missing data problem such as Synthetic generation of missing data (Howell, 2007); Data imputation using several strategies (García-Laencina et.al., 2010), and Evaluation of imputation algorithms (Santos et.al., 2017; García-Laencina et.al., 2015).

However, the problem of similarity of the symptoms causes a high degree of ambiguity in identifying pests and diseases, another problem in identifying pests and diseases is the incomplete symptoms (missing data) are told by the farmers because the symptoms conveyed have similarities with pests and other diseases making it difficult to identify. It requires the accuracy of analyzing the symptoms to get the right diagnosis. An effective way to handle the missing data problem is the classification approach. The classification method to solve the missing data problem is machine learning approach (Leng et.al., 2009) because it is often used to solve the problem of incomplete data (missing data) that told by the farmers. But as a comparison to solve the problem of incomplete data of symptoms in this study, we propose other methods namely Jaccard Similarity (JS), Cosine Similarity (CS), and Dice Similarity (DS). The purpose of the three methods is to find the best accuracy to solve the problem of incomplete data of symptoms to identify the pests and diseases of rice plants.

JS, CS, and DS have been applied in several fields to identified pest and diseases such as, identification of pests and plant diseases (Francis, 2016; Faria et.al, 2014), and biometrics (Sabab el. Al. 2016; Kurniawan 2014; Kaewthai, 2015). Some research was also done by (Gupta and Tiwari, 2016) using the CS method to diagnose a disease based on the similarity between symptoms and a pattern, as well as missing data. In several problem of missing data, the combination of similarity approach has proved to be effective compared to the single method. Similarity approach with JS, CS, and DS has provided the best performance.

Although there are many approaches to solve the missing data of pest and diseases, however, these problems often arise because unpredictable weather changes in recent years, human errors or system faults for collecting information. Therefore, the aim of this study is to identify pests and diseases based on missing data or incomplete data based on the dataset to reduce error rates and improve identification accuracy. The recommendation result is based on the False Identification Rate. The contribution of this research suggesting some recommendations based on the similarity approach testing with the value of the False Identification rate combination with the JS, CS, and DS.

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

2. Related Work

Missing data is the loss of information or data in a subject. Sometimes the missing data are caused by the research, such as, when data collection is done improperly or mistakes (Hand et.al., 2008). Many things cause missing data, which can be caused by incorrect input, related to the response of the respondent or there are obstacles in the data collection tool. Missing Data is a common obstacle researchers face in real-world contexts (Santos et.al, 2019).

Several studies to handle the problem of missing data have been carried out, such as missing data in the context of industrial data analytics to handling missing data in the industrial database (Ehrlinger et.al, 2018), approaches relational databases to handle the problem of missing data (Ezzine and Benhlima, 2018), missing data for TCM medical data in Data mining context (Zeng et.al, 2017). However, knowledge discovery is hindered because real data is often incomplete and noisy (Sessa and Syed, 2016).

Overall the type of missing data consist of Missing Completely at Random (MCAR) which means that missing data occur randomly from a complete sample (Polit and Beck, 2012; Deng, 2012); Missing not at Random (MNAR) which means that the probability of a missing observation is not related to the results of other observations (Polit and Beck, 2012; Missing at Random (MAR) which means that the probability of observation of missing data is usually related to information obtained with tools (Roderick et.al., 2002). This research focuses on the Missing at Random (MAR).

There are several methods used to measure distance similarities, such as CS, JS, DS, Hamming, and Minkowsky. The similarity testing has been implemented in some cases, such as Jaccard Similarity and TF-IDF for string comparison, Hamming Distance, and Relative Distance numeric value (Christen, 2018; Bilenko, 2018). Cosine similarity is a method used to measure the similarity between two objects, if the value of cosine is 0 then there is no similarity, while the similarity is 1 there is a high similarity. The research of (Chahal, 2016) to measure the performance of similarity between the data. The coefficient of the performing model in terms of the similarity of the model have the same model of recall and precision. However, coefficient CS is better compared to the coefficient of JS and DS in the complexity of algorithm computation. The comparison of the three coefficients of the three methods has been done by (Thada, 2013), the result showed the best performance of the three methods.

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them (Soyusiawaty and Zakaria, 2018). Let d = (d1, d2,…,dn), q = (q1, q2,…,qn) be two n-dimensional integer vectors. The formula for calculating similarity based on the vector similarity space measure are as follow:

$$Cosine\ Similarity(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij}.q_j}{\sqrt{\sum_{j=1}^{t} d_{ij^2}.\sum_{j=1}^{t} d_{qj^2}.}}$$

Where (D_i, Q) is a component of the vector $D_i\ and\ Q$, while d_{ij} is weight in the document. The Dice Similarity is a measurement between the number of elements in two samples.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

Where $X\ and\ Y$ are the numbers of elements in the two-samples. *Jaccard Similarity (JS)* is the one method of developing the Jaccard Coefficient to calculate the similarity of two continuous attribute vectors or count attributes with the following equation (Samatova, 2015):

$$T(p, q) = \frac{p.q}{||p||^2 + ||q||^2 - p.q}$$

Where $p.q$ are vector dot product, $||p||^2 + ||q||^2$ are the length of the vectors p and q.

3. Research Method

The database reference of pests and diseases in this study obtained from the research done by (Costa, Tjandrasa, and Djanali, 2018). The dataset consists of 60 data and 179 symptoms of pests and diseases. Table 1 showed the description of a rice pest and disease dataset based on the morphology of rice plants that consists of 9 parts. Each part labeling based on the symptoms of each morphology that were attacked by pests and diseases. There are 179 symptom attributes obtained from the field observations (Costa, Tjandrasa, and Djanali, 2018).

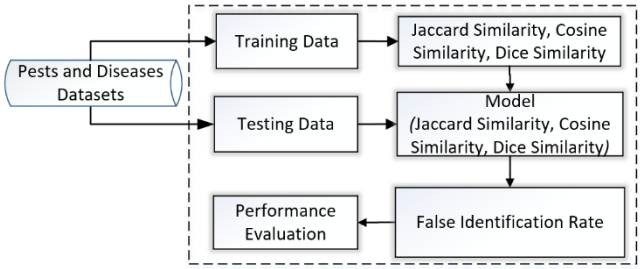

Figure 1. Blok Diagram of the Research

Figure 1. describes all the processes of our proposed study that consists of similarity model construction. In this process, all methods are used to build the similarity model based on

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

the three methods, such as JS, CS, and DC. Next, using the False Identification Rate (FIR) testing to measure the level of missing identification based on the missing data or incomplete data.

Table 1. Dataset Description

No	Attribute	Attribute Type	Number of Attribute
1	Leaf	Rotten leaves, dried leaves, roll leaf, thin leaves, oval leaf, brown leaf, leaf spot, striped leaf, scratched leaf, leaf spots, green leaf, broken leaf, leaves droop, yellow leaf, long leaf, narrow leaf, orange leaf, serrated leaves, uneven leaves, cracked leaves, floating leaf, dwarf leaf, ringed leaf, leaf hole, rhombus leaves, there are fig leaves, roll leaf, leaf regs, stiff leaves, pale leaf, short leaf, straight leaf, uneven leaves, broken leaves, dark leaf, twisted leaves, leaf edge is not balanced, white leaves, reddish leaves, small leaves, young leaves turn yellow, yellowed old leaves, burnt old leaves, wilted leaves, brown old leaves, chlorotic leaves, daun kebiruan, daun klorosis, tulang daun bercak coklat, jumlah daun besar, short old leaves, straight leaves, small leaves, dirty leaves, few leaves, yellow stripes, large leaf counts, long large leaves, dark brown leaves, light green leaves, short old leaves, straight leaves, green leaves, small leaves, dirty leaves, declining leaves, dead old leaves, purple leaves, yellow old leaves, brown old leaves, dry leaf tips, small leaves	a_1-a_{89}
2	Stem	short stems, dry stems, yellow stems, stab marks, floating stems, rot stems, hollow stems, weak stems, dwarf stems, burned stems, black stems, brown stems, thin stems, thin stems, broken stems	b_1-b_{16}
3	Root	root rot, black root, small root, young black root, dead root, long root, slight root, rough root, brown root	c_1-c_9
4	Seed	brown seeds, spotted seeds, young brown caterpillar seeds, mottled seeds, black seeds, hollow seeds, fig seeds, spore-filled seeds, thin seeds, clay seeds, scattered seeds, empty seeds, seed drops	d_1-d_{13}
5	Panicles	brown panicles, enlarged panicles, small panicles, broken panicles, spore panicles, green panicles, empty panicles, black panicle, orange panicle, rotten panicle, red panicle, short panicle, panicle filled, panicle broken, incomplete panicle, panicle drop	f_1-f_{16}
6	Shoots	yellow shoots, shoots withered, dry shoots, shoots easily pulled, reddish shoots, brown shoots	g_1-g_6
7	Grain	less grain, un filled grain, empty grain, empty grain	h_1-h_4
8	Midrib	rotten midrib, striped midrib, brown midrib	i_1-i_3
9	Tillers	tillers diminish, tillers, small tillers, tillers, late tillers	j_1-j_5

In this process, FIR is used to measure the success rate of the model built based on the proposed method. The result was obtained based on the percentage of errors in the testing process. Dataset is used in the testing process from data training (in-set testing).

The performance evaluation in this study uses the confusion matrix. The confusion matrix for a multi-class classification problem is a generalization of the binary case. Table 2 is an example of a multi-class confusion matrix (Makhtar et.al., 2011). For column X, the intersection with the first row is the True Positive (TP) values for class X. The sum of the value for the remaining cells of the column is the False Negative (FN) value for class X. The true positives for the second and third columns are the diagonal values of the confusion matrix.

Table 2. Confusion Matrix for Multi-Class

	Class X	Class Y	Class Z
Class X	$TP_{X(1,1)}$	$e_{XY(1,2)}$	$e_{XZ(1,3)}$
Class Y	$e_{YX(2,1)}$	$TP_{Y(2,2)}$	$e_{YZ(2,3)}$
Class Z	$e_{zx(3,1)}$	$e_{CB(3,2)}$	$TP_{Z(3,3)}$

The accuracy is evaluated based on the correct percentage of classification of the total amount of data. The performance evaluate uses the following formula:

$$Accuracy = \frac{all\ tp}{all\ data\ (n)}$$

where true positive (TP) is the number of testing data from a class that is correctly identified. False-positive (FP) is the number of testing data that incorrectly identified as from a class but actually from other classes.

Here are some cases of symptoms that have been identified as follows: brown leaf=1, leaf spot=1, floating leaf=0.25, and rhombus leaf=0.25. Diseases that are infected with these symptoms are Blast and Brown leaf spot. Then the calculation is as follows::

a. Jaccard Similarity

1) Blast

Looking for $p.q$ ($p\ dot\ q$)

$$p.q = 1 + 0.25 + 0.0625 + 0.0625 = 1.375$$

$$||p||^2 = 1^2 + 0.5^2 + 0.25^2 + 0.25^2 = 1.375$$

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

$$||q||^2 = 1^2 + 0.5^2 + 0.25^2 + 0.25^2 = 1.375$$
$$||q||^2 + ||q||^2 = 1.375 + 1.375 = 2.75$$
$$= 1.375 + 1.375 = 2.75$$
$$T(p,q) = \frac{1.375}{2.75-1.375} = 1$$

2) Brown leaf spot

With the same steps, the results obtained by brown leaf spot=0.428. Both of the calculations concluded that the highest similarity value was Blast.

b. Cosine Similarity

1). Blast

$$= \frac{(1x0.5) + (1x0.5) + (0.25x0.25) + (0.25x0.00)}{\sqrt{1^2 + 1^2 + 0.25^2 + 0.25^2} X \sqrt{1^2 + 1^2 + 0.25^2 + 0.25^2 + 1^2 + 0.25^2}}$$

$$= \frac{(0.5) + (0.5) + (0.0625) + (0)}{\sqrt{1 + 1 + 0.0625 + 0.0625} X \sqrt{1 + 1 + 0.0625 + 0.0625 + 1 + 0.0625}}$$

$$= \frac{1.0625}{\sqrt{2.125} X \sqrt{3.1875}} = 0.31$$

2). Brown leaf spot

With the same steps, the results obtained by brown leaf spot=0.28. Both of the calculation concluded that the highest similarities value was Blast.

c. Dice Similarity

1) Blast

$$= \frac{2|(1 + 1 + 0.25 + 0.25 + 1 + 1)x(0.5 + 1 + 1 + 0.25 + 1 + 0.25)|}{|1 + 1 + 0.25 + 0.25 + 0 + 0| + |0.5 + 1 + 1 + 0.25 + 1 + 0.25|}$$

$$= \frac{2|(2.50)x(4)|}{|2.50| + |4|}$$

$$= \frac{|20|}{|6.50|} = 3.20$$

2) Brown leaf spot

With the same steps, the results obtained by Brown leaf spot=2.33. Both of the calculation concluded that the highest similarities value was Blast.

4. Results and Analysis

For each data with the number of symptoms tested, the percentage of errors will be seen. Overall, the result of the experiment using error identification with False Identification Rate (FIR) for JS, DS, and CS are shown in Table 3. With 2 symptoms from 10 data of pests and diseases shows the result of FIR with JS shows the 6 data missing identified, while the resulting FIR for DS and CS are 7 data missing identified. The next, with 3 symptoms show the result of FIR with JS and CS shows the lower missing identification is 3 data, while DS is 4 data. The result identification with 4 symptoms, shows the resulting FIR with JS and CS still have the smallest missing level is 1 data. While the resulting FIR with 5 symptoms did not have a missing identification.

Table 4, shows the percentage of missing identification with JS DS and CS. the result shows the method of JE and CS obtains the highest performance compared to the DS.

Figure 2 shows the comparison level of the missing identification based on the number of symptoms of pests and diseases with JS, DS, and CS there are several things that to be analyzed. Firstly, based on the result testing with 2 symptoms showed the missing identification of JS lowest is 60% compared to CS and DS. The result shows that the error rate is still high. With the 3 symptoms, the missing identification with DS is relatively high is 40% compared to the JS and CS. While with the 5 symptoms the result identification showed the method of JS and CS has not error identification. Secondly, the result testing with JS, CS, and DS shows that the minimum symptoms for identification are three with an error rate is 30% because the result shows the error rate decreases significantly. Finally, overall the average error rate obtained by the JS method is smaller than DS and CS. This can be seen in the results of the testing (Figure 2), showing that from the number of symptoms testing, the error rate obtained was smaller at 60%. So it can be concluded that the result testing missing identification of the JS, DS, and CS showed the minimal symptoms for the pest and disease identification with 3 symptoms.

5. Evaluation Performa

Based on the result of testing showed that the JS method selected to test the accuracy of missing identification. Results of previous the system evaluation shows that JS showed the best FIR results. The rule of performance evaluation is to count the True Negative (TN) and False Positive (FP) to obtain the curve of Receiver Operating Characteristic (ROC). The ROC curve graphically represents the trade-off between TN and TP (Tom, 2003). The area under the curve measures the accuracy of the identification of pests and diseases based on the missing data. Figure 3 showed the better performance of JS, the higher values of precision and recall indicated that it was able to find the True Positive (TP) efficiently. With the 3 symptoms, the result shows that the performance of recall is higher compared to the precision and F-measure. However, the 4 symptoms showed the performance of F-measure is the highest compared to the precision and recall.

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

Table 3. Rezult Experiment of FIR With JS (Left), DS (Center), and CS (Right)

2 Symptoms		3 Symptoms		4 Symptoms		5 Symptoms	
Data	Result	Data	Result	Data	Result	Data	Result
1	0	1	0	1	0	1	0
2	1	2	0	2	0	2	0
3	1	3	0	3	0	3	0
4	0	4	0	4	0	4	0
5	0	5	0	5	0	5	0
6	1	6	0	6	0	6	0
7	1	7	1	7	1	7	0
8	1	8	1	8	0	8	0
9	1	9	1	9	0	9	0
10	0	10	0	10	0	10	0
FIR	60%	FIR	30%	FIR	10%	FIR	0%

2 Symptoms		3 Symptoms		4 Symptoms		5 Symptoms	
Data	Result	Data	Result	Data	Result	Data	Result
1	0	1	0	1	0	1	0
2	1	2	1	2	1	2	1
3	1	3	0	3	0	3	0
4	0	4	0	4	0	4	0
5	0	5	0	5	0	5	0
6	1	6	0	6	0	6	0
7	1	7	1	7	1	7	0
8	1	8	1	8	0	8	0
9	1	9	1	9	0	9	0
10	1	10	0	10	0	10	0
FIR	70%	FIR	40%	FIR	20%	FIR	10%

2 Symptoms		3 Symptoms		4 Symptoms		5 Symptoms	
Data	Result	Data	Result	Data	Result	Data	Result
1	0	1	0	1	0	1	0
2	1	2	0	2	0	2	0
3	1	3	0	3	0	3	0
4	0	4	0	4	0	4	0
5	0	5	0	5	0	5	0
6	1	6	0	6	0	6	0
7	1	7	1	7	1	7	0
8	1	8	1	8	0	8	0
9	1	9	1	9	0	9	0
10	1	10	0	10	0	10	0
FIR	70%	FIR	30%	FIR	10%	FIR	0%

**If 0=True and 1=False*

Table 4. Percentage of Missing Identification

Methods	Number of Data Testing	Missing Identification (%)			
		2 Symptoms	3 Symptoms	4 Symptoms	5 Symptoms
JS	10	60%	30%	10%	0%
CS	10	70%	30%	10%	0%
DS	10	70%	40%	20%	10%

6. Discussion

This study proposes a two-phase trial. The first trial uses complete data with a sample size of 300 datasets. The dataset has the highest similarity of symptoms. The objective of this testing is to identify pests and diseases based on missing data or incomplete data based on the dataset to reduce error rates and improve identification accuracy. The results showed that the problem of missing data of the pest and diseases caused by the decrease of accuracy identification, some research was also done by (Zieba, 2014). Another problem is also caused by the level of symptom similarity between diseases (Costa, Tjandrasa, and Djanali, 2018). The results showed that when the symptoms are told by the farmers or input by the user are missing or incomplete the system cannot identify pests and diseases correctly. Then the result testing of FIR and ROC in this research showed, to handle the problem of missing data of pests and diseases on the identification process, the minimum number of symptoms that must be inputted is 3. The result also shows some factors that caused the decrease of inaccuracy is the number of datasets and symptoms, the same problem was found by research conducted by (Capariño and Sison, 2019). Another problem is caused by the many database applications e.g., in data integration, data cleaning, or data exchange (Song et.al., 2018).

The second trial uses 500 datasets with the highest similarity of the symptoms. The objective of this testing is to identify the rate of ambiguity of our proposed method to test the missing identification based on the similarity of the symptoms. The result showed the similarity of the symptoms in the large dataset caused the highest of the missing identification of the pest and diseases, some research are also done by (Costa, Tjandrasa, and Djanali, 2018). To solve the problem the research provides an interactive system, but our testing showed that the interactive system is not enough to solve the similarity case using a large number of datasets with a high level of similarity. This is also proven by research conducted by (Chen, et.al., 2009). The result of testing also shows that the high level of similarity causes the long interactive process that makes the ambiguity system to carry out the identification process. So, to solve this problem the results of our study propose a similarity approach with the JS method, and the minimum symptoms inputted by the user is 3. These results are proven in FIR testing and performance evaluation with the recall of 84%, precision 81%, and F-measure is 78%. While the best performance obtained is 95% recall, precision 97%, and F-measure 92% (Figure 3).

Vol.1, Issue.1, pp.28-35, 2020
Available online at: http://tljes.org/index.php/tljes/data

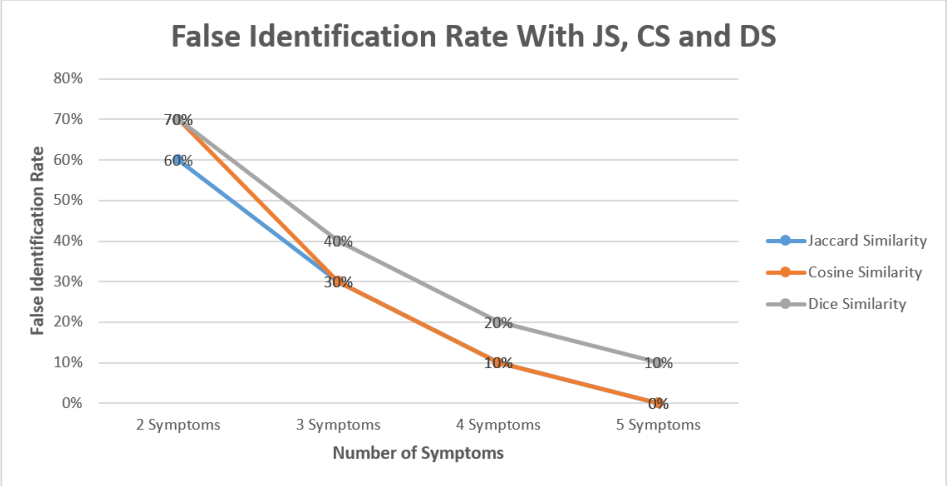

Figure 2. The Result Performance Comparison of FIR

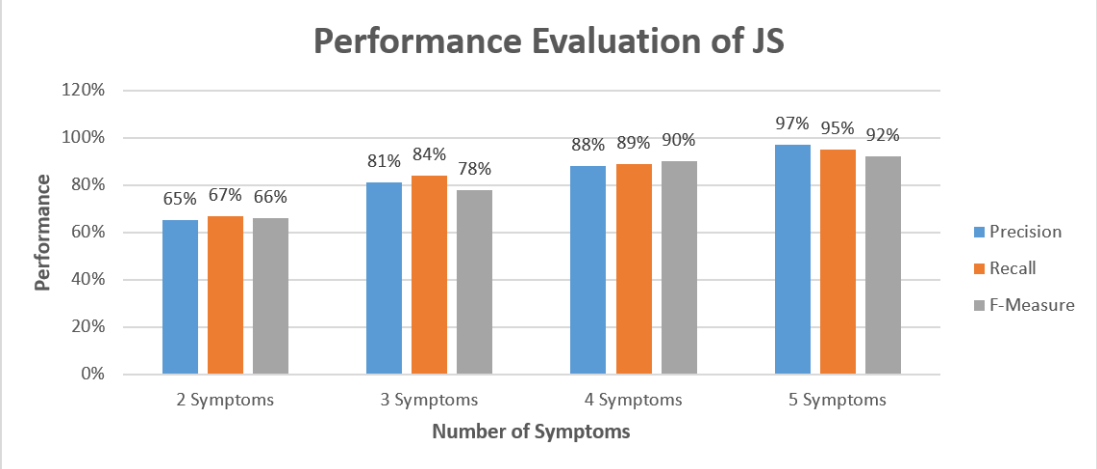

Figure 3. Performance Evaluation

7. Conclusion and Recommendation

The similarity approach with JS, CS, and DS have shown the varying results. With the FIR testing shown the average error rate obtained by the JS method is smaller than DS and CS. This can be seen in the results of the test show that from the number of symptoms, the error rate obtained was smaller at 60%. The result also showed the minimum symptoms that are use to identify pest and diseases are 3. The results testing of FIR recommended JS as the method that has the smallest misidentification rate. The area under the curve measures the accuracy of the identification of pests and diseases based on the missing data. So that the ROC test results showed the highest accuracy obtained by the JS for precision, recall and F-measure are 97%, 95% and 92%. Several problems that caused decrease the accuracy, such as the amount of data, the similarity of instances in the dataset, class labelling, and the like.

For future work, a combination of hybrid similarity method to reduce the error rate and increase the accuracy. It is also to improve the time of execution that as considered of the metrics performance.

References

Santos M. S., Pereira R. C., Costa A. F., Soares J. P., Santos J. and Abreu P. H.. (2019) Generating Synthetic Missing Data: A Review by Missing Mechanism," in IEEE Access, vol. 7, pp. 11651-11667, doi: 10.1109/ACCESS.201 9.2891360.

Capariño E. T., Sison A. M. and Medina R. P.. (2019), Application of the Modified Imputation Method to Missing Data to Increase Classification Performance. IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, pp. 134-139, doi: 10.1109/CCOMS.2019.8821632.

Sessa J. and Syed D. (2016). Techniques to deal with missing data," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, pp. 1-4, doi:10.1109/ICEDSA.2016.78184

Gupta P. and Tiwari P. (2016) "Measures of cosine similarity intended for fuzzy sets, intuitionistic and interval-valued intuitionistic fuzzy sets with application in medical diagnoses. International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 1846-1849.

Little T. D., Jorgensen T. D., Lang K. M. and Moore E. W. G. (2013). On the joys of missing data", *J. Pediatric Psychol.*, vol. 39, pp. 151-162.

Ezzine I. and Benhlima L. (2018) "A Study of Handling Missing Data Methods for Big Data.IEEE 5th International Congress on Information Science and Technology (CiSt), Marrakech, pp. 498-501, doi: 10.1109/CIST.2018.8596389.

Zeng D., Xie D., Liu R. and Li X. (2017). Missing value imputation methods for TCM medical data and its effect in the classifier accuracy. IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian, pp. 1-4, doi: 10.1109/HealthCom 2017.8210844.

Ehrlinger L., Grubinger T., Varga B., Pichler M., Natschläger T. and Zeindl J. (2019) Treating Missing Data in Industrial Data Analytics. *Thirteenth International Conference on Digital Information Management (ICDIM)*, Berlin, Germany, pp. 148-155, doi: 10.1109/ICDIM.2018.8846984.

Song S., Sun Y., Zhang A., Chen L. and Wang J., (2018) "Enriching Data Imputation under Similarity Rule Constraints," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 2, pp. 275-287, 1 Feb. 2020, doi: 10.1109/TKDE.2018.2883103.

Costa E., Tjandrasa H., Djanali S., (2018) Text mining for pest and disease identification on rice farming with interactive text messaging, International Journal of Electrical and Computer Engineering Vol.8 (3), pp.1671-1683.

Chahal M.. (2016). Information Retrieval Using Dice Similarity Coefficent. Internationl Journal of Advanced Research of Computer Science and Software Engineering. Vol. 6.

Tada V.. (2013). Comparision of Jaccard, Dice, Cosine Similarity Coefficient to Find the Best Fitness Value for Web Retrieve Documents Using Genetic Algorithm. International Journal of Inovation in Engineering and Technology. Vol. 2. pp. 202-205.

Francis, A.S. Dhas B.K., Anoop. (2016). Identification of Leaf Diseases in Pepper Plants Using Soft Computing Techniques." *International* Conference on Emerging Devices and Smart Systems (ICEDSS), pp.168-173.

García-Laencina P. J., P. Abreu H., Abreu M. H., and Afonoso N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,'' Comput. Biol. Med., vol. 59, pp. 125–133, Apr.

Howell D. C. (2007). The treatment of missing data,'' in The Sage Handbook of Social Science Methodology. London, U.K.: Sage, pp. 208–224.

García-Laencina P. J., Sancho-Gómez J.-L., and Figueiras-Vidal A. R.. (2010). ''Pattern classification with missing data: A review,'' Neural Comput. Appl., vol. 19, no. 2, pp. 263–282.

Santos M. S., Soares J. P., Abreu P. H., Araújo H., and Santos J. (2017). Influence of data distribution in missing data imputation,'' in Proc. Conf. Artif. Intell. Med. Eur. Vienna, Austria: Springer, pp. 285–294.

Deng (2016). "On Biostatistics and Clinical Trials". Archived from the original on 15 March 2016. *Retrieved 13 May 2016*.

Roderick J. A., Rubin, Donald B. (2002), Statistical Analysis with Missing Data (2nd ed.), *Wiley*

Makhtar M., Neagu D.C., Ridley M.J. (2011) Comparing Multi-class Classifiers: On the Similarity of Confusion Matrices for Predictive Toxicology Applications. In: Yin H., Wang W., Rayward-Smith V. (eds) Intelligent Data Engineering and Automated Learning - IDEAL 2011. IDEAL 2011. Lecture Notes in Computer Science, vol 6936. Springer, Berlin, Heidelb.

Faria A. (2014) Automatic identification of fruit flies (Diptera: Tephritidae). J. Vis. Commun. Image R, vol.25, pp.1516-1527.

Sabab A., Pritom A.I. (2016). Cardiovascular Disease Prognosis Using Effective Classification and Feature Selection Technique. *International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*. pp.1-6.

Soyusiawaty D. and Zakaria Y. "Book Data Content Similarity Detector With Cosine Similarity (Case study on digilib.uad.ac.id). (2018). 12th International Conference on Telecommunication Systems, Services, and Applications

(TSSA), Yogyakarta, Indonesia, pp. 1-6, doi: 10.1109/TSSA.2018.8708758.

Kurniawan N., Yanti M. Z. A., Nazri, and Zulvandri. (2014). Expert Systems for Self-Diagnosing of Eye Diseases Using Naïve Bayes. *International Conference of Advanced Informatics: Concept, Theory and Application,* pp. 113-116

Kaewthai S., Kiattisin S. (2015). Diabetes Dose Titration Identification Model." *Biomedical Engineering International Conference*. pp.1-5.

Polit D.F., Beck C.T.. (2012). Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed. Philadelphia, USA: Wolters Klower Health, Lippincott Williams & Wilkins

Lan et al. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, pp. 721-735.
1

Saltn C., Buckley, (1998). Term weighting approaches in automatic text retrieval." Information Processing and Management, vol. 24, pp. 513-523.

Chrsten. (2006). A comparison of personal name matching: techniques and practical issues, in: Workshops Proceedings of the 6th IEEE International Con- ference on Data Mining. pp. 290–294.

Bilenko, R. J. Mooney, W.W. Cohen, P.D. Ravikumar, S.E. Fienberg. (2003). Adaptive name matching in information integration, IEEE Intell. Syst. 18 (5), 16–23 .
F. Samatova, W. Hendrix, Jenkins, K. Padmanabhan, A Chakraborty. (2015). Graph-based Proximity Measures. Department of Computer Science North Carolina State University.

Zięba M., (2014) "Service-Oriented Medical System for Supporting Decisions With Missing and Imbalanced Data," in IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 5, pp. 1533-1540, Sept. 2014, doi: 10.1109/JBHI.2014.2322281.