Improved Sampling Data Workflow Using Smtmk To Increase The Classification Accuracy Of Imbalanced Dataset

Muhammad Syafiq Alza bin Alias[1], Norazlin Binti Ibrahim[2], Zalhan Bin Mohd Zin[3]

[1, 2, 3] Industrial Automation Section, UniKL Malaysia France Institute, Bangi, Malaysia

Email [1]syafiq.alias@s.unikl.edu.my, [2]norazlin@unikl.edu.my, [3]zalhan@unikl.edu.my&

Abstract:* One of the main challenges in machine learning classification is handling imbalanced data because imbalanced data can produce result bias towards the majority class and a poor performance of classification. Therefore, in this paper, an improved workflow is introduced to cater this issue. After combination of Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links or known as SMTmk method is performed, additional step is required to further increase the performance of machine learning classification especially in Specificity field. The step is completed by reducing the number of majority class based on the ratio of minority class. Three machine learning algorithms is used to test the classification result which are Extreme Gradient Boosting, Random Forest and Logistic Regression. Result recorded in this research shows that the ratio of 7 to 1 is better than the established methods which are SMOTE and hybrid method of SMOTE and Tomek Links.

Keywords: SMOTE, Tomek Links; Imbalance Data; Machine Learning;

1. INTRODUCTION

The existing real-world data in many areas such as credit card fraud, medical diagnosis and others consist of imbalanced data. The imbalanced data occurs simply because of an uneven balance in the number of positive and negative cases, or binary class labels, in a dataset [1]. It is one of the main challenges in machine learning classification because the classification would be bias toward the majority class [2] and suffers a poor performance according to receiver operator characteristic curve (ROC) [3]. Imbalanced data can offer high accuracy of the testing result because the testing sample will simply be classified to the majority class which occupied a high percentage of the total population [4]. It is an alarming issue because the main interest in dataset is the minority class (e.g., the credit card fraudulence transaction in financial industry). Therefore, any mistakes in classifying the data or failure to detect the true negative even one of the data might problem for the case study.

There are two main approaches to handle imbalanced data and ensure the accuracy of classification is good which are data level rebalance (data sampling) and modified learning algorithm approach [5]. This research is focus on the data sampling technique. Synthetic Minority Over-sampling Technique (SMOTE) is quite popular among machine learning researcher to cater the imbalanced data issue [6] [7] [8]. It proven to be effective and achieves high accuracy in many domains such as land cover, credit card fraud detection, bio informatics and others [3]. However, SMOTE will result in overgeneralization because it produces synthetic instance with the same number. Hence, the boundary between classes is uncertain [9].

Therefore, this research is conducted to improve the sampling data workflow by introducing the ratio method after the SMOTE and Tomek Link resampling is conducted. This additional step works by reducing the majority class associated to the minority class with ratio of 2 to 1, 3 to 1, 5 to 1 and 7 to 1.

RELATED WORKS

Numerous researches have been carried out to handle imbalanced data because learning from imbalanced data gives bad performance [10] and harmful to the classifier [2]. Normally, machine learning algorithm proposed based on the assumption that the training data is balanced [3]. Common method in dealing with mentioned scenario is data sampling method which is performed during the pre-processing data. The easiest sampling method are random under sampling (RUS) and Random Over Sampling (ROS) [11]. RUS works by selecting data randomly and delete it from the majority class. On the other hand, Random Over Sampling (ROS) works by randomly duplicates the data of the minority class. However, both methods have drawbacks as such RUS can lead to loss of information due to deleting data while ROS poses overfitting of data because of same data being repeated [12].

SMOTE method can perform better that RUS and ROS. It works by adding synthetic examples based on k-nearest neighbors which adjacent for each data in the minority class [9] [13] instead of replicating the existing ones [14]. The oversampling method is introduced by Chawla et. Al. [15] to increase the minority class data so that it will be balanced with the majority class data.

Another trendy approach to overcome the issue of imbalanced data is by using hybrid sampling technique with combination of SMOTE and Tomek Links (SMTmk). A data cleansing method Tomek Links introduced by Ivan Tomek [16] in 1976 is used as one of the combinations because Tomek links are able to remove the overlap between the classes either by removing both samples forming a Tomek link or by only removing the majority class sample [17]. Although Tomek Links approach may not give the best result, but when the method is combined with SMOTE method, the accuracy produced is high as mentioned in [7]. Basically, the mechanism of SMTmk is new synthetic instances is created by increasing the minority class instances using SMOTE and majority class instances is decreased by using Tomek Links algorithm to eliminate some of the class instances [18].

There are six evaluation metrics that are used to evaluate the sampling data workflow which are Accuracy, Sensitivity, Specificity, Precision, Matthews Correlation Coefficient (MCC) and Breakpoint Cluster Region (BCR). The evaluation metrics are calculated based on confusion matrix provided in Figure 1 [19].

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 1: Confusion Matrix

The formula of each evaluation metric is expressed as the following [20]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (1)$$

$$\text{Sensitiviy} = \frac{\text{TP}}{\text{TP+FN}} \qquad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+ FP}} \qquad (4)$$

$$\text{MCC} = \frac{\text{TP X TN} - \text{FP X FN}}{\sqrt{((\text{TP+fn})(\text{TP+FP})(\text{TN+FN})(\text{TN+FP}))}} \qquad (5)$$

$$\text{BCR} = \text{Sensitivity X Specificity} \qquad (6)$$

I. METHODOLOGY

The overview of the general methodology to develop machine learning algorithm is visualized in Figure 2. Basically, it is comprised of 5 steps. In this research our focus is on handling the imbalanced data that falls under the Pre-processing step where process of preparing the data take place. All the simulations in conducted in python using the Jupyter Notebook.

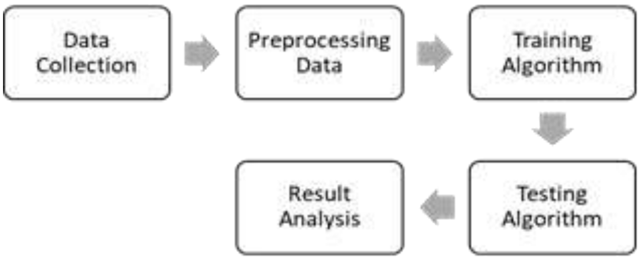
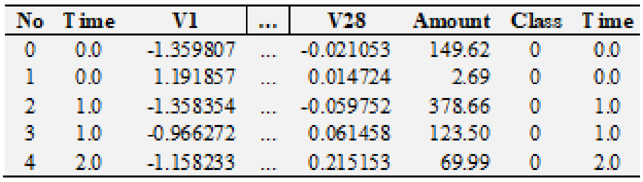


Figure 2: Overview of the general methodology to develop machine learning algorithm.

In data collection phase, a highly imbalanced data which was downloaded from Kaggle website [21] is used. The data composed of fraudulence and legitimate credit card transaction within 2 days that were made by European cardholders in September 2013. It contained 492 fraudulence transactions out of 284,807 transactions in total. The input variables contain only numerical value from the result of a PCA transformation due to confidentiality reason. The features are relabeled as Features V1, V2, V3 until V28 (Figure 3) except for two fields which are Time and Amount. In feature called 'Class', the output of the data whether the transaction is fraudulence or not is recorded.



No	Time	V1	...	V28	Amount	Class	Time
0	0.0	-1.359807	...	-0.021053	149.62	0	0.0
1	0.0	1.191857	...	0.014724	2.69	0	0.0
2	1.0	-1.358354	...	-0.059752	378.66	0	1.0
3	1.0	-0.966272	...	0.061458	123.50	0	1.0
4	2.0	-1.158233	...	0.215153	69.99	0	2.0

Figure 3: Sample data from Credit Card Transaction dataset.

The main highlight is the pre-processing phase. Any missing data will be replaced with certain number. However, this data is already been process with no missing data including transformation as mentions in previous paragraph. The proposed of imbalanced data techniques are applied in this step. In this study, a total of 7 ways to handle imbalanced data techniques are involved. First, the data is not changed. It is normally separated into training and testing data with ratio of 8 to 2. This first technique is called as 'Normal' in this research. For the next two simulations, the experiments are conducted by using well known imbalanced data techniques called Synthetic Minority Over-sampling Technique, SMOTE and the combination of SMOTE and Tomek links (SMTmk). Again, the dataset is separated into training and testing data with ratio of 8 to 2. Then, this experiment is continued with the proposed techniques to handle imbalanced data and increase the accuracy of fraud detection. This study proposed that after applying the SMTmk technique and the dataset is divided into training and testing data, the training data need to be further processed before the training of the machine learning algorithm is conducted. The added process is by decreasing the data in majority class based on ratio of detection class which is fraudulence class since this study aims to highlight the detection class. In this research the ratio used are 2 to 1 (called Prop2), 3 to 1 (called Prop3), 5 to 1 (called Prop5) and 7 to 1 (called Prop7). The fraudulence class will have more data compared to the legitimate class to have more bias on fraudulence class.

Then, three machine learning algorithms are used which are eXtreme Gradient Boosting (XGB), Random Forest (RF) and Logistic Regression (LR). The training algorithm and testing algorithm will be conducted using these 3 algorithms.

The last phase of this research is result analysis. Six evaluation metrics are used to evaluate the sampling data workflow which are Accuracy, Sensitivity, Specificity, Precision, Matthews Correlation Coefficient (MCC) and Breakpoint Cluster Region (BCR). All the result is recorded and presented in the following section.

II. RESULT AND DISCUSSION

The Credit Card transaction dataset used in this research contains of highly imbalanced data as shown in Figure 4 and Table 1. The fraudulence transaction data is only 0.17% from overall data and the proportion is about 578 to 1. This highly imbalanced data may lead to false indication of model accuracy.

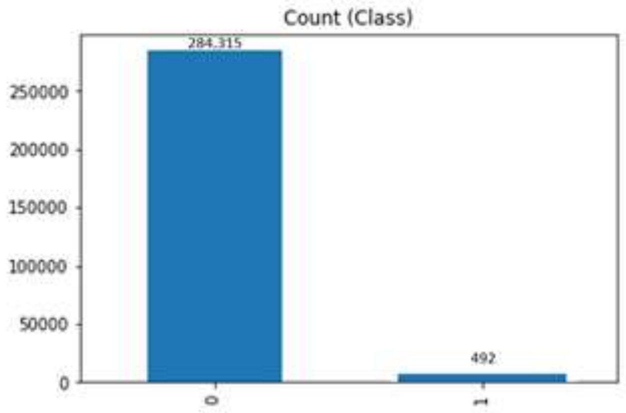


Figure 4: Graph of the highly imbalanced dataset between the classes of data.

Table 1: Overview of the class and number of transactions in the dataset.

Data	No of Transactions
Good	284315
Fraud	492
Total	284807

This experiment is conducted to handle imbalanced data. The first step in developing machine learning is to split the data into Training and Testing data. Table 2 shows the number of training data and the number of testing data is shown in Table 3.

Table 2: Overview of the class and number of transactions in the Training dataset.

No of Transactions	**Good**	**Fraud**	**Total**
Normal	227440	405	227845
SMOTE	227389	227515	454904
SMTmk	227389	227515	454904
Prop2	113932	227245	341177
Prop3	76116	227153	303269
Prop5	45775	227167	272942
Prop7	32842	227102	259944

Table 3: Overview of the class and number of transactions in the Testing dataset.

No of Transactions	**Good**	**Fraud**	**Total**
Normal	56875	87	56962
SMOTE	56926	56800	113726
SMTmk	56926	56800	113726
Prop2	28494	56801	85295
Prop3	19024	56794	75818
Prop5	11505	56731	68236
Prop7	8216	56771	64987

Based on Table 2 and Table 3, the Fraud data for Normal method is less that the Good data. This is because the separation will take the real cases from the dataset. However, for SMOTE and SMTmk, the Fraud data is balanced with the Good data. Although the Fraud data contained in the whole dataset is only 472, the number of Fraud cases in SMOTE and SMTmk is far more which is about 227515 for training and 56800 for testing. This is because both processes will synthetically add more data based on k-nearest neighbors to balance the number of Good and Fraud cases. On the other hand, all the proposed methods are having the same number of Fraud cases as SMOTE and SMTmk since the proposed method will undergo the SMOTE and SMTmk. The Good data is different because of the additional process that is proposed in this research. Each of the Good data will be reduced based on the current value of Fraud data. For example, in Prop2 the Good data will be divided by 2 from the Fraud data. Prop 3 will have 3 times less number of Good data than the Fraud cases followed by Prop5 divided by 5 and Prop7 divided by 7.

After the data is split into two, the data is then trained and tested with 3 machine learning algorithms which are XGB, RF and LR. The performance of the machine learning algorithm is analyzed by using the evaluation metric stated in the methodology which are Accuracy, Sensitivity, Specificity, Precision, Matthews Correlation Coefficient (MCC) and Breakpoint Cluster Region (BCR). The result for XGB, RF and LR algorithms is presented in Table 4, Table 5, and Table 6 respectively

Table 4: Result of the XGB algorithm.

Metrics	**Normal**	**SMOTE**	**SMTmk**	**Prop2**	**Prop3**	**Prop5**	**Prop7**
Accuracy	99.9561%	98.0356%	98.0304%	98.9155%	99.3827%	99.4651%	99.6045%
Sensitivity	99.9912%	99.0057%	98.9829%	98.5681%	98.4493%	97.5750%	97.2493%
Specificity	77.0115%	97.0634%	97.0757%	99.0898%	99.6954%	99.8484%	99.9454%
Precision	99.9648%	97.1255%	97.1366%	98.1925%	99.0848%	99.2397%	99.6135%
MCC	84.6335%	96.0892%	96.0780%	97.5653%	98.3554%	98.0843%	98.2005%
BCR	88.5014%	98.0346%	98.0293%	98.8290%	99.0724%	98.7117%	98.5973%

Table 5: Result of the RF algorithm.

Metrics	**Normal**	**SMOTE**	**SMTmk**	**Prop2**	**Prop3**	**Prop5**	**Prop7**
Accuracy	99.9508%	98.0989%	98.2563%	99.5029%	99.7850%	99.8315%	99.8492%
Sensitivity	99.9912%	99.8437%	99.8560%	99.5403%	99.4533%	99.0874%	98.8559%
Specificity	73.5632%	96.3504%	96.6532%	99.4842%	99.8961%	99.9824%	99.9930%
Precision	99.9596%	96.4811%	96.7640%	98.9775%	99.6891%	99.9124%	99.9508%
MCC	82.5801%	96.2563%	96.5620%	98.8852%	99.4278%	99.3981%	99.3160%
BCR	86.7772%	98.0970%	98.2546%	99.5122%	99.6747%	99.5349%	99.4244%

Table 6: Result of the LR algorithm.

Metrics	**Normal**	**SMOTE**	**SMTmk**	**Prop2**	**Prop3**	**Prop5**	**Prop7**
Accuracy	99.9157%	95.9728%	96.0387%	96.2108%	97.0456%	97.4705%	97.9319%
Sensitivity	99.9824%	98.2574%	98.2451%	96.6870%	96.1312%	93.8896%	92.3442%
Specificity	56.3218%	93.6831%	93.8275%	95.9719%	97.3518%	98.1968%	98.7406%
Precision	99.9332%	93.9720%	94.1009%	92.3319%	92.4010%	91.3488%	91.3876%
MCC	68.3546%	92.0411%	92.1667%	91.6387%	92.2774%	91.0891%	90.6805%
BCR	78.1521%	95.9702%	96.0363%	96.3295%	96.7415%	96.0432%	95.5424%

Overall, results show that aside from the Normal method, Prop7 method has the highest accuracy of detection followed by Prop5, Prop3 and Prop2. All the mentioned proposed method performed better than the conventional imbalanced data handler method. For XGB algorithm, Prop 7 has recorded the accuracy of 99.6045% which is better than SMOTE that has 98.0356% accuracy and SMTmk has 98.0304% of accuracy. Besides that, in RF algorithm also shows that the proposed method, Prop7 has better performance in terms of accuracy which is 99.8492% compared to SMOTE that has 98.0989% while SMTmk recorded 98.2563%. In the third algorithm that is LR, Prop7 has the accuracy of 97.9319% compared to the conventional methods which is SMOTE has accuracy of 95.9728% and SMTmk with accuracy of 96.0387%.

Although, the Normal method has recorded highest accuracy in all three of the algorithms which is 99.9561% for XGB, 99.9508% for RF and 99.9157 for LR, but the specificity of this method shows otherwise. The normal method is only good in sensitivity which mean the method is good in detecting a true positive class but performed poorly in detecting true negative class as recorded in every of their specificity fields. This is because the data is highly imbalanced toward the true positive class. All the specificity fields for Normal method is less than 80% while others data sampling method has higher than 90%. Therefore, the proposed methods are fit to be used because again Prop7 has the highest percentage of Specificity followed by Prop5, Prop3 and Prop2. Prop7 has the specificity of 99.9454% for XGB, 99.9930% for RF and 98.7406% for LR. The SMOTE perform lower than Prop7 which the specificity are 97.0634% for XGB, 96.3504% for RF and 93.6831% for LR. The improve version of SMOTE that is SMTmk has better performance thatn SMOTE, however it is still lower that Prop7 with value of specificity are 97.0757% for XGB, 96.6532% for RF and 93.8275% for LR.

All the proposed methods work better than the SMOTE and SMTmk methods. Besides that, based on the proposed method models, it can be said that the higher the ratio of targeted class (Fraudulence class) the higher the accuracy and specificity of detection. However, this research only covers up to 7 to 1 ratio of data because to avoid higher imbalanced data which can lead to false indication of accuracy.

2. CONCLUSION

Machine learning performed well in classification of data and has been widely used in many areas such as flood prediction, text classification, etc. However, real world data often recorded with highly imbalanced which result in poor performance in classification. In this research, an improved way to handle imbalanced data has been introduced. After SMTmk method is performed, additional step which is removing majority data based on ratio of minority is proposed to improve the Specificity of the result. Based on the simulation conducted on Credit Card transaction data, result shows that the proposed method, Prop7 produced the highest Specificity value compared to Normal, SMOTE and SMTmk methods with value of 99.9454% for XGB, 99.9930% for RF and 98.7406% for LR. Although, the Accuracy of detection is dominated by Normal method and proposed method is in the second place, however, the Specificity value recorded is very low which is less than 80% for all the 3 machine learning algorithms tested. Financial industry cannot afford to miss out the true negative detection and rather to have the false positive indication to detect the fraudulence transaction. Therefore, it can be said that the proposed method works well in handling imbalanced data with acceptable value of accuracy and high value of specificity. In this research, the proposed method, Prop7 has been proven to have a good detection on true negative and true positive values of the dataset which is better than the current conventional methods which are SMOTE and SMTmk.

ACKNOWLEDGEMENT

The authors would like to thank Ministry of Higher Education Malaysia and Universiti Kuala Lumpur (UniKL) for funding this research through research Grants (FRGS/1/2016/ICT02/UNIKL/02/2) and (UniKL STR15142).

REFERENCES

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

[1] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection using Random Forest with Class Imbalanced Big Data," IEEE Int. Conf. Inf. Reuse Integr. Data Sci. Medicare, pp. 80–87, 2018.

[2] C.-L. Liu and P.-Y. Hsieh, "Model-Based Synthetic Sampling for Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. PP, no. 1, pp. 1–15, 2019.

[3] M. Ahsan, R. Gomes, and A. Denton, "SMOTE Implementation on Phishing Data to Enhance Cybersecurity," IEEE Int. Conf. Electro Inf. Technol., vol. 2018-May, pp. 531–536, 2018.

[4] H. Y. Huang, Y. J. Lin, Y. S. Chen, and H. Y. Lu, "Imbalanced Data Classification using Random Subspace Method and SMOTE," 6th Int. Conf. Soft Comput. Intell. Syst. 13th Int. Symp. Adv. Intell. Syst. SCIS/ISIS 2012, pp. 817–820, 2012.

[5] F. Koto, "SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE: An Enhancement Strategy to Handle Imbalance in Data Level," Int. Conf. Adv. Comput. Sci. Inf. Syst., pp. 280–284, 2014.

[6] J. Li, H. Li, and J. L. Yu, "Application of Random-SMOTE on Imbalanced Data Mining," Proc. - 2011 4th Int. Conf. Bus. Intell. Financ. Eng. BIFE 2011, pp. 130–133, 2011.

[7] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective Prediction of Three Common Diseases by Combining SMOTE with Tomek Links Technique for Imbalanced Medical Data," Proc. 2016 IEEE Int. Conf. Online Anal. Comput. Sci. ICOACS 2016, vol. 2016, pp. 225–228, 2016.

[8] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, "An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set," ICEAST 2018 - 4th Int. Conf. Eng. Appl. Sci. Technol. Explor. Innov. Solut. Smart Soc., pp. 1–4, 2018.

[9] T. E. Tallo and A. Musdholifah, "The Implementation of Genetic Algorithm in SMOTE (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem," 2018 4th Int. Conf. Sci. Technol., vol. 1, pp. 1–4, 2018.

[10] M. Al Helal, M. S. Haydar, and S. A. M. Mostafa, "Algorithms Efficiency Measurement on Imbalanced Data using Geometric Mean and Cross Validation," IWCI 2016 - 2016 Int. Work. Comput. Intell., pp. 110–114, 2016.

[11] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "A Comparative Study of Data Sampling and Cost Sensitive Learning," Proc. - IEEE Int. Conf. Data Min. Work. ICDM Work. 2008, pp. 46–52, 2008.

[12] P. Sarakit, T. Theeramunkong, and C. Haruechaiyasak, "Improving Emotion Classification in Imbalanced YouTube Dataset using SMOTE Algorithm," ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., pp. 1–5, 2015.

[13] Y. Pristyanto, I. Pratama, and A. F. Nugraha, "Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification," 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, pp. 310–314, 2018.

[14] M. Habib, H. Faris, M. A. Hassonah, J. Alqatawna, A. F. Sheta, and A. M. Al-Zoubi, "Automatic Email Spam Detection using Genetic Programming with SMOTE," ITT 2018 - Inf. Technol. Trends Emerg. Technol. Artif. Intell., pp. 185–190, 2019.

[15] K. W. P. Chawla, N.V., Bowyer, K.W., Hall, L.O., "SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[16] I. Tomek, "Two Modifications of CNN," IEEE Trans. Syst. Man. Cybern., vol. SMC-6, no. 11, pp. 769–772, 1976.

[17] I. Dutta, "Data Mining Techniques to Identify Financial Restatements," 2018.

[18] Y. Sanguanmak and A. Hanskunatai, "DBSM: The Combination of DBSCAN and SMOTE for Imbalanced Data Classification," 2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2016, pp. 1–5, 2016.

[19] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," Int. Jt. Conf. Neural Networks, pp. 1322–1328, 2008.

[20] Anuranjeeta, K. K. Shukla, A. Tiwari, and S. Sharma, "Classification of Histopathological Images of Breast Cancerous and Non Cancerous Cells based on Morphological Features," Biomed. Pharmacol. J., vol. 10, no. 1, pp. 353–366, 2017.

[21] "Kaggle: Your Home for Data Science." [Online]. Available: https://www.kaggle.com/. [Accessed: 19-Jun-2019].