The Best Techniques to Deal with Unbalanced Sequential Text Data in Deep Learning

Sumarni Adi[1]*, Awaliyatul Hikmah[2], Bety Wulan Sari[3], Andi Sunyoto[4], Ainul Yaqin[5], Mardhiya Hayaty[6]
Department of Information Systems, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia[1, 3]
Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia[2, 5, 6]
Magister of Informatics Engineering, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia[4]

***Abstract*—Datasets with a balanced distribution of data are often difficult to find in real life. Although various methods have been developed and proven successful using shallow learning algorithms, handling unbalanced classes using a deep learning approach is still limited. Most of these studies only focus on image data using the Convolution Neural Network (CNN) architecture. In this study, we tried to apply several class handling techniques to three datasets of unbalanced text data. Both use a data-level approach with resampling techniques on word vectors and algorithm-level using Weighted Cross-Entropy Loss (WCEL) to handle cases of imbalanced text classification. With Bidirectional Long-Short Term Memory (BiLSTM) architecture. We tested each method using three datasets with different characteristics and levels of imbalance. Based on the experiments that have been carried out, each technique applied has a different performance on each dataset.**

***Keywords*—Imbalanced text classification; deep learning; resampling technique; weighted cross-entropy loss**

I. Introduction

Datasets with unbalanced class conditions are usual in real life, for example, in the case of fraud detection[1], cancer diagnosis[2], and spam detection[3]. These are challenges for machine learning models to perform classification tasks because samples are not the same in each class. As a result, the classifier may have high accuracy in the majority class. However, the classifier tends to ignore the minority class, so it has poor performance for detecting the minority class[4]. On the other hand, the minority class sometimes has a more important role because it has beneficial information, for example, when diagnosing cancer where patients are in the minority class. If the learning algorithm cannot detect the minority class properly, it can endanger someone's life.

Although various methods for dealing with class imbalance problems have been developed over the last two decades and have proven successful in various domains, most of them still focus on shallow learning algorithms[5]. Several researchers express similar things, among others: [6], [7], and [8], which states that the handling of unbalanced classes in deep learning has not been studied further. Based on a survey conducted by[5], more than 80% of research related to unbalanced classroom problems in deep learning still focuses on the field of computer vision using the Convolution Neural Network (CNN) architecture. Research by[9] as well, using Convolution Neural Network (CNN) to estimate the accuracy of the head pose angle based on deep learning in image recognition.

There are three approaches used to deal with unbalanced classes: data-level methods, algorithm-level methods, and hybrid methods[10]. Data-level methods are carried out by changing the data distribution in each class or resampling to achieve the desired condition. Resampling is done by reducing data from the majority class (under sampling) or adding data to the minority class (oversampling). Although it has been proven to be effective in overcoming the problem of data imbalance based on the survey conducted, on the other hand, under sampling has the potential to eliminate data that has important information. Oversampling can result in the learning model being overfitted so that the performance of the resulting model may not necessarily improve and increase computational effort [11]. The following method is algorithm-level, namely by making direct modifications to the learning algorithm to reduce bias in the majority class. Finally, the hybrid method is a combination of data-level and algorithm-level methods[8].

II. Literature Review

Handling unbalanced classes using a data-level approach is done by[6], [12], and[13]. Study[6] compared the Random Under Sampling (RUS), Random Oversampling (ROS), and Two-Phase Learning methods using multiclass image datasets trained with various CNN architectures. Overall, ROS has the best performance compared to the other two methods, RUS has poor results, and two-phase learning is considered less effective in dealing with class imbalance cases. The two-phase learning method was also proposed by[12] to classify WHOI-Plankton dataset images with a high level of imbalance. Different from[6], study[12] Instead, it concludes that this approach has been proven to be effective in improving the performance of the minority class while maintaining the performance of the majority class. Still using image dataset and CNN architecture, research[13] gets better performance by implementing the ROS method to handle unbalanced classes.

The algorithm-level approach is carried out by[14], [15], and[16]. Study[14] implemented a cost-sensitive CNN to classify various image datasets. This study also compares cost-sensitive with data-level approaches such as SMOTE and RUS. The proposed method is proven to have the most superior performance. Unfortunately, the performance metric used in this study is only accuracy, in which the evaluation method is not appropriate for measuring the performance of learning models with unbalanced classes[5]. Furthermore, research[15] used a cost-sensitive deep neural network (CSDNN) to predict hospital readmission. This study also compares the proposed

*Corresponding Author.

method with shallow learning algorithms such as Decision Tree and Support Vector Machine. The proposed method is proven to get better performance. CNN's cost-sensitive approach was also used in this study[16] for time-series classification with unbalanced data. The proposed method provides superior performance compared to the data-level approach.

The research that applies the hybrid method is[17] by combining SMOTE technique and weighted loss function to classify various image datasets with deep neural network architecture. The proposed method can improve the learning algorithm's performance, but on the other hand, the application of SMOTE also produces noisy data.

In this study, we intend to use the Bidirectional Long-Short Term Memory (BiLSTM) architecture to classify text data with various levels of imbalance. Several techniques to overcome class imbalances will be applied and observed how they affect the learning model's performance.

III. Research Method

A. Data Acquisition

We use three datasets in the form of labeled text with varying levels of imbalance. The level of imbalance is using the imbalance ratio (IR) formulated in Equation 1 by dividing the maximum class size against the minimum class[18].

$$IR(C)=\frac{\max_i C_i}{\min_j C_j} \quad (1)$$

Where:

$\max_i C_i$=maximum class size, IR(C)=imbalance ratio.

The first dataset we use is the Customer Support Tickets Dataset, which contains complaints about problems from users of an application obtained from the Google Play Store review column. The second dataset is the News Category Dataset, containing news headlines from 2012 to 2018 obtained from HuffPost[19]. Then the last dataset is the Drug Review Dataset which contains patient reviews related to drugs from the disease[20]. This study only used a few categories or classes from the three datasets with details, as shown in Table I.

TABLE I. Details of the Dataset

Datasets	Instances	Class	IR (C)
Cust. Support Tickets	5,150	3	4.99
News Category	50.879	3	15.72
Drug Review	57,463	5	8.52

B. Data Preprocessing

This stage includes the case folding process, data cleaning from special characters, normalization, to data representation. The cleaned text data will be converted into numeric digits by utilizing the Tokenizer library on Keras. Then the digits will be converted into word vectors using the word embedding technique with a dimension of 300. The word embedding technique applied is different based on the language contained in the dataset. The first dataset uses the fastText technique because the Indonesian language training model is available. In contrast, the second and third datasets use the GloVe technique because an English training model is available.

C. Unbalanced Class Handling

There are two approaches that we will use in this research, namely:

1) Data-Level: In simple terms, Fig. 1 shows the resampling technique in this study. Resampling of training data is 80% training and 20% testing. The resampled data is a word vector from the training dataset.

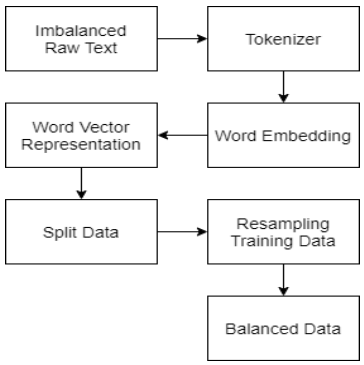


Fig. 1. Data Balancing Step.

We used RUS and Tomek Links (TL) for undersampling. RUS will delete some instances randomly in the majority class until the desired distribution is reached. In comparison, Tomek Links will delete a pair of closest neighbors but belong to a different class[21]. To perform oversampling, we used ROS and Synthetic Minority Oversampling Technique (SMOTE). ROS will randomly duplicate instances of the minority class, while SMOTE will generate synthetic data by linear interpolation between adjacent minority class samples[22]. The distribution of classes before and after resampling is shown in Fig. 2 for Customer Ticket Dataset, Fig. 3 for News Category Dataset, and Fig. 4 for the Drug Review Dataset.

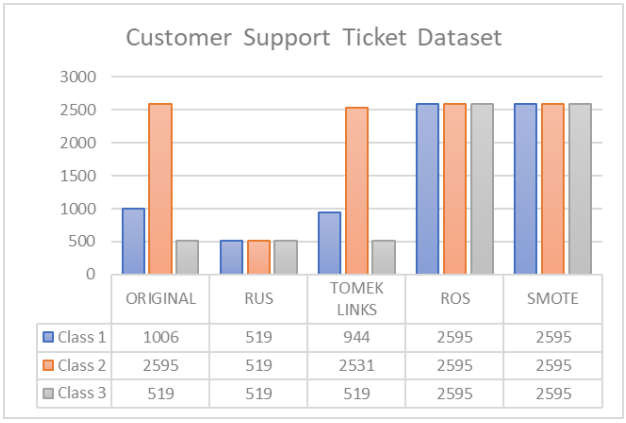
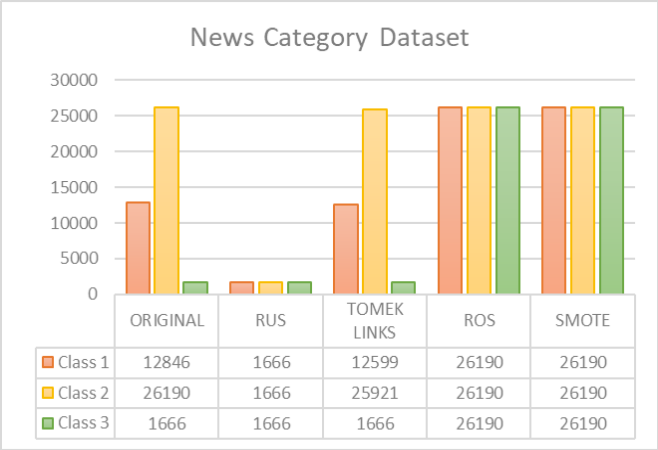
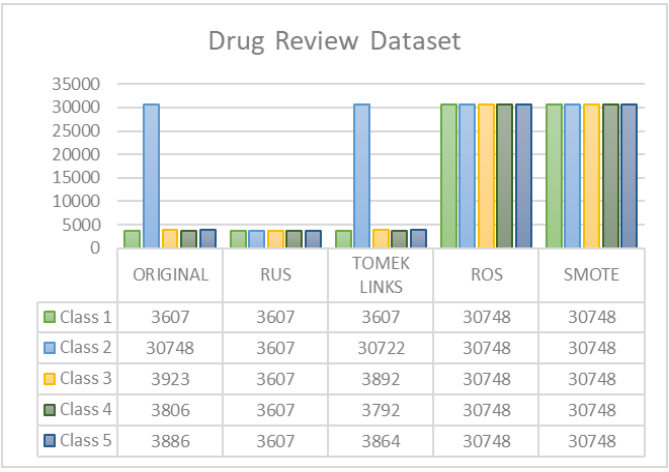


Fig. 2. Class Distribution before and after Resampling on the Customer Support Ticket Dataset.



	ORIGINAL	RUS	TOMEK LINKS	ROS	SMOTE
Class 1	12846	1666	12599	26190	26190
Class 2	26190	1666	25921	26190	26190
Class 3	1666	1666	1666	26190	26190

Fig. 3. Class Distribution before and after Resampling on the News Dataset.



	ORIGINAL	RUS	TOMEK LINKS	ROS	SMOTE
Class 1	3607	3607	3607	30748	30748
Class 2	30748	3607	30722	30748	30748
Class 3	3923	3607	3892	30748	30748
Class 4	3806	3607	3792	30748	30748
Class 5	3886	3607	3864	30748	30748

Fig. 4. Class Distribution before and after Resampling on the Drug Review Dataset.

2) Algorithm-Level: In the algorithm-level approach, we use Weighted Cross-Entropy Loss (WCEL) to increase the sensitivity of the learning model to minority classes. For this reason, significant weight will be applied to the minority class, while the majority class will be given a smaller weight. By applying this weighting, the cross-entropy loss can be formulated in Eq. 2.

$$L(y, o) = - w \sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (2)$$

With w is the weight of the calculated class using $w = \frac{Cmax}{Ci}$ where Cmax is the size of the majority class, and Ci is the number of data samples in the class c. Then M Then represent the number of classes, y is a binary (0 or 1) indicator if the class label *c* is the correct classification for the observation o and p is the predicted probability.

D. Build Neural Network

In this study, we use Bidirectional LSTM architecture because, according to[23], the highest accuracy value is when using bidirectional LSTM to perform multiclass text classification tasks. There are 64 neurons in each hidden layer (the more neurons, the more accuracy, but the computation time will be extended). We trained the network with a batch size of 8 based on[23], that smaller the batch size, the greater the accuracy, and applied a dropout with a probability of 0.6 to reduce the risk of overfitting. The higher the dropout value given, it will reduce overfitting, but if it is too high, it can decrease accuracy, so try n error only to determine the dropout value. The cost function we use is cross-entropy, specifically for applying the algorithm-level method in handling class imbalances. We use different weights for each class through the equations described previously. As a gradient descent optimization algorithm, we use the Adaptive Momentum Estimation (Adam) Optimizer with a learning rate of 0.001.

E. Evaluation

Evaluation with the right measuring tools is needed to compare the performance of several applied approaches. Considering that the minority class has a negligible impact on accuracy, apart from using Overall Accuracy, we also use other performance metrics, namely True Positive Rate (TPR), True Negative Rate (TNR), F1 Score, and Geometric Mean. F1 Score is a combination of Precision which calculates the positive class score classified in a positive class, and Recall, which represents how well the prediction of the positive class to be a single score. Next, by considering sensitivity (another term for Recall) and Specificity, which represents how well the prediction of the negative class is, we use the Geometric Mean (G-Mean) to combine the two into a single score. Since the deep learning algorithm is stochastic, we train five times for each method, then take the average Score of the five training results. The equations for calculating True Positive Rate (TPR) / Sensitivity / Recall / Hit Rate, True Negative Rate (TNR) / Specificity / Selectivity, Overall Accuracy, F1 Score, and G-Mean are as follows.

$$TPR = \frac{TP}{TP+FN} \quad (3)$$

$$TNR = \frac{TN}{TN+FP} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \quad (6)$$

$$G\text{-}Mean = \sqrt{Sensitivity \times Specificity} = = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (7)$$

IV. Results and Discussion

The results of this study are summarized in Table II. Based on our results, no method consistently provides the most superior performance using three predefined benchmarks. In the Customer Support Ticket Dataset, the highest accuracy was obtained without applying any method, while the highest F1 Score and G-Mean were obtained using WCEL. Furthermore, the highest accuracy was obtained using TL in the News Dataset, while the highest F1 Score and G-Mean were obtained using ROS. Finally, in the Drug Review Dataset, the highest accuracy and F1 Score were obtained without using any method, while the highest G Mean was obtained using WCEL.

TABLE II. COMPARISON OF THE PERFORMANCE IN EACH METHOD

Method	Cust. Support Ticket			News Category			Drug Review		
	Overall Accuracy	F1 Score (Average)	G-Mean (Average)	Overall Accuracy	F1 Score (Average)	G-Mean (Average)	Overall Accuracy	F1 Score (Average)	G-Mean (Average)
None	**0.836**	0.789	0.834	0.892	0.755	0.795	**0.930**	**0.837**	0.904
RUS	0.700	0.656	0.783	0.752	0.658	0.811	0.836	0.672	0.803
TL	0.796	0.722	0.767	**0.900**	0.787	0.825	0.878	0.697	0.801
ROS	0.826	0.784	0.828	0.892	**0.814**	**0.872**	0.916	0.832	0.904
SMOTE	0.810	0.760	0.799	0.896	0.803	0.859	0.904	0.806	0.886
WCEL	0.834	**0.798**	**0.867**	0.864	0.755	0.867	0.926	0.834	**0.906**

Table III shows the comparison of class-level performance for each method for the Customer Support Ticket Dataset. There are three classes/categories: Class 1: Account, Class 2: Customer Service, and Class 3: Transaction. Table IV for the News Category Dataset shows three classes/categories, namely Class 1: Entertainment, Class 2: Politics, Class 3: Tech. Table V the Drug Review Dataset shows five classes/categories, namely Class 1: ADHD, Class 2: Birth Control, Class 3: Tech, Class 4: Insomnia, Class 5: Weight Loss.

TABLE III. COMPARISON OF TPR, TNR, F1 SCORE, AND G-MEAN IN EACH METHOD ON THE NEWS CATEGORY DATASET

Method	True Positive Rate (TPR)			True Negative Rate (TNR)			F1 Score			G-Mean		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
None	0.746	0.892	0.722	**0.934**	0.786	0.958	0.770	0.882	0.714	0.834	0.838	0.830
RUS	0.708	0.648	**0.828**	0.810	0.866	0.864	0.618	0.752	0.598	0.756	0.750	0.844
TL	0.656	**0.896**	0.562	0.920	0.698	0.964	0.686	0.862	0.618	0.776	0.790	0.736
ROS	0.720	0.890	0.740	0.926	0.772	0.962	0.740	0.876	**0.736**	0.814	0.828	0.842
SMOTE	0.710	0.882	0.640	0.914	0.732	**0.972**	0.722	0.864	0.694	0.806	0.804	0.786
WCEL	**0.824**	0.848	0.806	0.908	**0.878**	0.942	**0.780**	**0.884**	0.730	**0.864**	**0.864**	**0.872**

In the Customer Support Ticket Dataset, WCEL has the most superior performance except for the F1 Score in Class 3. However, WCEL's performance is consistently superior to the original data or without applying any method. Although the accuracy produced by WCEL is still less than the original data, the resulting difference is not too far, which is 0.2%.

TABLE IV. COMPARISON OF TRUE POSITIVE RATE (TPR) AND TRUE NEGATVE RATE (TNR) IN EACH METHOD ON THE CUSTOMER TICKET DATASET

Method	True Positive Rate (TPR)			True Negative Rate (TNR)			F1 Score			G-Mean		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
None	0.870	0.934	0.360	0.926	0.848	**0.998**	0.856	0.924	0.484	0.898	0.890	0.598
RUS	0.788	0.736	**0.752**	0.812	0.876	0.922	0.718	0.816	0.440	0.798	0.804	**0.832**
TL	0.856	**0.948**	0.474	**0.944**	0.846	0.994	**0.864**	**0.930**	0.566	0.898	0.894	0.682
ROS	**0.886**	0.910	0.670	0.918	0.892	0.982	0.862	0.922	**0.658**	**0.904**	**0.900**	0.812
SMOTE	0.856	0.930	0.622	0.942	0.862	0.986	**0.864**	0.928	0.616	0.900	0.896	0.780
WCEL	0.864	0.874	0.718	**0.906**	0.954	0.954	0.844	0.908	0.514	0.890	0.886	0.826

In the News Dataset, SMOTE has consistently superior performance over the original data, although it is not superior to other methods. On the other hand, the TL performance is superior to the original data except for the G-Mean Score in Class 2. TL even has the highest accuracy compared to other methods.

In the Drug Review Dataset, the original dataset tends to be superior to the unbalanced class handling method. Although ROS has a reasonably high contribution to Class 4 and Class 5, ROS lowers the performance of Class 1, which is a minority class. In the case of this dataset, it seems that the minority class already has a pretty good performance. In theory, the learning algorithm will have difficulty detecting the minority class.

Based on the results we got, no method has the most superior performance over other methods. Overall, RUS had the worst performance, despite getting the highest G-Mean Score in Class 3 in the News Dataset. ROS and WCEL tend to improve the performance of the minority class, but at the same time, sometimes, these methods also reduce the performance of the majority class. Meanwhile, TL and SMOTE performed exceptionally well on the News Dataset, but not on the other two datasets that we used in this study.

Although the data-level method approach contributes to improving the performance of the minority class in some cases, the weakness of this approach is that it takes a long time to resample, except for RUS and ROS. The comparison of the time required for each data-level method is shown in Fig. 5. SMOTE and TL, which performed well on the News Dataset, took resampling time of more than 1.6 hours for SMOTE and more than 3 hours for TL. With a performance increase of <1%, this seems less applicable to the case of big data, where the amount of data available will be much larger, so the time required for resampling will also be longer.

TABLE V. COMPARISON OF TPR, TNR, F1 SCORE, AND G-MEAN IN EACH METHOD ON THE DRUG REVIEW DATASET

Method	True Positive Rate (TPR)					True Negative Rate (TNR)					F1 Score					G-Mean				
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 1	Class 2	Class 3	Class 4	Class 5	Class 1	Class 2	Class 3	Class 4	Class 5	Class 1	Class 2	Class 3	Class 4	Class 5
None	**0.952**	**0.990**	**0.962**	0.668	0.618	**1.000**	0.990	**0.992**	0.962	**0.970**	**0.956**	**0.996**	**0.950**	0.646	0.636	**0.976**	**0.990**	**0.978**	0.800	0.774
RUS	0.860	0.912	0.896	0.376	0.602	0.972	0.980	0.980	0.960	0.918	0.790	0.950	0.848	0.336	0.436	0.914	0.944	0.934	0.526	0.698
TL	0.830	0.984	0.876	0.252	0.706	0.988	0.964	0.988	**0.976**	0.928	0.840	0.980	0.872	0.254	0.538	0.906	0.972	0.930	0.416	0.782
ROS	0.904	0.974	0.886	0.668	**0.772**	0.990	0.978	0.990	**0.976**	0.962	0.886	0.980	0.896	**0.690**	**0.706**	0.946	0.972	0.936	0.806	**0.862**
SMOTE	0.874	0.968	0.888	**0.692**	0.628	0.988	0.964	0.988	0.962	0.968	0.874	0.976	0.886	0.650	0.642	0.930	0.968	0.936	**0.814**	0.780
WCEL	0.950	0.984	0.960	0.624	0.702	0.994	**0.994**	**0.992**	0.970	0.964	0.936	0.990	0.944	0.632	0.668	0.972	**0.990**	0.976	0.774	0.816

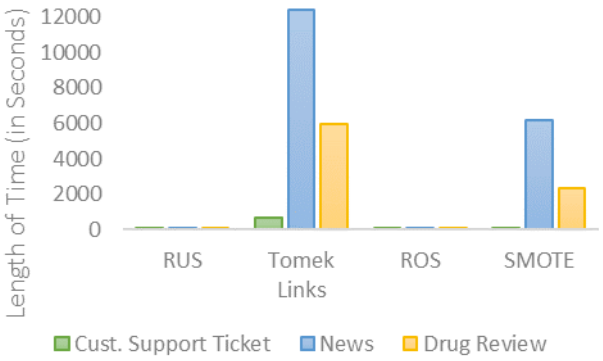


Fig. 5. Comparison of the Length of Time for Resampling.

V. CONCLUSION

This study discusses several testing methods for handling unbalanced classes, including RUS, TL, ROS, and SMOTE for data-level methods and WCEL for algorithm-level methods. We used three labeled text datasets with varying amounts of data and levels of balance. The neural network architecture we use is Bidirectional LSTM. We tested five times for each method. From the research results obtained, no method has the most superior performance because the quality of the data is also very influential on the learning model's performance. Both data-level and algorithm-level approaches can, in some cases, improve minority class performance. However, at the same time, The algorithm-level used in this study sometimes also reduces the performance of the majority class. Meanwhile, in the data-level approach, the resulting increase does not seem proportional to the time required for resampling.

Further research on the handling of unbalanced classes in deep learning is needed to produce consistent performance and be implemented effectively and efficiently.

ACKNOWLEDGMENT

Thanks to the Faculty of Computer Science, Universitas Amikom Yogyakarta, who helped in this study. This article was funded by Department of Information System, Universitas Amikom Yogyakarta - Indonesia.

REFERENCES

[1] S. Makki, Z. Assaghir, Y. Taher, R. Haque, MS Hacid, and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," IEEE Access, vol. 7, pp. 93010–93022, 2019, doi:10.1109/ACCESS.2019.2927266.

[2] S. Fotouhi, S. Asadi, and MW Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," J. Biomed. information., vol. 90, no. December 2018, p. 103089, 2019, doi:10.1016/j.jbi.2018.12.003.

[3] P. Ratadiya and R. Moorthy, "Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification," arXiv, 2019.

[4] T. G. Pratama, R. Hartanto, and N. A. Setiawan, "Machine learning algorithm for improving performance on 3 AQ-screening classification," Commun. Sci. Technol., vol. 4, no. 2, pp. 44–49, 2019, doi: 10.21924/cst.4.2.2019.118.

[5] JM Johnson and TM Khoshgoftaar, "Survey on deep learning with class imbalance," J. Big Data, vol. 6, no. 1, 2019, doi:10.1186/s40537-019-0192-5.

[6] M. Buda, A. Maki, and MA Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural Networks, vol. 106, pp. 249–259, 2018, doi:10.1016/j.neunet.2018.07.011.

[7] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and PJ Kennedy, "Training deep neural networks on imbalanced data sets," Proc. int. jt. Conf. Neural Networks, vol. 2016-Octob, pp. 4368–4374, 2016, doi:10.109/IJCNN.2016.7727770.

[8] S. Pouyanfar et al., "Dynamic Sampling in Convolutional Neural Networks for Imbalanced Data Classification," Proc. - IEEE 1st Conf. Multimed. inf. Process. Retrieval, MIPR 2018, pp. 112–117, 2018, doi:10.109/MIPR.2018.00027.

[9] K. Arai, A. Yamashita, and H. Okumura, "Head Position and Pose Model and Method for Head Pose Angle Estimation based on Convolution Neural Network" International Journal of Advanced Computer Science and Applications(IJACSA), vol. 12, no. 10, pp. 42–49, 2021. Available: http://dx.doi.org/10.14569/IJACSA.2021.0121006

[10] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," prog. Arti. Intell., vol. 5, no. 4, pp. 221–232, 2016, doi:10.1007/s13748-016-0094-0.

[11] P. Branco, L. Torgo, and R. Ribeiro, "A Survey of Predictive Modeling under Imbalanced Distributions," pp. 1–48, 2015, [Online]. Available: http://arxiv.org/abs/1505.01658.

[12] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer

learning," in 2016 IEEE International Conference on Image Processing (ICIP), Sept. 2016, pp. 3713–3717, doi:10.1109/ICIP.2016.7533053.

[13] P. Hensman and D. Masko, "The Impact of Imbalanced Training Data for Convolutional Neural Networks," Ph.D., 2015, [Online]. Available: https://www.kth.se/social/files/588617ebf2765401cfcc478c/PHensmanDMasko_dkand15.pdf.

[14] SH Khan, M. Hayat, M. Bennamoun, FA Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 8, pp. 3573–3587, 2018, doi:10.1109/TNNLS.2017.2732482.

[15] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, and A. Kronzer, "Predicting Hospital Readmission via Cost-Sensitive Deep Learning," IEEE/ACM Trans. Comput. Biol. Bioinformatics., vol. 15, no. 6, pp. 1968–1978, 2018, doi:10.109/TCBB.2018.2827029.

[16] Y. Geng and X. Luo, "Cost-sensitive convolution-based neural networks for imbalanced time-series classification," arXiv, 2018.

[17] R. Harliman and K. Uchida, "Data- and algorithm-hybrid approach for imbalanced data problems in deep neural networks," int. J. Mach. Learn. Comput., vol. 8, no. 3, pp. 208–213, 2018, doi:10.18178/ijmlc.2018.8.3.689.

[18] J. Ortigosa-Hernández, I. Inza, and JA Lozano, "Measuring the class-balance extent of multi-class problems," Pattern Recognition. Lett., vol. 98, pp. 32–38, 2017, doi:10.1016/j.patrec.2017.08.002.

[19] R. Misra, "News Category Dataset," 2018. https://www.kaggle.com/rmisra/news-category-dataset.

[20] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning," in Proceedings of the 2018 International Conference on Digital Health, Apr. 2018, pp. 121–125, doi:10.1145/3194658.3194677.

[21] R. M. Pereira, Y. M. G. Costa, and C. N. Silla, "MLTL: A multi-label approach for the Tomek Link undersampling algorithm: MLTL: The Multi-Label Tomek Link," Neurocomputing, vol. 383, pp. 95–105, 2020, Available: https://doi.org/10.1016/j.neucom.2019.11.076.

[22] F. Charte, A. J. Rivera, M. J. Del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," Knowledge-Based Syst., vol. 89, pp. 385–397, 2015, [Online]. Available: https://doi.org/10.1016/j.knosys.2015.07.019.

[23] A. Hikmah, S. Adi, and M. Sulistiyono, "The Best Parameter Tuning on RNN Layers for Indonesian Text Classification," in the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 94–99.