APTIKOM
ASOSIASI PENDIDIKAN TINGGI INFORMATIKA DAN KOMPUTER
INDONESIAN ASSOCIATION OF HIGHER EDUCATION IN INFORMATICS AND COMPUTING

PROCEEDING

THE NINTH INTERNATIONAL CONFERENCE ON INFORMATICS AND COMPUTING (ICIC)

2024

24-25 October 2024
MEDAN, INDONESIA

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956352

ISBN: 979-8-3315-1760-1

2024 Ninth International Conference on Informatics and Computing (ICIC)

Medan, Indonesia

(Hybrid Conference)

October 24-25, 2024

ISBN: 979-8-3315-1760-1

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10957181

2024 Ninth International Conference on Informatics and Computing (ICIC)

Medan, Indonesia (Hybrid)

Phone: +6281384175979

Email: conference.icic@gmail.com

Website: https://icic-aptikom.org

October 24-25, 2024

ISBN: 979-8-3315-1760-1

2024 Ninth International Conference on Informatics and Computing (ICIC)

ISBN: 979-8-3315-1760-1

Additional copies of this publication are available from

Curran Associates, Inc.

57 Morehouse Lane

Red Hook, NY 12571 USA

+1 845 758 0400

+1 845 758 2633 (FAX)

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956314

The 2024 International Conference on Informatics and Computing (ICIC) Committee

Steering Committee:
Achmad Benny Mutiara, Gunadarma University, Indonesia
Husni Teja Sukmana, UIN Syarif Hidayatullah Jakarta, Indonesia

General Chair:
Yusuf Durachman, UIN Syarif Hidayatullah Jakarta, Indonesia

Program Committee Co-chair:
Achmad Nizar Hidayanto, Universitas of Indonesia, Indonesia
Husni Teja Sukmana, UIN Syarif Hidayatullah Jakarta, Indonesia

Program Co-chair:
Husni Teja Sukmana, UIN Syarif Hidayatullah Jakarta, Indonesia
Prihandoko, Gunadarma University, Indonesia
Sandy Kosasi STMIK Pontianak, Indonesia
Henderi, Raharja University, Indonesia

Publication Co-chairs:
Shinta Oktaviana, Universitas Nusa Mandiri, Indonesia
Dewi Khairani, UIN Syarif Hidatullah Jakarta, Indonesia
Evi Triandini, STIKOM Bali, Indonesia

Publicity Co-chairs:
Solikin, University of Bina Insani, Indonesia
Hanny Hikmayanti Handayani, University of Buana Perjuangan Karawang, Indonesia

Web Development:
Dewi Khairani, UIN Syarif Hidatullah Jakarta, Indonesia

TPC members:
Arli Aditya Parikesit, Indonesia International Institute for Life Sciences
Bagus Andra, Universitas Nusa Mandiri
Rifiana Arief, Gunadarma University
Artika Arista, Universitas Pembangunan Nasional Veteran Jakarta
Ria Astriratma, Universitas Pembangunan Nasional Veteran Jakarta
Lintang Yuniar Banowosari, Gunadarma University
Indra Budi, Fasilkom UI
Nurul Chamidah, Universitas Indonesia
Aang Darmawan, Prodi Sistem Informasi Universitas Islam Madura
David David, STMIK Pontianak
Yusuf Durachman, State Islamic University of Syarif Hidayatullah Jakarta
Made Harta Dwijaksara, Universitas Indonesia
Windu Gata, Universitas Nusa Mandiri
Rahmadya Handayanto, universitas islam 45
Robby Kurniawan Harahap, Gunadarma University
Muhammad Haris, Universitas Nusa Mandiri

PREFACE

Dear esteemed guests, distinguished speakers, researchers, and participants,

It is my distinct pleasure to welcome you all to the 2024 International Conference on Informatics and Computing (ICIC 2024), held in the vibrant city of Medan, North Sumatera. As the Chair of this year's conference, I am honored to see such an exceptional turnout from around the globe, where we gather to share knowledge, exchange ideas, and explore the cutting-edge developments in informatics, computing, and related fields.

This year's conference has been particularly competitive, with 458 paper submissions from various academic institutions and research centers worldwide. After a rigorous peer-review process, we are proud to have accepted 246 papers, resulting in an acceptance rate of approximately 53.4%. This reflects the high standard of quality and innovation in the research presented at ICIC 2024. We are thrilled to showcase these outstanding contributions, which will undoubtedly advance the discourse in our field. All accepted papers are submitted to IEEE Xplore. IEEE Conference Number: #64337 ISBN: 979-8-3315-1760-1

The theme of ICIC 2024 resonates deeply with our collective goal of advancing informatics and computing technologies, especially in areas such as artificial intelligence, data science, cybersecurity, and cloud computing. These fields are pivotal in driving not only academic progress but also societal growth, economic development, and global competitiveness.

I would like to extend my heartfelt gratitude to the organizing committee, our reviewers, authors, and all contributors for their dedication and hard work in making this conference possible. I am confident that the discussions, presentations, and collaborations forged during ICIC 2024 will lead to groundbreaking advancements that will shape the future of informatics and computing.

Once again, welcome to ICIC 2024 in Medan, North Sumatera. I hope you all have a rewarding and insightful experience, and I look forward to the exciting innovations and ideas that will emerge over the course of the conference.

Sincerely,

Dr. Yusuf Durachman, M.I.T.

Chair, ICIC 2024

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956560

Table of Contents

2024 Ninth International Conference on Informatics and Computing (ICIC) | 979-8-3315-1760-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC64337.2024.10956881

Editing Team

A Comparison of Random Forest and Support Vector Machine Classification Algorithms for Imbalanced and Balanced Rodent Tuber Dataset with Random Oversampling Method

Iwan Binanto
Informatics, Sanata Dharma University
Yogyakarta, Indonesia
Corresponding author: iwan@usd.ac.id

Marselinus Sandimus Jamlu
Informatics, Sanata Dharma University
Yogyakarta, Indonesia
sandigroot@gmail.com

Robertus Denyva Adibuana Wibisono
Informatics, Sanata Dharma University
Yogyakarta, Indonesia

Nesti Fronika Sianipar
Biotechnology Departement,
Faculty of Engineering, Bina Nusantara University.
Research Center Food Biotechnology,
Bina Nusantara University,
Jakarta, Indonesia
nsianipar@binus.edu

Abstract*— Class distributions that are assumed to be balanced by conventional machine learning algorithms make them susceptible to biases that favor the dominant class throughout the classification process. This study primarily discusses how classification results may be impacted by the dataset's state, whether it is balanced or unbalanced. The Rodent Tuber's unbalanced dataset, which includes chemical data useful for cancer diagnosis, was utilized in this investigation. Support Vector Machine with three kernels and Random Forest will be used to classify this dataset. To see the differences, Random Oversampling is used to do classification on both an imbalanced and balanced dataset. The outcomes demonstrate that better results with shorter running times are obtained from a balanced dataset. The Random Forest method yielded the best classification results on the balanced dataset.

Keywords—Random Forest, Support Vector Machine, SVM, Rodent Tuber, Imbalanced Data, Random Oversampling, Balancing Algorithm

I. INTRODUCTION

One of the deadliest diseases in the world is still cancer, and traditional medicine has long employed medicinal herbs to treat a variety of ailments because of its natural properties and minimal potential side effects compared to synthetic drugs. An indigenous plant of Indonesia known for its detoxifying and anti-cancer qualities is the rodent tuber [1]. The Rodent Tuber (Typhonium flagelliforme) was one of the therapeutic plants that caught the interest of researchers. The whole portion of the plant contains anti-cancer compounds in the roots, tubers, stems, and leaves [2]. It has been used to treat the cervix [3], liver [4], and breast cancer [5].

Data extraction from this plant revealed a data imbalance that led to significant discrepancies between majority and minority classes [6]. One significant issue with unbalanced data is that, in the instance of the Rodent Tuber dataset, the number of anti-cancer compounds found may be far lower than the total number of common compounds in the plant. Models may overlook the significant minority class of anti-cancer medications because of this disparity.

Several studies on the classification of the Rodent Tuber dataset using different classification algorithms and balancing data techniques compared to this research have been conducted, with results showing that some classification algorithms are greatly affected by imbalance data [6]-[10].

Resampling using Random Over Sampling (ROS) is one approach to solving the issue. This technique duplicates data from the minority class to bring it up to par with the majority class, hence reducing the ratio of inequality across classes [11],[12].

II. LITERATURE REVIEW

Conventional machine learning algorithms often presume a balanced distribution of classes, which leaves them vulnerable to biases during the classification process that favours the dominant class. When dealing with unbalanced datasets like as the Rodent Tuber data, this bias might be very harmful. A major problem with imbalanced data is that, in the case of the Rodent Tuber dataset, there may be many less discovered anti-cancer chemicals than there are common compounds in the plant. Because of this discrepancy, models may fail to include the important minority class of anti-cancer drugs [6], [7].

In recent years, several studies have explored methods for addressing imbalanced data classification problems [1], [3], [4]. One of them is a study conducted by Iwan Binanto et al. [4] on data from the Rodent Tuber (Typhonium flagelliforme) plant, highlighting significant disparities in the number of target classes between anti-cancer compounds and common compounds.

The key characteristic of imbalanced data lies in the potential for neglecting crucial data points from the minority class during classification. This happens because these data streams are often disregarded because of their low number and

perceived lower priority in addressing real-world problems. As a result, classifying imbalanced datasets can be challenging, often resulting in subpar performance for the minority class, even when the overall model achieves high accuracy [5], [6].

Support Vector Machine (SVM) model trained on imbalanced data resulted in poor performance for the minority class (class 0), with both recall and precision being very low [6]. However, applying random oversampling (ROS) to balance the data significantly improved the model's performance on the minority class, achieving perfect precision (100%) and reasonable recall (60%) for class 0. This suggests that random oversampling can be an effective technique for dealing with imbalanced data, especially compared to under-sampling approaches [11].

Conversely, Random Forest (RF) has greater robustness against data imbalance [2], [11], [13]. Random Forest (RF) can easily handle missing data and performs well in a variety of uneven datasets. According to a study by A.S. More et al. [8], Random Forest (86.448%) predicts unbalanced data more accurately than Support Vector Machine (SVM) (63.551%).

This research will analyze the performance of Support Vector Machine (SVM) and Random Forest (RF) models on the Rodent Tuber dataset with both balanced and unbalanced data distributions. By employing random oversampling to balance the data, we aim to compare the effectiveness of these techniques using various performance metrics [14],[15].

III. Research Method

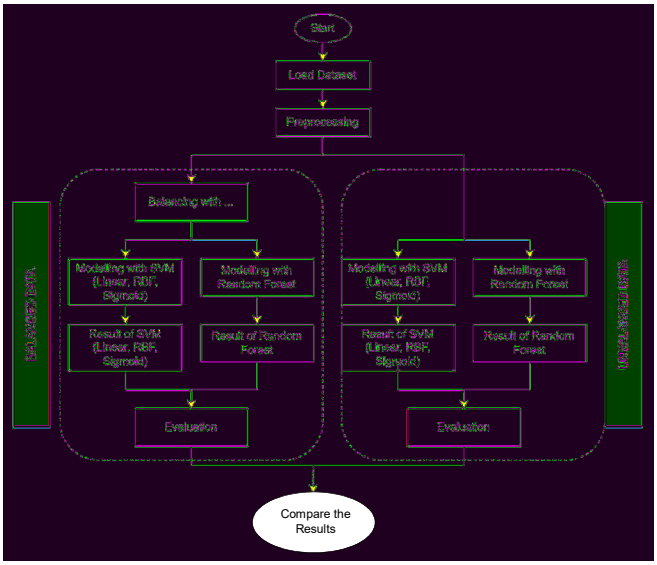

Fig. 1. Research Method

The research method for this study is illustrated in Figure 1, which shows that Support Vector Machine (SVM) with its three kernels and Random Forest (RF) will be compared on both balanced and imbalanced datasets. The results will then be compared.

An instance of an imbalanced dataset included in this research is the Rodent Tuber (Typhonium Flagelliforme) data [16], which has 663,228 rows and 7 columns, namely "Retention Time", "m/z", "Intensity", "Real_m/z", "Compound Name", "Formula", "Status". This dataset was obtained from the extraction of Rodent Tuber (Typhonium Flagelliforme) using the LCMS technique.

The features from the dataset that will be used are "Retention Time", "m/z", "Intensity", "Real_m/z", "Status" because these are numeric.

After that, this data will go through a splitting phase [12] where 80% of the data will be used for training and 20% for testing. The random state of 42 is set for both the Random Forest (RF) and Support Vector Machine (SVM). For these experiments, laptops with an NVIDIA GeForce RTX 2050 GPU, and 8GB RAM were employed.

The processing time for each algorithm is calculated as the total time taken to compute accuracy, precision, recall, and F1-score.

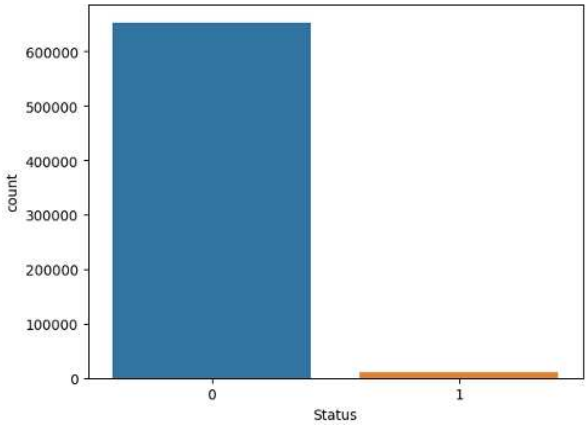

Fig. 2. Imbalanced dataset of Typhonium Flagelliforme

IV. Results and Discussions

It can be seen on Table 1 that the accuracy of Support Vector Machine (SVM) and Random Forest (RF) does not encounter any issues in classification, even with imbalanced data. However, accuracy alone is not sufficient, as it can lead to bias. Therefore, Precision, Recall, and F1-score are also calculated. The results show that Support Vector Machine (SVM) cannot provide results with any of its kernels. This happens because Support Vector Machine (SVM) is biased towards the majority class. Support Vector Machine (SVM) works by finding a hyperplane with the maximum margin between classes. In imbalanced data, the majority class tends to dominate the determination of this hyperplane. Support Vector Machine (SVM) generally assumes that the dataset has a balanced class distribution, which is not the case with imbalanced data.

TABLE I. Model Performance on Imbalance Dataset

	Imbalance Dataset			
	RF	SVM		
		linear	RBF	sigmoid
Accuracy	0.99	0.99	0.99	0.98
Recall	0.99	0.00	0.03	0.00
Precision	0.99	0.00	0.00	0.00
F1-Score	0.99	0.00	0.00	0.00
Process Time (s)	493	1.179	1.235	795,0261

This phenomenon where accuracy is high but precision, recall, and F1 score are very low (even 0) often occurs in imbalanced datasets. This is known as the "accuracy paradox".

It is required to balancing the dataset because of its extremely imbalanced nature between target 0 and 1. Random Over Sampling (ROS) is used to accomplish balancing by randomly selecting and duplicating data point from the minority class. The advantages of Random Over Sampling (ROS) are its simplicity and ease of implementation, as well as not losing information. The dataset is shown in Fig. 3 following the Random Over Sampling (ROS) implementation.

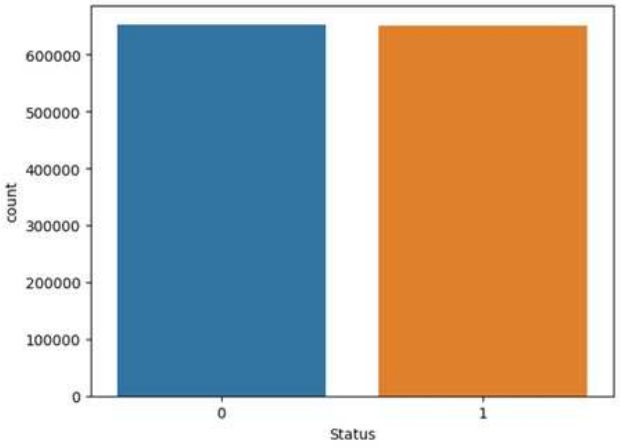

Fig. 3. Balanced dataset of Typhonium Flagelliforme

After balancing the data with Random Over Sampling (ROS), a significant difference in precision, recall, and F1 score values is observed in Support Vector Machine (SVM) as shown in Table 2.

After the data was balanced, there were differences in processing time. Some were slower, while others were faster. Random Forest (RF) became slower, while the Support Vector Machine (SVM) with the sigmoid kernel became significantly faster, although the other two kernels became slower. Overall, Support Vector Machine (SVM) has a faster processing time compared to Random Forest (RF).

We can see in Table 2 that in Random Forest (RF), recall is lower than precision, because there is a possibility of overfitting caused by Random Over Sampling (ROS).

TABLE II. MODEL PERFORMANCE ON BALANCE DATASET

	Balanced Dataset			
	RF	**SVM**		
		linear	**RBF**	**sigmoid**
Accuracy	0.69	0.65	0.73	0.60
Recall	0.62	0.99	0.91	0.59
Precision	0.99	0.62	0.51	0.60
F1-Score	0.76	0.76	0.65	0.60
Process Time (s)	704	2.400	1.568	4.834

Recall is higher than precision in Support Vector Machine (SVM) in balanced dataset due to the nature of the model's performance in identifying positive instances.

The results of the experiments with imbalanced dataset indicate that Random Forest (RF) better than Support Vector Machine (SVM) in terms of overall accuracy and process time. But, in a balanced dataset, the processing time of Random Forest (RF) is longer and significantly greater compared to Support Vector Machine (SVM). This may be because the model requires more trees to achieve the desired accuracy, which can extend the processing time and increasing the possibility of overfitting.

However, Random Over Sampling (ROS) strategy is another essential component that makes the model capable of correctly classifying the input.

V. CONCLUSIONS

Random Forest (RF) is an effective method for handling imbalanced datasets, as in previous research [6] – [9] because each tree is trained on a different bootstrap sample, allowing it to better detect patterns specific to minority classes. Support Vector Machine (SVM) models achieve high accuracy and recall on balanced data, but struggle on imbalanced data. Beside of that, it can be concluded that the best SVM kernel is the RBF kernel in balanced dataset.

This research is still preliminary in nature, there is a lack of literature review and in-depth knowledge about the algorithms used so we cannot discuss the results of this research properly.

Future research should explore other techniques like cost-sensitive learning or under-sampling methods, as well as the impact of feature selection or engineering on model performance.

REFERENCES

[1] I. Binanto et al., "Comparison of Similarity Coefficients on Morphological Rodent Tuber," 2019, doi: 10.1109/INAPR.2018.8627050.

[2] C. Y. Choo, K. L. Chan, K. Takeya, and H. Itokawa, "Cytotoxic activity of Typhonium flagelliforme (Araceae)," Phytotherapy Research, vol. 15, no. 3, pp. 260–262, May 2001, doi: 10.1002/ptr.717.

[3] E. Purwaningsih, "UNIVERSA MEDICINA," 2014.

[4] J. Teknologi, N. Fronika Sianipar, R. Purnamaningsih, D. Lolita Gumanti, and M. Vidianty, "Analysis of Gamma Irradiated Fourth Generation Mutant of Rodent Tuber (Typhonium Flagelliforme Lodd.) Based On Morphology and RAPD Markers", 2016. [Online]. Available: www.jurnalteknologi.utm.my

[5] A. Nurrochmad, E. Lukitaningsih, and E. Meiyanto, "Anti cancer activity of rodent tuber (Thyphonium flagelliforme (lodd.) Blume on human breast cancer t47d cells," International Journal of Phytomedicine, vol. 3, pp. 138–146, 2011, [Online]. Available: http://www.arjournals.org/index.php/ijpm/index

[6] I. Binanto, N. F. Sianipar, F. Dea, M. N. Primadani, and T. W. Kartikasari, "Klasifikasi Senyawa Keladi Tikus Menggunakan Algoritma KNN, Gaussian Naïve Bayes dengan Menerapkan Imbalance Data Borderline SMOTE", Prosiding Sains Nasional dan Teknologi, vol. 13, no. 1, p. 377, Nov. 2023, doi: 10.36499/psnst.v13i1.9005.

[7] I. Binanto, N. F. Sianipar, N. M. D. Aprilianti, J. Gein, and P. D. Paska, "Analisis Perbandingan Algoritma KNN, Gaussian Naïve Bayes, Random Forest untuk Data Tidak Seimbang dan Data yang Diseimbangkan dengan Metode TomekLink Undersampling pada Dataset LCMS Tanaman Keladi Tikus", Prosiding Sains Nasional dan Teknologi, vol. 13, no. 1, p. 156, Nov. 2023, doi: 10.36499/psnst.v13i1.9002.

[8] I. Binanto and N. F. Sianipar, "Perbandingan Algoritma Klasifikasi Random Forest, Gaussian Naive Bayes, dan KNearest Neighbor untuk Data Tidak Seimbang dan Data yang diseimbangkan dengan Metode Adaptive Synthetic", doi: 10.35842/sintaks.v2i1.30.

[9] I. Binanto and N. F. Sianipar, "Perbandingan Algoritma Klasifikasi Random Forest, Gaussian Naive Bayes, dan K-Nearest untuk Data

Tidak Seimbang dan Data yang diseimbangkan dengan metode Random Undersampling pada", doi: 10.35842/sintaks.v2i1.28.

[10] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," IEEE Access, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[11] Z. Zheng, Y. Cai, Y. Li, Z. Zheng, Y. Cai, and Y. Li, "Oversampling Method for Imbalanced Classification", 2015.

[12] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in 2020 11th International Conference on Information and Communication Systems, ICICS 2020, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

[13] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," in Procedia Computer Science, Elsevier B.V., 2019, pp. 503–513. doi: 10.1016/j.procs.2019.12.017.

[14] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans Knowl Data Eng, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[15] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, no. 4. Springer Verlag, pp. 221–232, Nov. 01, 2016. doi: 10.1007/s13748-016-0094-0.

[16] N. Fronika Sianipar, R. Purnamaningsih, J. Teknologi Pangan, U. Bina Nusantara, J. Alam Sutera Boulevard No, and B. Besar Penelitian dan Pengembangan Bioteknologi dan Sumber Daya Genetik Pertanian, "Ethos (Jurnal Penelitian dan Pengabdian M asyarakat): 65-74 Pengembangan Tanaman Keladi Tikus (Typhonium Flagelliforme Lodd.) Asal Indonesia sebagai Obat Antikanker"