Studies in Computational Intelligence

Volume 1227

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI AG (Switzerland), zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Ford Lumban Gaol · Tokuro Matsuo · Takayuki Ito
Editors

Advances in Smart Knowledge Computing

Towards Post Artificial Intelligence Era

Springer

Editors
Ford Lumban Gaol
Department of Informatics Engineering
and Information System
BINUS University
Jakarta, Indonesia

Tokuro Matsuo
Advanced Institute of Industrial Technology
Shinagawa, Japan

Takayuki Ito
Department of Social Informatics
Kyoto University
Kyoto, Japan

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-032-01132-9 ISBN 978-3-032-01133-6 (eBook)
https://doi.org/10.1007/978-3-032-01133-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Exploring the Data Balance Effect: Artificial Neural Network Classification on Rodent Tuber's Liquid Chromatography Mass Spectrometry Data

Iwan Binanto[1,*], Antonius Miquel Aureliano[1], Antonius Yoga Chris Raharja[1], Martinus Angger Budi Wicaksono[1], Nesti F. Sianipar[2]

[1]Department of Informatics, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia
iwan@usd.ac.id, {antoniusmiquel.39,niusyoga78,angger-budi9}@gmail.com
[2]Food Technology Department, Faculty of Engineering
Research Interest Group Biotechnology
Bina Nusantara University
Jakarta, Indonesia
nsianipar@binus.edu

Abstract. This paper highlights the challenges associated with unbalanced data by examining how balanced data affects Artificial Neural Network (ANN) classification on Rodent Tuber's Liquid Chromatography Mass Spectrometry data. Chemical disparity in the dataset drives a reexamination of traditional algorithms and their preference for balanced data. We present ANN as an alternative, discussing its benefits, such as fault tolerance and flexible computation. The study assesses ANN performance using balancing data strategies: Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Random Under-Sampling (RUS), and NearMiss. ADASYN achieved the best outcome and ANN still cannot provide optimal results with balanced data. So, additional studies and continuous improvements are needed.

Keywords: Rodent Tuber, balanced and imbalanced data, balancing algorithms, Artifical Neural Networks.

1 Introduction

Rodent Tuber (Typhonium flagelliforme Lodd.) is a native plant in Indonesia known for its detoxification and anticancer properties used in traditional medicine for years[1]. Liquid Chromatography Mass Spectrometry data from Rodent Tuber plants, previously researched by Sianipar et al [2]–[7], was used in research conducted by Binanto et al [8]. The disparity between the anti-cancer and common chemicals shown in binary targets throws this data out of balance. This is due to the specific properties unique to a few chemicals contained in Rodent Tuber that have the ability to prevent cancer cell growth or destroy cancer cells.

The use of unbalanced data in classification using machine learning can produce inaccurate results. Other research on this topic has predominantly used Random Forest, Gaussian Naïve Bayes, and K-Nearest Neighbor algorithms, all focusing on balanced data [9]–[12].

This emphasis on balanced data in the existing literature prompts a reevaluation of methodology. Therefore, this study aims to introduce Artificial Neural Networks (ANN) as an alternative approach. In contrast to other traditional algorithms, ANN are widely utilized as classification models for binary outcomes in medicine. Neural networks have several advantages, including the ability to store information throughout the network, work with incomplete knowledge, and fault tolerance. They also have distributed memory and parallel processing capabilities [13].

This study explores class imbalance in Rodent Tuber dataset using Artificial Neural Networks (ANN) and balancing strategies, highlighting the benefits and challenges of using ANN with unbalanced data.

2 Literature Review

Previous studies on this topic have only used conventional algorithms, including Random Forest, Gaussian Naïve Bayes, and K-Nearest Neighbor, focusing on balanced data using various balancing techniques [9]–[12]. From previous research, the KNN, Gaussian Naïve Bayes, and Random Forest algorithms on real data (unbalanced data) produced accuracy values of 0.984, 0.985, and 1, respectively. On balanced data using borderline SMOTE, the accuracy was 0.967, 0.499, and 0.984, respectively [12].

In response to the limitations posed by traditional algorithms and the emphasis on balanced data, this study proposes the integration of Artificial Neural Networks (ANN) as an alternative approach. New systems and computational techniques for machine learning, knowledge demonstration, and ultimately the application of information learned to maximize the output responses of complex systems are known as artificial neural networks (ANN) [14]. ANN is based on how nervous systems, such as the brain, manage and process information. On a much smaller scale, they are concentrated on the neuronal architecture of the cerebral cortex. Many experts in artificial intelligence regard ANN as the greatest, if not the only, option for creating intelligent machines [15].

ANN have several key features. ANN offer flexible computation, similar to biological neural networks, and can handle incomplete knowledge. They are fault-tolerant and have distributed memory, requiring examples for learning [13], [16]. However, their success depends on the quality of examples provided, as incorrect outputs may occur if certain aspects are not shown. There are many studies on ANN, including the use of ANN for cervical cancer classification [17], the comparison of ANN with CNN [18], and decision-making of health care organization using ANN [19].

The accuracy of many classifications is severely affected by imbalanced datasets in machine learning approaches. The actual time environment is comparatively more prone to inconsistencies like imbalance and data with noise [20]. Despite their effectiveness, traditional methods like Random Forest, Gaussian Naïve Bayes, and K-Nearest Neighbor have limitations in handling imbalanced data. Recent advancements, such as Generative Adversarial Networks (GANs), have shown promise in addressing class imbalance but have not been thoroughly explored in this context.

3 Research Method

The dataset of Rodent Tuber [2]–[7] will be used in this work and preprocessed ready, which will result in an initially unbalanced dataset as shown in Fig 1 and Fig 2.

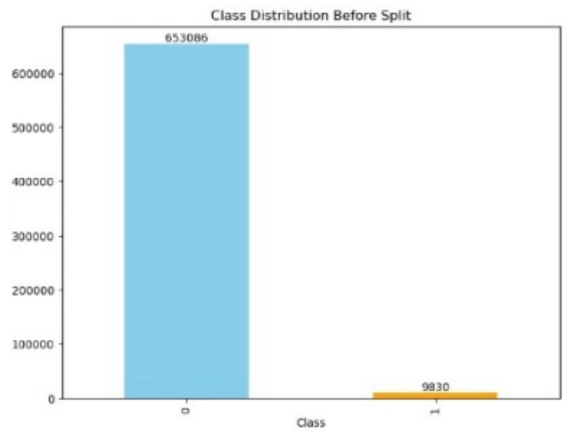

Fig. 1. Unbalanced Dataset Graph

	Retention Time	Intensity	m/z	Real_m/z	NamaSenyawa	RumusSenyawa	Status
0	5.022	281952	60.057205	60.03236	Urea	CH4N2O	0
1	5.022	198032	60.246964	60.03236	Urea	CH4N2O	0
2	5.022	43064	66.888519	67.04220	Pyrrol	C4H5N	0
3	5.022	63172	67.268036	67.04220	Pyrrol	C4H5N	0
4	5.022	57852	68.280075	68.03745	Imidazole	C3H4N2	0
...	...	...	...	...	...	...	...
663223	2401.000	456	1188.677734	1188.42798	Man2XylManGlcNAcFucGlcNAc	C45H76N2O34	0
663224	2401.000	191	1191.460938	1191.38086	Foetoside C	C58H94O25	0
663225	2401.000	191	1193.358398	1193.40686	Man6GlcNAc-I	C44H75NO36	0
663226	2401.000	191	1195.192871	1195.35303	Furostane base -2H + O-Hex, O-Hex-dHex-dHex-dHex	C57H94O26	0
663227	2401.000	419	1195.572266	1195.35303	Furostane base -2H + O-Hex, O-Hex-dHex-dHex-dHex	C57H94O26	0

663228 rows × 7 columns

Fig. 2. Unbalanced Preprocessed Dataset [8]

The dataset will be transformed into multiple balanced representation by applying Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Random Under-Sampling (RUS), and Near Miss Under-sampling (NearMiss) to solve the class imbalance. The type of data distribution (balanced or imbalanced) and its effect on the effectiveness of artificial neural networks (ANN) for predicting are the two main factors that drive our research. On both imbalanced and balanced datasets, in-depth model validation and training involving several neural network designs were carried out. A thorough analysis was conducted using confusion matrices and performance indicators such as accuracy, precision, recall, and F1 score.

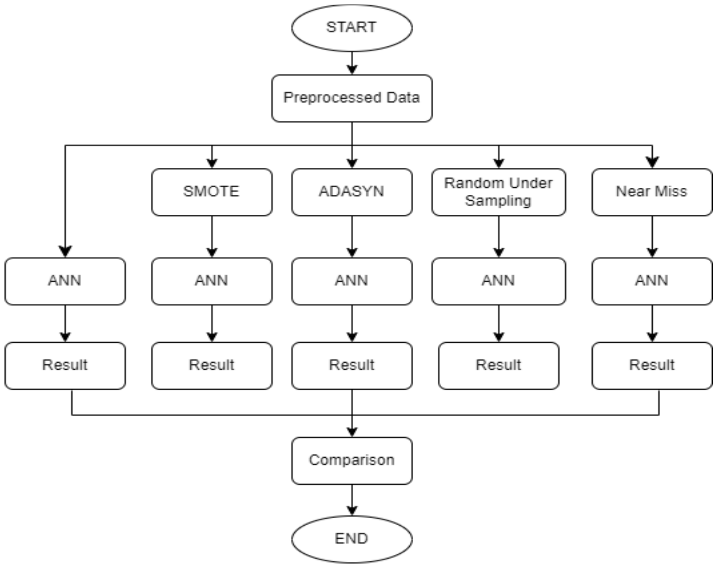

Fig. 3. Research Method

Fig. 3 shows the research method that shows the workflow. These images provide a visual overview of the experimental setup with default parameters, assisting in a thorough understanding of the study's methodology.

4 Results and Discussions

Before balancing, the dataset is split into two datasets: the training dataset and the testing dataset. The training dataset is then balanced, while the testing dataset will be used to test the created ANN model.

The results of balancing of Rodent Tuber’s dataset using SMOTE, ADASYN, Random Under Sampling (RUS), and Near Miss are shown in Fig. 4 - 7. It can be seen that the number of data points changes when balanced. This is related to the way the balancing algorithms work.

Fig. 4. Balanced Dataset with SMOTE

Fig. 5. Balanced Dataset with ADASYN

Fig. 6. Balanced Dataset with Random Under Sampling

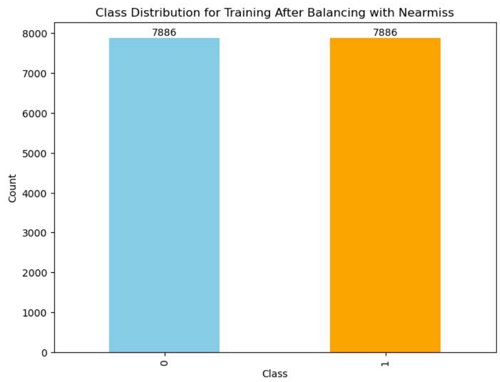

Fig. 7. Balanced Dataset with Near Miss

For the ANN model trained using the original imbalanced dataset, the accuracy is 99%. While the accuracy is very high compared to the rest, the confusion matrix in Fig. 8 shows the flaw of the model. It predicted everything as Positive and none as Negative, resulting a staggering amount of True Positive, and False Positive. As stated before, the success of ANN depends on the quality of data that is feed into the model. With the dataset being highly imbalanced, the model treats the data with the class of 1 as noise, and predicted it as class 0.

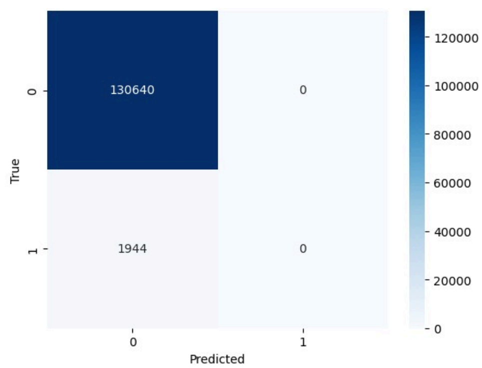

Fig. 8. Confusion Matrix of Rodent Tuber's Imbalanced Data

For ANN model trained using balanced dataset with SMOTE, the accuracy is 56%. While not as high as the imbalanced dataset, the confusion matrix in Fig. 9 shows that the model predicted was able to predict the True Negative. But, the model struggles to differentiate between True Positive and True Negative. Meaning that a lot of normal compound is predicted as the anti-cancer compound.

For ANN model trained using dataset balanced with ADASYN, it resulted in an accuracy of 65%. The model did a better job than with SMOTE, but still struggles to differentiate between True Positive and True Negative. The similarity in results between SMOTE and ADASYN raises intriguing questions about the underlying dynamics of oversampling techniques in the context of our dataset. One potential explanation is that, despite their distinct mechanisms, both SMOTE and ADASYN introduce synthetic instances to the minority class, influencing the model in similar ways. This shared characteristic might contribute to the observed difficulty in accurately discriminating between positive and negative instances.

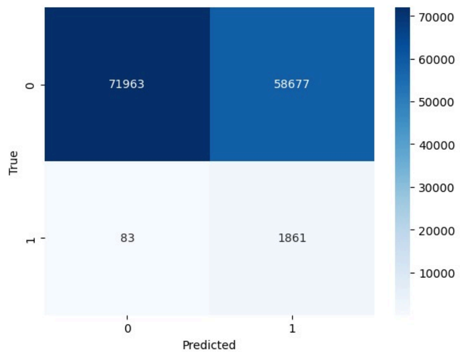

Fig. 9. Confusion Matrix of Rodent Tuber's Balance Data with SMOTE

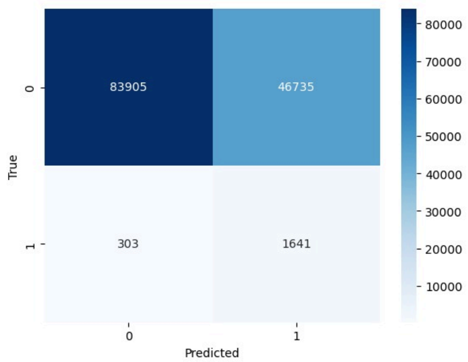

Fig. 10. Confusion Matrix of Rodent Tuber's Balance Data with ADASYN

For ANN model trained using dataset balanced with Random Undersampling (RUS), the accuracy is 37%. The model did a worse job than with SMOTE and ADASYN. The confusion matrix in Fig. 11 shows that It is the opposite of both SMOTE and ADASYN, meaning, it predicted more False Negative than True Positive.

For ANN model trained using dataset balanced with Near Miss, the accuracy is 62%. The model did a fairly decent job, but still worse than the model using ADASYN. The confusion matrix in Fig. 12 also shows that it also struggles differentiating True Positive and True Negative, similar to both of the model using ADASYN and SMOTE. The inferior performance of the undersampling methods compared to their oversampling counterparts may stem from the inherent reduction in the dataset's size. By discarding instances from the majority class during under-sampling, crucial information might be lost, hindering the model's ability to discern subtle patterns within the data.

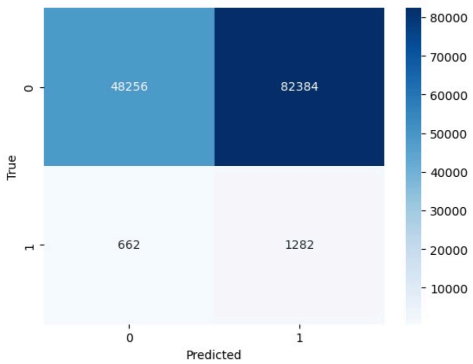

Fig. 11. Confusion Matrix of Rodent Tuber's Balance Data with Random Under Sampling

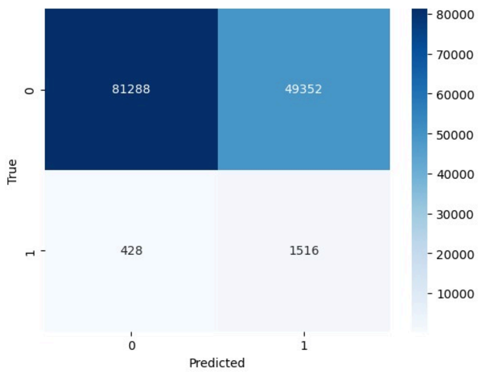

Fig. 12. Confusion Matrix of Rodent Tuber's Balance Data with Near Miss

Table 1 shows the evaluation metrics for all of the models trained using imbalanced and balanced data using various balancing algorithms. We used default parameters on each balancing algorithms and ANN.

In the imbalanced dataset, it appears that the accuracy is 99%, but upon closer inspection, it turns out to be only for class 0. Class 1 has a value of 0 across the board. This indicates overfitting. The issues encountered by the ANN models in effectively handling the Rodent Tuber dataset are complex, influenced by the inherent bias in artificial neural networks and the synthetic nature of the balancing algorithms used. Bias in ANNs from their ability to capture and learn patterns from training data, which frequently leads to a preference for the majority class, which in this case consists of normal compounds. In the context of anti-cancer compound prediction, this inherent bias can manifest as a tendency to predict instances as normal compounds, resulting in a significant number of False Negatives.

Table 1. Each balancing strategies evaluation metric score

Balancing strategies	Class	Evaluation Metric			
		Precision	Recall	F1-Score	Accuracy
None (Imbalanced)	0	0.99	1.0	0.99	0.99
	1	0.00	0.00	0.00	
SMOTE	0	1.00	0.55	0.71	0.56
	1	0.03	0.96	0.06	
ADASYN	0	1.00	0.64	0.78	0.65
	1	0.03	0.84	0.07	
RUS	0	0.99	0.37	0.54	0.37
	1	0.02	0.66	0.03	
NearMiss	0	0.99	0.62	0.77	0.62
	1	0.03	0.78	0.06	

Moreover, the synthetic nature of balancing algorithms, such as SMOTE and ADASYN, adds additional layers of complexity. While these algorithms are designed to address class imbalance by generating synthetic instances for the minority class, their effectiveness can be hindered by the intricacies of the dataset. The synthetic instances created may not perfectly encapsulate the subtle patterns and characteristics of genuine anti-cancer compounds, posing difficulties in accurate classification. As a result, ANN models trained on datasets balanced with these techniques may struggle to generalize effectively to unseen data, especially given the nuanced nature of the anti-cancer compound prediction task.

Furthermore, the challenges posed by bias and the synthetic nature of balancing algorithms underscore the need for a meticulous examination of model predictions. The observed struggles in differentiating between True Positive and True Negative instances emphasize the complexities associated with discerning between normal and anti-cancer compounds.

ADASYN achieved high true positives but higher false negatives, while producing the second best true negative after SMOTE. Random sampling was the least effective, trailing behind NearMiss.

Balancing algorithms like SMOTE and ADASYN, designed to address class imbalance, can be complicated due to the intricacies of the dataset. These synthetic instances may not accurately represent genuine anti-cancer compounds, making it difficult for ANN models to generalize effectively.

In our exploration of data balance effects on Neural Network predictions for Rodent Tuber, the challenges posed by an imbalanced dataset were evident. Despite employing SMOTE, RUS, ADASYN, and NearMiss, F1 scores revealed performance ranging from 0.03 to 0.78. Notably, ADASYN demonstrated an impressive ability for true positives at the expense of increased false negatives, while producing the second best true negative after SMOTE. ADASYN, while generating more true positives and fewer false negatives than SMOTE, exhibited a trade-off with higher false positives. Random sampling emerged as the least effective, with NearMiss trailing behind ADASYN in performance.

The synthetic nature of balancing techniques, combined with ANN biases, emerged as critical factors in shaping the predictive outcomes. This highlights the intricate interplay between data balance and the nuanced task of tuber prediction.

5 Conclusion

This study shows that data balancing and the use of ANN in the classification of Rodent Tuber have not produced good results yet. Although it is shown that balancing data using ADASYN has better results when used by ANN. Therefore, continuous studies are needed in the selection of parameters and model architecture for optimal pharmacognosy classification. Also, contribute to the discussion of optimizing Neural Network predictions in pharmacognosy contexts, laying the groundwork for future research in imbalanced data scenarios.

References

1. 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
2. Z. Muflikhati, N. F. Sianipar, M. R. A. A. Syamsunarno, and Anas, “Understanding of genes encoding bioactive compounds from potential medicinal plants in Indonesia as cancer cell inhibitors,” Biodiversitas, vol. 24, no. 8, pp. 4645–4660, 2023, doi: 10.13057/biodiv/d240846.
3. N. F. Sianipar, R. Purnamaningsih, I. Darwati, D. Laurent, and . C., “The Effects of Gamma Irradiation and Somaclonal Variation on Morphology Variations of Mutant Rodent Tuber (Typhonium flagelliforme Lodd.) Lines,” KnE Life Sciences, vol. 2, no. 1, p. 637, Sep. 2015, doi: 10.18502/kls.v2i1.236.
4. N. F. Sianipar and R. Purnamaningsih, TROPICAL AGRICULTURAL SCIENCE, "Enhancement of the Contents of Anticancer Bioactive Compounds in Mutant Clones of Rodent Tuber (Typhonium flagelliforme Lodd.) based on GC-MS Analysis” Pertanika J. Trop. Agric. Sci, vol. 41, no. 1, pp. 305–320, 2018, [Online]. Available: http://www.pertanika.upm.edu.my/
5. N. F. Sianipar, D. L. Gumanti, Rosaria, and Vidianty, TROPICAL AGRICULTURAL SCIENCE, "Analysis of Gamma Irradiated-Third Generation Mutants of Rodent Tuber (Typhonium flagelliforme Lodd.) Based on Morphology, RAPD, and GC-MS Markers,” Pertanika J. Trop. Agric. Sci, vol. 40, no. 1, pp. 185–202, 2017, [Online]. Available: http://www.pertanika.upm.edu.my/
6. N. F. Sianipar, D. Laurent, R. Purnamaningsih, and I. Darwati, “Genetic Variation of the First Generation of Rodent Tuber (Typhonium flagelliforme Lodd.) Mutants Based on RAPD Molecular Markers,” Hayati, vol. 22, no. 2, pp. 98–104, Apr. 2015, doi: 10.4308/hjb.22.2.98.

7. N. F. Sianipar, R. Purnamaningsih, and Rosaria, “Bioactive compounds of fourth generation gamma-irradiated Typhoniumflagelliforme Lodd. mutants based on gas chromatography-mass spectrometry,” in IOP Conference Series: Earth and Environmental Science, Institute of Physics Publishing, Sep. 2016. doi: 10.1088/1755-1315/41/1/012025.
8. I. Binanto, H. L. H. S. Warnars, N. F. Sianipar, and W. Budiharto, “Anticancer Compound Identification Model of Rodent Tuber’s Liquid Chromatography-Mass Spectrometry Data,” ICIC Express Letters, vol. 16, no. 1, pp. 9–16, Jan. 2022, doi: 10.24507/icicel.16.01.9.
9. G. A. Batan, M. J. Keytimu, F. L. Katumbo, I. Binanto, and N. F. Sianipar, “Penerapan Metode Random Forest, Gaussian NB, Dan KNN Terhadap Data Unbalance dan Data Balance Menggunakan Random Over Sampling Untuk Klasifikasi Senyawa Keladi Tikus,” Prosiding Seminar Nasional Teknik Elektro, Informatika & Sistem Informasi (SINTaKS), vol. 2, no. 1, pp. 1–5, Sep. 2023, doi: 10.35842/SINTAKS.V2I1.26.
10. V. Meriana et al., “Perbandingan Algoritma Klasifikasi KNN, Gaussian Naive Bayes, Dan Random Forest Pada Dataset LCMS Tanaman Keladi Tikus Yang Diseimbangkan Dengan Metode Synthetic Minority Over-Sampling Technique,” Prosiding Seminar Nasional Teknik Elektro, Informatika & Sistem Informasi (SINTaKS), vol. 2, no. 1, pp. 6–10, Sep. 2023, doi: 10.35842/SINTAKS.V2I1.27.
11. N. Made et al., “Analisis Perbandingan Algoritma KNN, Gaussian Naive Bayes, Random Forest untuk Data Tisak Seimbang dan Data yang Diseimbangkan dengan Metode Tomek Link Undersampling pada Dataset LCMS Tanaman Keladi Tikus,” Prosiding Seminar Sains Nasional dan Teknologi, vol. 13, no. 1, pp. 156–160, Nov. 2023, doi: 10.36499/PSNST.V13I1.9002.
12. F. Dea, M. N. Primadani, T. W. Kartikasari, I. Binanto, and N. F. Sianipar, Prosiding Seminar Nasional Sains dan Teknologi, "Klasifikasi Senyawa Keladi Tikus Menggunakan Algoritma KNN, Gaussian Naive Bayes dengan Menerapkan Imbalance Data Borderline SMOTE”, doi: 10.36499/psnst.v13i1.9005.
13. M. M. Mijwil and M. M. Mijwel, “Artificial Neural Networks Advantages and Disadvantages,” 2018. [Online]. Available: https://www.researchgate.net/publication/323665827
14. M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, “Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial,” IEEE Communications Surveys and Tutorials, vol. 21, no. 4, pp. 3039–3071, Oct. 2019, doi: 10.1109/COMST.2019.2926625.
15. R. Dastres and M. Soori, “Artificial Neural Network Systems,” International Journal of Imaging and Robotics (IJIR), vol. 2021, no. 2, pp. 13–25, 2021, Accessed: Nov. 28, 2023. [Online]. Available: www.ceserp.com/cp-jour
16. R. Hidayat, S. Koesnaryo, B. Dwinagara, S. Saptono, and E. Winarno, “Analisis Kestabilan Lereng Menggunakan Artificial Neural Network : Sebuah Tinjauan Pustaka,” pp. 209–215, 2022, [Online]. Available: http://journal.itny.ac.id/index.php/ReTII
17. M. Anousouya Devi, S. Ravi, J. Vaishnavi, S. Punitha, “Classification of Cervical Cancer Using Artificial Neural Networks”, Procedia Computer Science, 2016, Volume 89, Pages 465-472, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.06.105.
18. Vawda, Mohamed Ismail, Romano Lottering, Onisimo Mutanga, Kabir Peerbhay, and Mbulisi Sibanda. 2024. "Comparing the Utility of Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) on Sentinel-2 MSI to Estimate Dry Season Aboveground Grass Biomass" *Sustainability* 16, no. 3: 1051. https://doi.org/10.3390/su16031051.
19. Shahid N, Rappon T, Berta W. “Applications of artificial neural networks in health care organizational decision-making: A scoping review”. PLoS One. 2019 Feb

19;14(2):e0212356. doi: 10.1371/journal.pone.0212356. PMID: 30779785; PMCID: PMC6380578.
20. P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," IOP Conf Ser Mater Sci Eng, vol. 1099, no. 1, p. 012077, Mar. 2021, doi: 10.1088/1757-899x/1099/1/012077.