Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

Abu Kwaik, K., Chatzikyriakidis, S., & Dobnik, S. (2022). Pre-trained models or feature engineering: The case of dialectal Arabic. Dalam H. Al-Khalifa, T. Elsayed, H. Mubarak, A. Al-Thubaity, W. Magdy, & K. Darwish (Ed.), *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection* (hlm. 41–50). European Language Resources Association. https://aclanthology.org/2022.osact-1.5/

Alqulaity, E. Y., Yafooz, W. M. S., Alourani, A., & Jaradat, A. (2024). Arabic dialect identification in social media: A comparative study of deep learning and transformer approaches. *Intelligent Automation & Soft Computing*, *39*(5), 907–928. https://doi.org/10.32604/iasc.2024.055470

Alsudais, A., Alotaibi, W., & Alomary, F. (2022). Similarities between Arabic dialects: Investigating geographical proximity. *Information Processing & Management*, *59*(1), 102770. https://doi.org/10.1016/j.ipm.2021.102770

Alsuwaylimi, A. A. (2024). Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM. *Heliyon*, *10*(17), e36280. https://doi.org/10.1016/j.heliyon.2024.e36280

AlYami, R., & AlZaidy, R. (2020). Arabic Dialect Identification in Social Media. *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 1–2. https://doi.org/10.1109/ICCAIS48893.2020.9096847

Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2022). RN-SMOTE: Reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 5059–5074. https://doi.org/10.1016/j.jksuci.2022.06.005

Ariasa, K., Gunadi, I. G. A., & Candiasa, I. M. (2020). Optimasi algoritma klaster dinamis pada K-means dalam pengelompokan kinerja akademik mahasiswa (studi kasus: Universitas Pendidikan Ganesha). *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, *9*(2), 181–192. https://ejournal.undiksha.ac.id/index.php/janapati/article/view/23491

Aruleba, I., & Sun, Y. (2025). Enhanced credit risk prediction using deep learning and SMOTE-ENN resampling. *Machine Learning with Applications*, *21*, 100692. https://doi.org/10.1016/j.mlwa.2025.100692

Baimukan, N., Bouamor, H., & Habash, N. (2022). Hierarchical aggregation of dialectal data for Arabic dialect identification. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 4586–4596. https://aclanthology.org/2022.lrec-1.489/

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bishop, C. M. (2006). *Pattern recognition and machine learning* (M. Jordan, J. Kleinberg, & B. Schölkopf, Ed.). Springer: New York. https://link.springer.com/book/9780387310732

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1), 106. https://doi.org/10.1186/1471-2105-14-106

Buckwalter, T. (2002). *Arabic morphological analyzer version 1.0*.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). *Safe-Level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem* (hlm. 475–482). https://doi.org/10.1007/978-3-642-01307-2_43

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, B., Xia, S., Chen, Z., Wang, B., & Wang, G. (2021). RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise. *Information Sciences*, *553*, 397–428. https://doi.org/10.1016/j.ins.2020.10.013

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2 ed.). Routledge: New York. https://doi.org/10.4324/9780203771587

Cover, T. M., & Thomas, J. A. (2005). *Elements of information theory*. John Wiley & Sons: New York. https://doi.org/10.1002/047174882X

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., Bouzoubaa, K., Cavalli-Sforza, V., El-Beltagy, S. R., El-Hajj, W., Jarrar, M., & Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, *64*(4), 72–81. https://doi.org/10.1145/3447735

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

Elnagar, A., Yagi, S. M., Nassif, A. B., Shahin, I., & Salloum, S. A. (2021). Systematic Literature Review of Dialectal Arabic: Identification and Detection. *IEEE Access*, *9*(1), 31010–31042. https://doi.org/10.1109/ACCESS.2021.3059504

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Dalam D. ~W. Pfitzner & J. ~K. Salmon (Ed.), *Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Proceedings of a conference held August 2-4* (hlm. 226–331). https://dl.acm.org/doi/10.5555/3001460.3001507

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Fränti, P., Cariou, C., & Zhao, Q. (2025). Cluster overlap as objective function. *Computers, Materials & Continua*, *85*(3), 4687–4704. https://doi.org/10.32604/cmc.2025.066534

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Gunadi, A., Dewi Oktofa Rahmawati, & Nurfa Risha. (2025). Evaluation of the performance of the SMOTE, SMOTE ENN, and Borderline SMOTE resampling methods based on the number of outlier data with Z score. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *16*(2), 128–140. https://doi.org/10.24843/LKJITI.2025.v16.i02.p05

Hairani, H., Widiyaningtyas, T., & Dwi Prasetya, D. (2024). Addressing class imbalance of health data: A systematic literature review on modified synthetic minority oversampling technique (SMOTE) strategies. *JOIV : International Journal on Informatics Visualization*, *8*(3), 1310–1318. https://doi.org/10.62527/joiv.8.3.2283

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Dalam *Advances in Intelligent Computing* (hlm. 878–887). Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3 ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Hasan, Y., Amerehi, F., Healy, P., & Ryan, C. (2023). Stem rebalance: A novel approach for tackling imbalanced datasets using SMOTE, edited nearest neighbour, and mixup. *2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 3–9. https://doi.org/10.1109/ICCP60212.2023.10398660

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, H., & Ma, Y. (Ed.). (2013). *Imbalanced learning: Foundations, algorithms, and applications*. John Wiley & Sons. https://doi.org/http://dx.doi.org/10.1002/9781118646106

Hemmatian, J., Hajizadeh, R., & Nazari, F. (2025). Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE. *PLOS ONE*, *20*(2), 1–20. https://doi.org/10.1371/journal.pone.0317396

Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A study on the performance of resampling algorithms for addressing class imbalance in regression models. *Algorithms*, *18*(1), 1–16. https://doi.org/10.3390/a18010037

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. Dalam N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, & S. Touileb (Ed.), *Proceedings of the*

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Sixth Arabic Natural Language Processing Workshop (hlm. 92–104). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.10/

Jeong, Y., Lee, K., Park, Y. W., & Han, S. (2024). A systematic approach for learning imbalanced data: Enhancing zero-inflated models through boosting. *Machine Learning*, *113*(10), 8233–8299. https://doi.org/10.1007/s10994-024-06558-3

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. Dalam C. Nédellec & C. Rouveirol (Ed.), *Machine Learning: ECML-98* (hlm. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Jose, B. A., & Das, P. (2022). A multi-objective approach for inter-cluster and intra-cluster distance analysis for numeric data. Dalam R. Kumar, C. W. Ahn, T. K. Sharma, O. P. Verma, & A. Agarwal (Ed.), *Soft Computing: Theories and Applications* (hlm. 319–332). Springer Nature Singapore. https://doi.org/https://doi.org/10.1007/978-981-19-0707-4_30

Jurafsky, D., & Martin, J. H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3 ed.). Stanford: Stanford University Press. https://web.stanford.edu/~jurafsky/slp3

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Dalam I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Ed.), *Advances in Neural Information Processing Systems* (hlm. 3149–3157). https://dl.acm.org/doi/10.5555/3294996.3295074

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 1–16. https://doi.org/10.1016/j.asoc.2019.105662

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kwaik, K. A., Saad, M., Chatzikyriakidis, S., & Dobnik, S. (2018). Shami: A corpus of Levantine Arabic dialects. Dalam N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo,

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Ed.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1576/

Le, T. (2022). A comprehensive survey of imbalanced learning methods for bankruptcy prediction. *IET Communications*, *16*(5), 433–441. https://doi.org/10.1049/cmu2.12268

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Li, Y., Yang, Y., Song, P., Duan, L., & Ren, R. (2025). An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. *Scientific Reports*, *15*(1), 23521. https://doi.org/10.1038/s41598-025-09506-w

Lichouri, M., & Abbas, M. (2020). Simple vs Oversampling-based Classification Methods for Fine Grained Arabic Dialect Identification in Twitter. Dalam I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh, & W. Zaghouani (Ed.), *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (hlm. 250–256). Association for Computational Linguistics. https://aclanthology.org/2020.wanlp-1.24/

Lu, S., & Ye, J. (2024). Imbalanced data classification scheme based on G-SMOTE. *Procedia Computer Science*, *247*, 1295–1303. https://doi.org/10.1016/j.procs.2024.10.155

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297. http://projecteuclid.org/euclid.bsmsp/1200512992

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Matharaarachchi, S., Domaratzki, M., & Muthukumarana, S. (2024). Enhancing SMOTE for imbalanced data with abnormal minority instances. *Machine Learning with Applications*, *18*, 100597. https://doi.org/10.1016/j.mlwa.2024.100597

Matrane, Y., Benabbou, F., & Sael, N. (2023). A systematic literature review of Arabic dialect sentiment analysis. *Journal of King Saud University - Computer*

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

and Information Sciences, *35*(6), 101570. https://doi.org/10.1016/j.jksuci.2023.101570

Mohamed, E. (2010). *Orthographic enrichment for arabic grammatical analysis* [Indiana University]. https://dl.acm.org/doi/10.5555/2019865

Mohammed, A., Jiangbin, Z., & Murtadha, A. (2023). A three-stage neural model for Arabic Dialect Identification. *Computer Speech & Language*, *80*(2), 101488. https://doi.org/10.1016/j.csl.2023.101488

Moniz, N., & Cerqueira, V. (2021). Automated imbalanced classification via meta-learning. *Expert Systems with Applications*, *178*, 1–50. https://doi.org/10.1016/j.eswa.2021.115011

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press: Cambridge. http://probml.github.io/book1

Nayel, H., Hassan, A., Sobhi, M., & El-Sawy, A. (2021). Machine learning-based approach for Arabic dialect identification. Dalam N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, & S. Touileb (Ed.), *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (hlm. 287–290). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.34

Nouas, S., Oukid, L., & Boumahdi, F. (2025). Enhancing imbalanced text classification: an overlap-based refinement approach. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2025.03.001

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2009). Data mining with imbalanced class distributions: Concepts and methods. *4th Indian International Conference on Artificial Intelligence (IICAI-09)*, 349–376. https://api.semanticscholar.org/CorpusID:16651273

Qin, X., Zhang, S., Dong, X., Shi, H., & Yuan, L. (2024). Classification of high-dimensional imbalanced biomedical data based on spectral clustering SMOTE and marine predators algorithm. *Journal of Intelligent & Fuzzy Systems*, *46*(4), 8709–8728. https://doi.org/10.3233/JIFS-237538

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(1). https://dl.acm.org/doi/10.5555/3455716.3455856

Ridwan, R., Heni Hermaliani, E., & Ernawati, M. (2024). Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian. *Computer Science (CO-SCIENCE)*, *4*(1), 80–88. https://doi.org/10.31294/coscience.v4i1.2990

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. SAGE Publications, Inc.: Thousand Oaks. https://doi.org/10.4135/9781412984997

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saleh, H., AlMohimeed, A., Hassan, R., Ibrahim, M. M., Alsamhi, S. H., Hassan, M. R., & Mostafa, S. (2025). Advancing Arabic dialect detection with hybrid stacked transformer models. *Frontiers in Human Neuroscience*, *19*, 1–15. https://doi.org/10.3389/fnhum.2025.1498297

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sculley, D. (2010). Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web*, 1177–1178. https://doi.org/10.1145/1772690.1772862

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences: Madison. http://digital.library.wisc.edu/1793/60660

Shafeeq, A., & Hareesha. (2012). Dynamic clustering of data with modified K-means algorithm. *International Conference on Information and Computer Networks (ICICN 2012)*, 221–225. https://doi.org/10.13140/2.1.4972.3840

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1), 70. https://doi.org/10.1186/s40537-020-00349-y

UNIVERSITAS
GADJAH MADA
**Model Oversampling Hibrida Berbasis Klaster dan Densitas untuk Penanganan Data Tidak Seimbang pada
Kasus Klasifikasi Dialek Arab**
Maulana Ihsan Ahmad, Arif Nurwidyantoro, S. Kom., M.Cs., Ph.D.; Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Tao, Y., Zhang, Y., & Jiang, B. (2020). DBCSMOTE: a clustering-based oversampling technique for data-imbalanced warfarin dose prediction. *BMC Medical Genomics*, *13*(S10), 152. https://doi.org/10.1186/s12920-020-00781-2

Wang, S., & Manning, C. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. Dalam H. Li, C. Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Ed.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (hlm. 90–94). Association for Computational Linguistics. https://aclanthology.org/P12-2018/

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., & Han, X. (2021). A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data. *Information Sciences*, *572*, 574–589. https://doi.org/10.1016/j.ins.2021.02.056

Yadav, A. K., Alam, B., & Pal, O. (2025). Evaluating SMOTE-ENN impact on EEG signal seizure detection and classification. *Procedia Computer Science*, *258*, 2920–2929. https://doi.org/10.1016/j.procs.2025.04.552

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171–202. https://doi.org/10.1162/COLI_a_00169

Zhang, B., Liang, M., Zhou, Y., Ji, B., Han, M., Li, H., Lang, X., Song, Y., Gao, R., Zhou, Z., & Qiao, X. (2026). A refined SMOTE-ENN optimization method based on machine learning for heart rate variability data classification. *Frontiers in Digital Health*, *8*, 1–12. https://doi.org/10.3389/fdgth.2026.1749570

Zhang, Y., Zhu, C., & Wang, Q. (2020). LightGBM-based model for metro passenger volume forecasting. *IET Intelligent Transport Systems*, *14*(13), 1815–1823. https://doi.org/10.1049/iet-its.2020.0396