AN IMPROVED SYNTHETIC DATA GENERATION FOR RICE YIELD PREDICTION MODEL

MUHAMMAD KHALIFA UMANA
830636

MASTER OF SCIENCE
SCHOOL OF COMPUTING
UNIVERSITI UTARA MALAYSIA
2025

AN IMPROVED SYNTHETIC DATA GENERATION FOR RICE YIELD PREDICTION MODEL

BY
MUHAMMAD KHALIFA UMANA

Thesis Submitted to
School of Computing,
Universiti Utara Malaysia,
in Fulfilment of the Requirement for the Degree of Master of Science

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

UMANA MUHAMMAD KHALIFA

calon untuk Ijazah **MSc**
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"AN IMPROVED SYNTHETIC DATA GENERATION FOR RICE YIELD PREDICTION MODEL"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***15 Mei 2025.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
15 May 2025.

Pengerusi Viva: *(Chairman for VIVA)*	**Assoc. Prof. Noraziah Che Pa**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Dr. Samry @ Mohd Shamrie Sainin**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Ts. Wan Hussain Wan Ishak**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Mohamad Farhan Mohamad Mohsin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Ts. Dr. Mohamad Ghozali Hassan**	Tandatangan *(Signature)*

Tarikh:
(Date) ***15 May 2025***

Permission to Use

In presenting this thesis in fulfilment of the requirements for a Post Graduate degree from the Universiti Utara Malaysia (UUM), I agree that the library of this university may make it freely available for inspection. I further agree that permission for copying this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s) or in their absence, by the Director of Postgraduate Studies Unit, School of Computing where I did my thesis. It is understood that any copying or publication or use of this thesis or parts of it for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the UUM in any scholarly use which may be made of any material in my thesis.

Request for permission to copy or to make other use of materials in this thesis in whole or in part should be addressed to:

Director of Postgraduate Studies Unit, School of Computing
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

Abstrak

Beras memainkan peranan penting dalam memastikan keselamatan makanan dan kestabilan ekonomi di Malaysia. Walau bagaimanapun, ramalan hasil padi menghadapi cabaran kerana data pertanian yang terhad dan tidak mencukupi. Walaupun penjanaan data sintetik boleh menjadi penyelesaian seperti varian rangkaian lawan generatif (GAN), ia masih menghadapi masalah jika ia dilatih oleh data terhad. Untuk memperkayakan data dan mengukuhkan keupayaan GAN dalam menjana data, melengkapkan latihan model generatif oleh penjanaan data berasaskan statistik, seperti SMOTE, boleh disepadukan. Oleh itu, kajian ini bertujuan untuk membangunkan model ramalan hasil padi yang tepat dengan mencadangkan Penjanaan Dipertingkat SMOTE-CTABGAN (SCEG), integrasi baru SMOTE-ENC dan CTAB-GAN, untuk menangani had data. Objektif termasuk mereka bentuk kaedah penjanaan data sintetik baharu, melaksanakan prosedur pemprosesan pasca penjanaan untuk memastikan pembolehubah kekal praktikal, kekal di antara julat nilai yang berkaitan dengan pertanian, dan menilai prestasi model ramalan yang dipertingkat menggunakan data sintetik yang dijana. Menggunakan data hasil padi dari Jabatan Perangkaan Malaysia (2010-2021) dan data iklim dari Jabatan Meteorologi Malaysia (2010-2021), kajian ini mensintesis dan mempertingkat set data dengan kaedah yang dicadangkan. Penemuan menunjukkan keupayaan SCEG untuk menjana set data sintetik berkualiti tinggi yang berkait rapat dengan set data dunia sebenar. contohnya, SCEG mampu mengurangkan MSE sehingga 98.44% dalam regressor Pokok Keputusan. Tambahan pula, kaedah pasca pemprosesan seperti Pembolehubah Had dan Julat Antara Kuartil meningkatkan prestasi SCEG, dengan Had dan IQR memberikan peningkatan terbaik: pengurangan 60.20% dalam MAE, 86.65% dalam MSE dan 54.86% pengurangan dalam RMSE, serta 8.15. % peningkatan dalam R-kuadrat. Penyelidikan membuktikan bahawa SCEG boleh mencipta set data sintetik yang kukuh dan meningkatkan data dunia sebenar, dengan model yang lebih tepat dan boleh dipercayai untuk meramalkan pengeluaran beras. Perkaitan penyelidikan terletak pada kapasitinya untuk mengubah unjuran pertanian, menghasilkan pertimbangan yang lebih termaklum yang boleh meningkatkan dengan ketara usaha keselamatan makanan negara.

Kata kunci: Data Sintetik, Model Ramalan Pertanian, Rangkaian Adversarial Generatif.

Abstract

Rice plays a crucial role in ensuring food security and economic stability in Malaysia. However, rice yield prediction face challenges due to limited and inadequate agricultural data. Even though synthetic data generation can be a solution such as variant of generative adversarial network (GAN), it still has a problem if it is trained by limited data. To enrich the data and strengthen GAN ability in generating data, complementing the generative model training by statistical-based data generation, such as SMOTE, can be integrated. Therefore, this study aims to develop an improved rice yield prediction model by proposing the SMOTE-CTABGAN Enhanced Generation (SCEG), a novel integration of SMOTE-ENC and CTAB-GAN, to address data limitation. This study includes designing a new synthetic data generation method, implementing a post-generation processing procedure to ensure variable remain practical, stays between agriculturally relevant range of value, and evaluating the improved prediction models performance using the generated synthetic data. Using rice yield data from Department of Statistic Malaysia (2010-2021) and climate data from Malaysia Meteorological Department (2010-2021), this study synthesized and enhance the dataset with proposed method. The findings demonstrate SCEG's ability to generate a high-quality synthetic dataset that is closely related to real-world datasets. for example, SCEG able to reduce MSE by up to 98.44% in Decision Tree regressors. Furthermore, post-processing methods such as Variable-base Value Limitation and Interquartile Range improve SCEG performance, with Limitation and IQR delivering the best improvements: a 60.20% reduction in MAE, 86.65% in MSE, and 54.86% reduction in RMSE, as well as an 8.15% increase in R-squared. The research proves that SCEG can create a strong synthetic dataset and enhance real-world data, with a more precise and dependable model for predicting rice output. The research's relevance rests in its capacity to transform agricultural projections, resulting in more informed judgments that can significantly enhance the nation's food security endeavors.

Keywords: Synthetic Data, Agricultural Preediction Model, Generative Adversarial Network

Acknowledgement

In the name of Allah, the Beneficent, the Merciful. Alhamdulillah, I am eternally grateful to Almighty Allah (SWT) for His countless blessings, strength, and guidance throughout this journey, allowing me to finally complete this thesis.

My deepest gratitude and appreciation go to my esteemed supervisors, Dr. Mohamad Farhan Mohamad Mohsin and Associate Prof. Ts. Dr. Mohamad Ghozali Bin Hassan, for their invaluable guidance, encouragement, and unwavering support. Their expertise and constructive feedback have been instrumental in shaping this work.

I would also like to extend my heartfelt thanks to Universiti Utara Malaysia (UUM) for providing a conducive learning environment, excellent facilities, and all the necessary resources to support my study.

To my family—my beautiful adorable wife, loving caring parents, supportive siblings, and caring relatives—your continuous prayers, patience, and unconditional support have been my greatest motivation. I dedicate this achievement to you all.

Lastly, I express my gratitude to my friends and peers for their encouragement and companionship throughout this academic journey. Thank you all for being part of this success.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	-	Artificial Neural Network
CTAB-GAN	-	Conditional Table - Generative Adversarial Networks
GAN	-	Generative Adversarial Networks
SMOTE-ENC		Synthetic Minority Over-sampling Technique - Encoded Nominal and Continuous
MAE	-	Mean absolute error
MSE	-	Mean Squared Error
RMSE	-	Rooted Mean Squared Error
LR	-	Linear Regression
DTR	-	Decision Tree Regression

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Agriculture plays a vital role in Malaysia's economy and society, providing income and food security for its population (B. T. Tan et al., 2021). Among all crops, rice holds strategic importance as a staple food and a focus of agricultural policy. Sustaining rice productivity is critical for economic resilience and national food availability.

In line with this, accurate rice yield prediction has become an essential tool to support Malaysia's national food security efforts (Fatah, 2017; Vaghefi et al., 2013), especially under frameworks like the Twelfth Malaysian Plan (2021–2025) and the National Agro-Food Policy 2.0 (2021–2030) (Nazibudin et al., 2025; Sayed et al., 2025).. Reliable prediction systems help balance supply and demand, manage resources efficiently, and reduce the risks of shortages and price volatility.

Machine learning (ML) models are increasingly adopted to address these challenges in food production systems (Sarr & Sultan, 2023; Torsoni et al., 2023). In agriculture, predictive analytics enables more informed decision-making, allowing for better cultivation strategies and distribution planning.

A visual example of the rice production trend is presented in Figure 1.1, which shows the annual paddy output in Malaysia from 1980 to 2022 from research conducted by Mohd Shafri, et al (2023). This study applied several predictive models, such as Multivariate Adaptive Regression Splines (MARS), Multiple Linear Regression (MLR), and Support Vector Regression (SVR). The aim was to assess the relative efficacy of each model in capturing yield trends. The resultant projections offer

REFERENCES

Abdel-Basset, M., El-Shahat, D., Jameel, M., & Abouhawwash, M. (2023). Exponential distribution optimizer (EDO): a novel math-inspired algorithm for global optimization and engineering problems. *Artificial Intelligence Review*, *56*(9), 9329–9400. https://doi.org/10.1007/s10462-023-10403-9

Abdullah, B. U. D., Khanday, S. A., Islam, N. U., Lata, S., Fatima, H., & Nengroo, S. H. (2024). Comparative Analysis Using Multiple Regression Models for Forecasting Photovoltaic Power Generation. *Energies*, *17*(7), 1564. https://doi.org/10.3390/en17071564

Ahmad, M., Farooq, M. A., Zunnurain Hussain, M., Hasan, M. Z., Mustafa, M., Khalid, A., Awan, R., Hussain, U., Khan, Z. A., & Javaid, A. (2024). Car Price Prediction using Machine Learning. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–5. https://doi.org/10.1109/I2CT61223.2024.10544124

Ajibola-James, O., & Okeke, F. I. (2024). *An approach for good modelling and forecasting of sea surface salinity in a coastal zone using machine learning LASSO regression models built with sparse satellite time-series datasets*. https://doi.org/10.21203/rs.3.rs-4016353/v1

Alharbi, A. H., Khafaga, D. S., Zaki, A. M., El-Kenawy, E.-S. M., Ibrahim, A., Abdelhamid, A. A., Eid, M. M., El-Said, M., Khodadadi, N., Abualigah, L., & Saeed, M. A. (2024). Forecasting of energy efficiency in buildings using multilayer perceptron regressor with waterwheel plant algorithm hyperparameter. *Frontiers in Energy Research*, *12*. https://doi.org/10.3389/fenrg.2024.1393794

Almazroi, A. A., & Ayub, N. (2023a). Online Payment Fraud Detection Model Using Machine Learning Techniques. *IEEE Access*, *11*, 137188–137203. https://doi.org/10.1109/ACCESS.2023.3339226

Almazroi, A. A., & Ayub, N. (2023b). Online Payment Fraud Detection Model Using Machine Learning Techniques. *IEEE Access*, *11*, 137188–137203. https://doi.org/10.1109/ACCESS.2023.3339226

Alshantti, A., Varagnolo, D., Rasheed, A., Rahmati, A., & Westad, F. (2024). CasTGAN: Cascaded Generative Adversarial Network for Realistic Tabular Data Synthesis. *IEEE Access*, *12*, 13213–13232. https://doi.org/10.1109/ACCESS.2024.3356913

Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J., & Rathnayake, U. (2020). Artificial Neural Network to Estimate the Paddy Yield Prediction Using Climatic Data. *Mathematical Problems in Engineering*, *2020*, 1–11. https://doi.org/10.1155/2020/8627824

An, S., & Jeon, J.-J. (2023). Distributional Learning of Variational AutoEncoder: Application to Synthetic Data Generation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 57825–57851). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/b456a00e145ad56f6f251f79f8c8a7de-Paper-Conference.pdf

Anuradha, D., Kuchipudi, R., Ashreetha, B., Ramesh, J. V. N., & Rami, A. (2024). Enhancing Agricultural Yield Forecasting with Deep Convolutional Generative Adversarial Networks and Satellite Data. *International Journal of Advanced*

Computer Science and Applications, *15*(2). https://doi.org/10.14569/IJACSA.2024.0150269

Asiedu, S. T., Nyarko, F. K. A., Boahen, S., Effah, F. B., & Asaaga, B. A. (2024). Machine learning forecasting of solar PV production using single and hybrid models over different time horizons. *Heliyon*, *10*(7), e28898. https://doi.org/10.1016/j.heliyon.2024.e28898

B M, Dr. S., N K, Dr. C., T, Dr. P., & R, Dr. R. (2022). Rice and Wheat Yield Prediction in India Using Decision Tree and Random Forest. *Computational Intelligence and Machine Learning*, *3*(2), 1–8. https://doi.org/10.36647/CIML/03.02.A001

Baheti, A. S., Sawarkar, A. D., Shaikh, U. A., & Shrimankar, D. D. (2024). Alcohol Quality Analysis Using Machine Learning Regression Technique. *Cureus Journals*. https://doi.org/10.7759/s44389-024-00227-1

Banerjee, S., & Bishop, T. R. P. (2022). dsSynthetic: synthetic data generation for the DataSHIELD federated analysis system. *BMC Research Notes*, *15*(1), 230. https://doi.org/10.1186/s13104-022-06111-2

Bansal, Y., Lillis, D., & Kechadi, T. (2023). *Winter Wheat Crop Yield Prediction on Multiple Heterogeneous Datasets using Machine Learning*. https://contour.ag-space.com/

Barth, R., IJsselmuiden, J., Hemming, J., & Henten, E. J. Van. (2018). Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset. *Computers and Electronics in Agriculture*, *144*, 284–296. https://doi.org/10.1016/j.compag.2017.12.001

Batool, D., Shahbaz, M., Shahzad Asif, H., Shaukat, K., Alam, T. M., Hameed, I. A., Ramzan, Z., Waheed, A., Aljuaid, H., & Luo, S. (2022). A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. *Plants*, *11*(15). https://doi.org/10.3390/plants11151925

Baudhanwala, D., Mehta, D., & Kumar, V. (2024). Machine learning approaches for improving precipitation forecasting in the Ambica River basin of Navsari District, Gujarat. *Water Practice & Technology*, *19*(4), 1315–1329. https://doi.org/10.2166/wpt.2024.079

Belhiti, B., Milushev, J., Gupta, A., Breedis, J., Dinh, J., Pisel, J., & Pyrcz, M. (2021). *StackGAN: Facial Image Generation Optimizations*.

Berger, V. W., & Zhou, Y. (2014). Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*. Wiley. https://doi.org/10.1002/9781118445112.stat06558

Bermano, A. H., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., Patashnik, O., & Cohen-Or, D. (2022). State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Computer Graphics Forum*, *41*(2), 591–611. https://doi.org/10.1111/cgf.14503

Bharadiya, J. P., Tzenios, N. T., & Reddy, M. (2023). Forecasting of Crop Yield using Remote Sensing Data, Agrarian Factors and Machine Learning Approaches. *Journal of Engineering Research and Reports*, *24*(12), 29–44. https://doi.org/10.9734/jerr/2023/v24i12858

Bilal, M., Alrasheedi, M. A., Aamir, M., Abdullah, S., Norrulashikin, S. M., & Rezaiy, R. (2024). Enhanced forecasting of rice price and production in Malaysia using novel multivariate fuzzy time series models. *Scientific Reports*, *14*(1), 29903. https://doi.org/10.1038/s41598-024-77907-4

Birim, S., Kazancoglu, I., Mangla, S. K., Kahraman, A., & Kazancoglu, Y. (2024). The derived demand for advertising expenses and implications on sustainability: a comparative study using deep learning and traditional machine learning methods. *Annals of Operations Research*, *339*(1–2), 131–161. https://doi.org/10.1007/s10479-021-04429-x

Boomgard-Zagrodnik, J. P., & Brown, D. J. (2022). Machine learning imputation of missing Mesonet temperature observations. *Computers and Electronics in Agriculture*, *192*, 106580. https://doi.org/10.1016/j.compag.2021.106580

Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A., & Zahariadis, T. (2021). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information*, *12*(9), 375. https://doi.org/10.3390/info12090375

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2021). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agricultural and Forest Meteorology*, *297*, 108275. https://doi.org/10.1016/j.agrformet.2020.108275

Charitou, C., Garcez, A. d'Avila, & Dragicevic, S. (2020). Semi-supervised GANs for Fraud Detection. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206844

Chu, Z., & Yu, J. (2020). An end-to-end model for rice yield prediction using deep learning fusion. *Computers and Electronics in Agriculture*, *174*, 105471. https://doi.org/10.1016/j.compag.2020.105471

Coffie, G. H., & Cudjoe, S. K. F. (2024). Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *International Journal of Construction Management*, *24*(16), 1742–1750. https://doi.org/10.1080/15623599.2023.2289754

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, *19*(5), 1181. https://doi.org/10.3390/s19051181

Das, B., Nair, B., Reddy, V. K., & Venkatesh, P. (2018). Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. *International Journal of Biometeorology*, *62*(10), 1809–1822. https://doi.org/10.1007/s00484-018-1583-6

Daub, E. G., Trugman, D. T., & Johnson, P. A. (2015). Statistical tests on clustered global earthquake synthetic data sets. *Journal of Geophysical Research: Solid Earth*, *120*(8), 5693–5716. https://doi.org/10.1002/2014JB011777

De Luna, R. G., Bandong, J. E. O., Garcia, S. A. R., Mirabel, T. L., Panaligan, M. C. L., & Sunga, J. C. N. (2024). Optimizing Rooftop Solar Potential Prediction: A Comprehensive Study Leveraging Machine Learning Techniques. *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, 90–96. https://doi.org/10.1109/ICICoS62600.2024.10636905

Desai, A., Freeman, C., Wang, Z., & Beaver, I. (2021). *TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation*.

DeVore, M. D., & O'Sullivan, J. A. (2001). *Statistical assessment of model fit for synthetic aperture radar data* (E. G. Zelnio, Ed.; pp. 379–388). https://doi.org/10.1117/12.438231

DeVries, T., & Taylor, G. W. (2017). *Dataset Augmentation in Feature Space*.

Dhillon, M. S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I., & Ullmann, T. (2023). Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. *Frontiers in Remote Sensing*, *3*. https://doi.org/10.3389/frsen.2022.1010978

Diamantopoulou, M. J., & Papamichail, D. M. (2024). Performance Evaluation of Regression-Based Machine Learning Models for Modeling Reference Evapotranspiration with Temperature Data. *Hydrology*, *11*(7), 89. https://doi.org/10.3390/hydrology11070089

Dorairaj, D., & Govender, N. T. (2023). Rice and paddy industry in Malaysia: governance and policies, research trends, technology adoption and resilience. *Frontiers in Sustainable Food Systems*, *7*. https://doi.org/10.3389/fsufs.2023.1093605

Drechsler, J. (2010). *Using Support Vector Machines for Generating Synthetic Datasets* (pp. 148–161). https://doi.org/10.1007/978-3-642-15838-4_14

Ebrahimianghazani, F. (2023). *DP-Select: Improving Utility and Privacy in Tabular Data Synthesis with Differentially Private Generative Adversarial Networks and Differentially Private Selection*.

Ed-Daoudi, R., Alaoui, A., Ettaki, B., & Zerouaoui, J. (2023). Improving Crop Yield Predictions in Morocco Using Machine Learning Algorithms. *Journal of Ecological Engineering*, *24*(6), 392–400. https://doi.org/10.12911/22998993/162769

Ehteram, M., Najah Ahmed, A., Khozani, Z. S., & El-Shafie, A. (2023). Graph convolutional network – Long short term memory neural network- multi layer perceptron- Gaussian progress regression model: A new deep learning model for predicting ozone concertation. *Atmospheric Pollution Research*, *14*(6), 101766. https://doi.org/10.1016/j.apr.2023.101766

Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*. https://doi.org/10.1007/s10994-022-06296-4

Er-Retby, H., Zoubir, Z., Kaitouni, S. I., Mghazli, M. O., Elmankibi, M., & Benzaazoua, M. (2024). *Indoor Environment's Quality IEQ Forecasting for a Residential Building Using Machine Learning Models* (pp. 249–260). https://doi.org/10.1007/978-981-99-8501-2_23

Fallahian, M., Dorodchi, M., & Kreth, K. (2024). GAN-Based Tabular Data Generator for Constructing Synopsis in Approximate Query Processing: Challenges and Solutions. *Machine Learning and Knowledge Extraction*, *6*(1), 171–198. https://doi.org/10.3390/make6010010

Fassò, A., Rodeschini, J., Moro, A. F., Shaboviq, Q., Maranzano, P., Cameletti, M., Finazzi, F., Golini, N., Ignaccolo, R., & Otto, P. (2023). Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data*, *10*(1), 143. https://doi.org/10.1038/s41597-023-02034-0

Fatah, F. A. (2017). *Competitiveness and Efficiency of Rice Production in Malaysia* [Georg-August-University Göttingen]. https://doi.org/10.53846/goediss-6160

Fatchurrachman, Rudiyanto, Soh, N. C., Shah, R. M., Giap, S. G. E., Setiawan, B. I., & Minasny, B. (2023). Automated near-real-time mapping and monitoring of rice growth extent and stages in Selangor Malaysia. *Remote Sensing Applications: Society and Environment*, *31*, 100993. https://doi.org/10.1016/j.rsase.2023.100993

Fatima, G., Khan, S., Aadil, F., Kim, D. H., Atteia, G., & Alabdulhafith, M. (2024). An autonomous mixed data oversampling method for AIOT-based churn recognition and personalized recommendations using behavioral segmentation. *PeerJ Computer Science*, *9*, e1756. https://doi.org/10.7717/peerj-cs.1756

Finnøy, I. (2023). *Can Tabular Generative Models Generate Realistic Synthetic Near Infrared Spectroscopic Data?* Norwegian University of Life Sciences.

Forestier, G., Petitjean, F., Dau, H. A., Webb, G. I., & Keogh, E. (2017). Generating synthetic time series to augment sparse datasets. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2017-November*, 865–870. https://doi.org/10.1109/ICDM.2017.106

Friedberg, R., & Rogers, R. (2023). Privacy Aware Experimentation over Sensitive Groups: A General Chi Square Approach. In A. Dieng, M. Rateike, G. Farnadi, F. Fioretto, M. Kusner, & J. Schrouff (Eds.), *Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Privacy* (Vol. 214, pp. 23–66). PMLR. https://proceedings.mlr.press/v214/friedberg23a.html

Frifra, A., Maanan, M., Maanan, M., & Rhinane, H. (2024). Harnessing LSTM and XGBoost algorithms for storm prediction. *Scientific Reports*, *14*(1), 11381. https://doi.org/10.1038/s41598-024-62182-0

Gafrej, O. (2024). Predicting customer deposits with machine learning algorithms: evidence from Tunisia. *Managerial Finance*, *50*(3), 578–589. https://doi.org/10.1108/MF-02-2023-0135

Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–5. https://doi.org/10.1109/JCSSE.2016.7748856

Gangwar, A., Singh, S., Mishra, R., & Prakash, S. (2023). The State-of-the-Art in Air Pollution Monitoring and Forecasting Systems Using IoT, Big Data, and Machine Learning. *Wireless Personal Communications*, *130*(3), 1699–1729. https://doi.org/10.1007/s11277-023-10351-1

Gharibi, A., Babazadeh, R., & Hasanzadeh, R. (2024). Machine learning and multi-criteria decision analysis for polyethylene air-gasification considering energy and environmental aspects. *Process Safety and Environmental Protection*, *183*, 46–58. https://doi.org/10.1016/j.psep.2023.12.069

Gill, K. S., Gupta, R., Kumar, M., Rawat, R., & Chanti, Y. (2024). Healthcare Insurance Cost Classification Using Machine Learning's Linear Regression Approach. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT61223.2024.10544081

Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. *Neural Computing and Applications*, *33*(22), 15693–15707. https://doi.org/10.1007/s00521-021-06189-y

Goldreich, O. (2020). *The Uniform Distribution Is Complete with Respect to Testing Identity to a Fixed Distribution* (pp. 152–172). https://doi.org/10.1007/978-3-030-43662-9_10

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Robin Bryant, C., & Senthilnath, J. (2021). Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecological Indicators*, *120*, 106935. https://doi.org/10.1016/j.ecolind.2020.106935

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hammouch, H., El-Yacoubi, M., Qin, H., Berrahou, A., Berbia, H., & Chikhaoui, M. (2021). A two-stage Deep Convolutional Generative Adversarial Network-based data augmentation scheme for agriculture image regression tasks. *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, 1–6. https://doi.org/10.1109/ICCSI53130.2021.9736230

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning* (pp. 878–887). https://doi.org/10.1007/11538059_91

Han, X., Liu, F., He, X., & Ling, F. (2022). Research on Rice Yield Prediction Model Based on Deep Learning. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/1922561

Hasan, M. R. (2024). Addressing Seasonality and Trend Detection in Predictive Sales Forecasting: A Machine Learning Perspective. *Journal of Business and Management Studies*, *6*(2), 100–109. https://doi.org/10.32996/jbms.2024.6.2.10

Hashim, N., Ali, M. M., Mahadi, M. R., Abdullah, A. F., Wayayok, A., Mohd Kassim, M. S., & Jamaluddin, A. (2024). Smart Farming for Sustainable Rice Production: An Insight into Application, Challenge, and Future Prospect. *Rice Science*, *31*(1), 47–61. https://doi.org/10.1016/j.rsci.2023.08.004

Hematibahar, M., Kharun, M., Beskopylny, A. N., Stel'makh, S. A., Shcherban', E. M., & Razveeva, I. (2024). Analysis of Models to Predict Mechanical Properties of High-Performance and Ultra-High-Performance Concrete Using Machine Learning. *Journal of Composites Science*, *8*(8), 287. https://doi.org/10.3390/jcs8080287

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. In *Neurocomputing* (Vol. 493, pp. 28–45). Elsevier B.V. https://doi.org/10.1016/j.neucom.2022.04.053

Hernández–Bermejo, B., & Sorribas, A. (2001). Analytical Quantile Solution for the S-distribution, Random Number Generation and Statistical Data Modeling. *Biometrical Journal*, *43*(8), 1007. https://doi.org/10.1002/1521-4036(200112)43:8<1007::AID-BIMJ1007>3.0.CO;2-F

Hesamian, G., Johannssen, A., & Chukhrova, N. (2024). An explainable fused lasso regression model for handling high-dimensional fuzzy data. *Journal of*

Computational and Applied Mathematics, *441*, 115721. https://doi.org/10.1016/j.cam.2023.115721

Hu, Y., Wu, F., Li, Q., Long, Y., Garrido, G. M., Ge, C., Ding, B., Forsyth, D., Li, B., & Song, D. (2024). SoK: Privacy-Preserving Data Synthesis. *2024 IEEE Symposium on Security and Privacy (SP)*, 4696–4713. https://doi.org/10.1109/SP54263.2024.00002

Iantovics, L. B., & Enăchescu, C. (2022). Method for Data Quality Assessment of Synthetic Industrial Data. *Sensors*, *22*(4), 1608. https://doi.org/10.3390/s22041608

Ilyas, Q. M., Ahmad, M., & Mehmood, A. (2023). Automated Estimation of Crop Yield Using Artificial Intelligence and Remote Sensing Technologies. *Bioengineering*, *10*(2). https://doi.org/10.3390/bioengineering10020125

Islam, M. T., Ayon, E. H., Ghosh, B. P., Chowdhury, M. S., Shahid, R., puja, A. R., Rahman, S., Akter, A., Rahman, M., & Bhuiyan, M. S. (2024). Revolutionizing Retail: A Hybrid Machine Learning Approach for Precision Demand Forecasting and Strategic Decision-Making in Global Commerce. *Journal of Computer Science and Technology Studies*, *6*(1), 33–39. https://doi.org/10.32996/jcsts.2024.6.1.4

Jabeur, S. Ben, Mefteh-Wali, S., & Viviani, J.-L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*, *334*(1–3), 679–699. https://doi.org/10.1007/s10479-021-04187-w

Jayanthi, S., Priyadharshini, V., Kirithiga, V., & Premalatha, S. (2024). Mental health status monitoring for people with autism spectrum disorder using machine learning. *International Journal of Information Technology*, *16*(1), 43–51. https://doi.org/10.1007/s41870-023-01524-z

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques. *Procedia Computer Science*, *218*, 406–417. https://doi.org/10.1016/j.procs.2023.01.023

Jogunuri, S., F.T, J., Stonier, A. A., Peter, G., Jayaraj, J., S, J., J, J. J., & Ganji, V. (2024). Random forest machine learning algorithm based seasonal multi-step ahead short-term solar photovoltaic power output forecasting. *IET Renewable Power Generation*. https://doi.org/10.1049/rpg2.12921

Jonayet, M., Das, A. C., Mozumder, M. S. A., Hasan, M. A., Bhuiyan, M., Islam, M. R., Hossain, M. N., Akter, S., & Alam, M. I. (2024). Machine Learning Approaches for Demand Forecasting: The Impact of Customer Satisfaction on Prediction Accuracy. *The American Journal of Engineering and Technology*, *6*(10), 42–53. https://doi.org/10.37547/tajet/Volume06Issue10-06

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. *Knowledge-Based Systems*, *207*, 106368. https://doi.org/10.1016/j.knosys.2020.106368

Karunanithi, M., Chatasawapreeda, P., & Khan, T. A. (2024). A predictive analytics approach for forecasting bike rental demand. *Decision Analytics Journal*, *11*, 100482. https://doi.org/10.1016/j.dajour.2024.100482

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., & Markuzon, N. (2021). Application of Bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association*, *28*(4), 801–811. https://doi.org/10.1093/jamia/ocaa303

Kawade, R. R., & Jagtap, S. K. (2023). Optimal trained ensemble of classification model for speech emotion recognition: Considering cross-lingual and multi-lingual scenarios. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-17097-9

Kebede, E. A., Abou Ali, H., Clavelle, T., Froehlich, H. E., Gephart, J. A., Hartman, S., Herrero, M., Kerner, H., Mehta, P., Nakalembe, C., Ray, D. K., Siebert, S., Thornton, P., & Davis, K. F. (2024). Assessing and addressing the global state of food production data scarcity. *Nature Reviews Earth & Environment*, *5*(4), 295–311. https://doi.org/10.1038/s43017-024-00516-2

Ketha Dhana Veera Chaitanya, & Manas Kumar Yogi. (2023). Role of Synthetic Data for Improved AI Accuracy. *Journal of Artificial Intelligence and Capsule Networks*, *5*(3), 330–345. https://doi.org/10.36548/jaicn.2023.3.008

Khorrami, B. M., Soleimani, A., Pinnarelli, A., Brusco, G., & Vizza, P. (2024). Correction: Forecasting heating and cooling loads in residential buildings using machine learning: a comparative study of techniques and influential indicators.

Asian Journal of Civil Engineering, *25*(2), 2349–2351. https://doi.org/10.1007/s42107-023-00865-1

Kim, S., Lee, J., Jeong, K., Lee, J., Hong, T., & An, J. (2024). Automated door placement in architectural plans through combined deep-learning networks of ResNet-50 and Pix2Pix-GAN. *Expert Systems with Applications*, *244*, 122932. https://doi.org/10.1016/j.eswa.2023.122932

Kornish, D., Ezekiel, S., & Cornacchia, M. (2018). DCNN Augmentation via Synthetic Data from Variational Autoencoders and Generative Adversarial Networks. *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–6. https://doi.org/10.1109/AIPR.2018.8707390

Kozaklı, Ö., Ceyhan, A., & Noyan, M. (2024). Comparison of machine learning algorithms and multiple linear regression for live weight estimation of Akkaraman lambs. *Tropical Animal Health and Production*, *56*(7), 250. https://doi.org/10.1007/s11250-024-04049-0

Kumar, V., Kedam, N., Kisi, O., Alsulamy, S., Khedher, K. M., & Salem, M. A. (2024). A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting. *Water Resources Management*. https://doi.org/10.1007/s11269-024-03969-8

Kunar, A. (2021). *Effective and Privacy preserving Tabular Data Synthesizing*.

Lall, A. (2015). Data streaming algorithms for the Kolmogorov-Smirnov test. *2015 IEEE International Conference on Big Data (Big Data)*, 95–104. https://doi.org/10.1109/BigData.2015.7363746

Latif, S. D. (2023). Evaluating deep learning and machine learning algorithms for forecasting daily pan evaporation during COVID-19 pandemic. *Environment, Development and Sustainability*, *26*(5), 11729–11742. https://doi.org/10.1007/s10668-023-03469-6

Lederrey, G., Hillel, T., & Bierlaire, M. (2021, September 12). DATGAN: Integrating expert knowledge into deeplearning for population synthesis. *Proceedings of the 21st Swiss Transport Research Conference*. https://discovery.ucl.ac.uk/id/eprint/10174120/

Lederrey, G., Hillel, T., & Bierlaire, M. (2022). *DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data*.

Li, D., Liu, S., Lyu, Z., Xiang, W., He, W., Liu, F., & Zhang, Z. (2021). Use mean field theory to train a 200-layer vanilla GAN. *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 420–426. https://doi.org/10.1109/ICTAI52525.2021.00068

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2023). ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12181–12193. https://doi.org/10.1109/TKDE.2022.3159580

Liang, C., Yaohong, L., Gang, C., Tong, X., & Wei, Y. (2022). Insulator image data enhancement based on BIG-GAN. *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, 398–402. https://doi.org/10.1109/ICPECA53709.2022.9719111

Liang, G., On, B.-W., Jeong, D., Heidari, A. A., Kim, H.-C., Choi, G. S., Shi, Y., Chen, Q., & Chen, H. (2021). A text GAN framework for creative essay recommendation. *Knowledge-Based Systems*, *232*, 107501. https://doi.org/10.1016/j.knosys.2021.107501

Liang, H., Sun, L., Wei, J., Huang, X., Sun, L., Yu, B., He, C., & Zhang, W. (2024). *Synth-Empathy: Towards High-Quality Synthetic Empathy Data*.

Liao, S., Ni, H., Sabate-Vidales, M., Szpruch, L., Wiese, M., & Xiao, B. (2024). Sig-Wasserstein GANs for conditional time series generation. *Mathematical Finance*, *34*(2), 622–670. https://doi.org/10.1111/mafi.12423

Limanto, S., Buliali, J. L., & Saikhu, A. (2024). GLoW SMOTE-D: Oversampling Technique to Improve Prediction Model Performance of Students Failure in Courses. *IEEE Access*, *12*, 8889–8901. https://doi.org/10.1109/ACCESS.2024.3351569

Liu, C. (2024). Implied Volatility Forecasting for American Options Based on Random Forest Regressor, Linear Regression Model. *Advances in Economics, Management and Political Sciences*, *85*(1), 154–160. https://doi.org/10.54254/2754-1169/85/20240867

Liu, C.-H., Peng, C.-H., Huang, L.-Y., Chen, F.-Y., Kuo, C.-H., Wu, C.-Z., & Cheng, Y.-F. (2024). Comparison of multiple linear regression and machine learning methods in predicting cognitive function in older Chinese type 2 diabetes patients. *BMC Neurology*, *24*(1), 11. https://doi.org/10.1186/s12883-023-03507-w

Livieris, I. E., Alimpertis, N., Domalis, G., & Tsakalidis, D. (2024). An Evaluation Framework for Synthetic Data Generation Models. In *IFIP Advances in*

Information and Communication Technology (Vol. 713, pp. 320–335). Springer. https://doi.org/10.1007/978-3-031-63219-8_24

Lotfi, R., Changizi, M., MohajerAnsari, P., Hosseini, A., Javaheri, Z., & Ali, S. S. (2024). A robust, resilience machine learning with risk approach: a case study of gas consumption. *Annals of Operations Research*. https://doi.org/10.1007/s10479-024-05986-7

Lu, Y., Chen, L., Zhang, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2025). Machine Learning for Synthetic Data Generation: A Review. *Machine Learning (Cs.LG)*.

Lu, Z., Liu, H., Wei, P., & Zorko, D. (2024). *A Strategy for Contact Fatigue Life Prediction of Polymer Gears via an Experimental-simulated Hybrid Data-driven Model*. https://ssrn.com/abstract=4735031

Mahmoud, A., & Mohammed, A. (2024). Leveraging Hybrid Deep Learning Models for Enhanced Multivariate Time Series Forecasting. *Neural Processing Letters*, *56*(5), 223. https://doi.org/10.1007/s11063-024-11656-3

Mandapati, R., Gumma, M. K., Metuku, D. R., & Maitra, S. (2023). Field-level rice yield estimations under different farm practices using the crop simulation model for better yield. *Plant Science Today*. https://doi.org/10.14719/pst.2690

Manoli, G., Bonetti, S., Scudiero, E., Morari, F., Putti, M., & Teatini, P. (2015). Modeling Soil–Plant Dynamics: Assessing Simulation Accuracy by Comparison with Spatially Distributed Crop Yield Measurements. *Vadose Zone Journal*, *14*(12), 1–13. https://doi.org/10.2136/vzj2015.05.0069

Mansourifar, H., Moskovitz, A., Klingensmith, B., Mintas, D., & Simske, S. J. (2022). GAN-Based Satellite Imaging: A Survey on Techniques and Applications. *IEEE Access*, *10*, 118123–118140. https://doi.org/10.1109/ACCESS.2022.3221123

Marong, M., Husin, N. A., Zolkepli, M., & Affendey, L. S. (2024). Data-Driven Rice Yield Predictions and Prescriptive Analytics for Sustainable Agriculture in Malaysia. *International Journal of Advanced Computer Science and Applications*, *15*(3). https://doi.org/10.14569/IJACSA.2024.0150337

Meher, B. K., Singh, M., Birau, R., & Anand, A. (2024). Forecasting stock prices of fintech companies of India using random forest with high-frequency data. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(1), 100180. https://doi.org/10.1016/j.joitmc.2023.100180

Meyer, D., & Nagler, T. (2021). Synthia: multidimensional synthetic data generation in Python. *Journal of Open Source Software*, *6*(65), 2863. https://doi.org/10.21105/joss.02863

Meyer, D., Nagler, T., & Hogan, R. J. (2021). Copula-based synthetic data augmentation for machine-learning emulators. *Geoscientific Model Development*, *14*(8), 5205–5215. https://doi.org/10.5194/gmd-14-5205-2021

Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., & Yadav, R. (2024). Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. *Environmental Earth Sciences*, *83*(6), 163. https://doi.org/10.1007/s12665-024-11481-w

Mohanty, N., Behera, B. K., Ferrie, C., & Dash, P. (2025). A quantum approach to synthetic minority oversampling technique (SMOTE). *Quantum Machine Intelligence*, *7*(1), 38. https://doi.org/10.1007/s42484-025-00248-6

Mohd Shafie, S. N., Ab. Aziz, N., Nafi, M. N. A., A.Malek @ Abdul Malek, S., Amran, A., & Mohd Shafie, S. N. A. (2024). Predicting Paddy Production in Malaysia: A Comparative Analysis between Arima and Neural Network Autoregressive (NNAR) Models. *International Journal of Academic Research in Business and Social Sciences*, *14*(12). https://doi.org/10.6007/IJARBSS/v14-i12/23351

Mohd Shafri, A. S., Mohd Nor, S. R., & Siti Mariam Norrulashikin. (2025). Enhancing Paddy Production in Malaysia: A Comparative Analysis of Multiple Regression with External Factors. *Malaysian Journal of Fundamental and Applied Sciences*, *21*(1), 1750–1763. https://doi.org/10.11113/mjfas.v21n1.3991

Mohsin, M. F. M., Hassan, M. G., Mohd Sharif, K. I., Ismail, M. A., Shahron, S. A., & Ja'afar, N. A. (2023). Modelling of Ladyfinger Plantations Using Open Data Malaysia: A Comparative Study of Machine Learning Techniques. *Emerging Advances in Integrated Technology*, *4*(2).

Mokryn, O., Abbey, A., Marmor, Y., & Shahar, Y. (2024). Evaluating the dynamic interplay of social distancing policies regarding airborne pathogens through a temporal interaction-driven model that uses real-world and synthetic data. *Journal of Biomedical Informatics*, *151*, 104601. https://doi.org/10.1016/j.jbi.2024.104601

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Morales-García, J., Bueno-Crespo, A., Terroso-Sáenz, F., Arcas-Túnez, F., Martínez-España, R., & Cecilia, J. M. (2023). Evaluation of synthetic data generation for intelligent climate control in greenhouses. *Applied Intelligence*, *53*(21), 24765–24781. https://doi.org/10.1007/s10489-023-04783-2

Mouli, K. C., Raghavendran, Ch. V., Mallikarjuna Rao, Ch., Ushasree, D., Indupriya, B., Vatin, N. I., & Negi, A. S. (2024). Performance analysis of linear and non-linear machine learning models for forecasting compressive strength of concrete. *Cogent Engineering*, *11*(1). https://doi.org/10.1080/23311916.2024.2368101

Mukherjee, M., & Khushi, M. (2021a). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Applied System Innovation*, *4*(1), 18. https://doi.org/10.3390/asi4010018

Mukherjee, M., & Khushi, M. (2021b). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Applied System Innovation*, *4*(1), 18. https://doi.org/10.3390/asi4010018

Munkhammar, J., van der Meer, D., & Widén, J. (2021). Very short term load forecasting of residential electricity consumption using the Markov-chain mixture distribution (MCM) model. *Applied Energy*, *282*, 116180. https://doi.org/10.1016/j.apenergy.2020.116180

Museru, M. L., Nazari, R., Giglou, A. N., Opare, K., & Karimi, M. (2024). Advancing flood damage modeling for coastal Alabama residential properties: A multivariable machine learning approach. *Science of the Total Environment*, *907*. https://doi.org/10.1016/j.scitotenv.2023.167872

Nawa, V. M., & Nadarajah, S. (2023). New Closed Form Estimators for the Beta Distribution. *Mathematics*, *11*(13), 2799. https://doi.org/10.3390/math11132799

Nazibudin, N. A., Sabri, S. A. M., & Manaf, L. A. (2025). Aligning with sustainable development goals (SDG) 12: A systematic review of food waste generation in Malaysia. *Cleaner Waste Systems*, *10*, 100205. https://doi.org/10.1016/j.clwas.2025.100205

Nguyen, V.-H., Tuyet-Hanh, T. T., Mulhall, J., Minh, H. Van, Duong, T. Q., Chien, N. Van, Nhung, N. T. T., Lan, V. H., Minh, H. B., Cuong, D., Bich, N. N., Quyen, N. H., Linh, T. N. Q., Tho, N. T., Nghia, N. D., Anh, L. V. Q., Phan, D. T. M., Hung, N. Q. V., & Son, M. T. (2022). Deep learning models for forecasting dengue fever based on climate data in Vietnam. *PLOS Neglected Tropical Diseases*, *16*(6), e0010509. https://doi.org/10.1371/journal.pntd.0010509

Nguyen, X. C., Nguyen, T. P., Lam, V. S., Le, P.-C., Vo, T. D. H., Hoang, T.-H. T., Chung, W. J., Chang, S. W., & Nguyen, D. D. (2024). Estimating ammonium changes in pilot and full-scale constructed wetlands using kinetic model, linear regression, and machine learning. *Science of The Total Environment*, *907*, 168142. https://doi.org/10.1016/j.scitotenv.2023.168142

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software*, *174*, 105971. https://doi.org/10.1016/j.envsoft.2024.105971

Nirmala, J., Sravya, S. G., & Tharun, K. (2024). Housing Market Intelligence: Data Science for Rental Price Forecasting. *2024 3rd International Conference for*

Innovation in Technology (INOCON), 1–9. https://doi.org/10.1109/INOCON60754.2024.10511646

Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLOS ONE*, *18*(3), e0283452. https://doi.org/10.1371/journal.pone.0283452

Nuthalapati, S. babu, & Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. *International Journal of Science and Research Archive*, *12*(2), 408–422. https://doi.org/10.30574/ijsra.2024.12.2.1246

Nwanze, N. E., Iweka, S. C., Madu, K. E., & Edafiadhe, E. D. (2024). Solar Radiation Forecasting Models and their Thermodynamic Analysis in Asaba: Least Square Regression and Machine Learning Approach. *Journal of Energy Research and Reviews*, *16*(2), 9–21. https://doi.org/10.9734/jenrr/2024/v16i2333

Oikonomidis, A., Catal, C., & Kassahun, A. (2023). Deep learning for crop yield prediction: a systematic literature review. In *New Zealand Journal of Crop and Horticultural Science* (Vol. 51, Issue 1, pp. 1–26). Taylor and Francis Ltd. https://doi.org/10.1080/01140671.2022.2032213

Olofintuyi, S. S., Olajubu, E. A., & Olanike, D. (2023). An ensemble deep learning approach for predicting cocoa yield. *Heliyon*, *9*(4). https://doi.org/10.1016/j.heliyon.2023.e15245

Oriani, F., Stisen, S., Demirel, M. C., & Mariethoz, G. (2020). Missing Data Imputation for Multisite Rainfall Networks: A Comparison between Geostatistical

Interpolation and Pattern-Based Estimation on Different Terrain Types. *Journal of Hydrometeorology*, *21*(10), 2325–2341. https://doi.org/10.1175/JHM-D-19-0220.1

Ørting, S. N., Stephensen, H. J. T., & Sporring, J. (2023). Morphology on Categorical Distributions. *Journal of Mathematical Imaging and Vision*, *65*(6), 861–873. https://doi.org/10.1007/s10851-023-01146-x

Osborne, J. W. (2017). Simple Linear Models With Categorical Dependent Variables: Binary Logistic Regression. In *Regression & Linear Modeling: Best Practices and Modern Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781071802724

Oukhouya, H., & El Himdi, K. (2023). Comparing Machine Learning Methods—SVR, XGBoost, LSTM, and MLP— For Forecasting the Moroccan Stock Market. *IOCMA 2023*, 39. https://doi.org/10.3390/IOCMA2023-14409

Patil, A., & Venkatesh. (2021). DCGAN: Deep Convolutional GAN with Attention Module for Remote View Classification. *2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)*, 1–10. https://doi.org/10.1109/FABS52071.2021.9702655

Patrikar, A. M., Mahenthiran, A., & Said, A. (2023). Leveraging synthetic data for AI bias mitigation. In K. E. Manser, R. M. Rao, & C. L. Howell (Eds.), *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications* (p. 21). SPIE. https://doi.org/10.1117/12.2662276

Pham, H. T., Awange, J., Kuhn, M., Nguyen, B. Van, & Bui, L. K. (2022). Enhancing Crop Yield Prediction Utilizing Machine Learning on Satellite-Based Vegetation Health Indices. *Sensors*, *22*(3), 719. https://doi.org/10.3390/s22030719

Prakash, T. D., & Nagaraju, V. (2024). *Improving accuracy in road accidents prediction by comparing support vector machine with Lasso regression*. 020154. https://doi.org/10.1063/5.0197585

Rafiei, A., Ghiasi Rad, M., Sikora, A., & Kamaleswaran, R. (2024). Improving mixed-integer temporal modeling by generating synthetic data using conditional generative adversarial networks: A case study of fluid overload prediction in the intensive care unit. *Computers in Biology and Medicine*, *168*, 107749. https://doi.org/10.1016/j.compbiomed.2023.107749

Rajabi, A., & Garibay, O. O. (2022). TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks. *Machine Learning and Knowledge Extraction*, *4*(2), 488–501. https://doi.org/10.3390/make4020022

Ramachandran, K., Kayathwal, K., Wadhwa, H., & Dhama, G. (2023). FraudAmmo: Large Scale Synthetic Transactional Dataset for Payment Fraud Detection. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN54540.2023.10191990

Rauf, O., Supervisor, A., Zhao, Z., & Eemcs, L. C. (n.d.). *Extending CTAB-GAN with StackGAN*.

Razghandi, M., Zhou, H., Erol-Kantarci, M., & Turgut, D. (2022). Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in

Smart Home. *ICC 2022 - IEEE International Conference on Communications*, 4781–4786. https://doi.org/10.1109/ICC45855.2022.9839249

Reddy, A. P., Tarafdar, A., & Bera, U. K. (2023). Regression based Machine Learning approach to predict Flight Price between Bangalore and Kolkata. *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT57861.2023.10126456

Richert, I., & Buch, R. (2024). Interpolation of missing swaption volatility data using variational autoencoders. *Behaviormetrika*, *51*(1), 291–317. https://doi.org/10.1007/s41237-023-00213-2

Rigden, A. J., Golden, C., & Huybers, P. (2022). Retrospective Predictions of Rice and Other Crop Production in Madagascar Using Soil Moisture and an NDVI-Based Calendar from 2010–2017. *Remote Sensing*, *14*(5), 1223. https://doi.org/10.3390/rs14051223

Riza, L. S., Yudianita, A. H., Nugraha, E., Somantri, L., Sitanggang, I. S., Abu Samah, K. A. F., & Nazir, S. (2022). Remote sensing and machine learning for yield prediction of lowland paddy crops. *F1000Research*, *11*, 682. https://doi.org/10.12688/f1000research.110608.1

Rossi, D., Mascolo, A., Mancini, S., Breton, J. G. C., Breton, R. M. C., & Guarnaccia, C. (2023). Modelling and Forecast of Air Pollution Concentrations during COVID Pandemic Emergency with ARIMA Techniques: the Case Study of Two Italian Cities. *WSEAS TRANSACTIONS ON ENVIRONMENT AND DEVELOPMENT*, *19*, 151–162. https://doi.org/10.37394/232015.2023.19.13

Şahin, B., Evren, A. A., Tuna, E., Şahinbaşoğlu, Z. Z., & Ustaoğlu, E. (2023). Parameter Estimation of the Dirichlet Distribution Based on Entropy. *Axioms*, *12*(10), 947. https://doi.org/10.3390/axioms12100947

Saini, P., & Nagpal, B. (2024). Spatiotemporal Landsat-Sentinel-2 Satellite Imagery-based Hybrid Deep Neural Network for Paddy Crop Prediction using Google Earth Engine. *Advances in Space Research*. https://doi.org/10.1016/j.asr.2024.02.032

Sangeetha, J. M., & Alfia, K. J. (2024). Financial stock market forecast using evaluated linear regression based machine learning technique. *Measurement: Sensors*, *31*, 100950. https://doi.org/10.1016/j.measen.2023.100950

Santos de Jesus, E. dos, & Silva Gomes, G. S. da. (2023). Machine learning models for forecasting water demand for the Metropolitan Region of Salvador, Bahia. *Neural Computing and Applications*, *35*(27), 19669–19683. https://doi.org/10.1007/s00521-023-08842-0

Satpathi, A., Setiya, P., Das, B., Nain, A. S., Jha, P. K., Singh, S., & Singh, S. (2023). Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India. *Sustainability (Switzerland)*, *15*(3). https://doi.org/10.3390/su15032786

Sauber-Cole, R., & Khoshgoftaar, T. M. (2022). The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00648-6

Sayed, K., Syakir, M. I., Othman, A. A., Azhar, B., Tohiran, K. A., & Nobilly, F. (2025). Introducing the RICE framework in paddy-duck farming: a novel approach to enhance the adaptive capacity of paddy farmers among asnaf community in managing the risk of climate change. *Archives of Agronomy and Soil Science*, *71*(1), 1–16. https://doi.org/10.1080/03650340.2025.2465744

Shafqat, W., & Byun, Y.-C. (2022). A Hybrid GAN-Based Approach to Solve Imbalanced Data Problem in Recommendation Systems. *IEEE Access*, *10*, 11036–11047. https://doi.org/10.1109/ACCESS.2022.3141776

Shaon, Md. S. H., Karim, T., Shakil, Md. S., & Hasan, Md. Z. (2024). A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction. *Healthcare Analytics*, *6*, 100353. https://doi.org/10.1016/j.health.2024.100353

Sharma, D. K., Singh, B., Raja, M., Regin, R., & Rajest, S. S. (2021). An Efficient Python Approach for Simulation of Poisson Distribution. *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2011–2014. https://doi.org/10.1109/ICACCS51430.2021.9441895

Shin, D., Han, D., & Kyeong, S. (2022). Performance Enhancement of Malware Classifiers Using Generative Adversarial Networks. *2022 IEEE International Conference on Big Data (Big Data)*, 6528–6533. https://doi.org/10.1109/BigData55660.2022.10020505

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Singh, P., Adebanjo, A., Shafiq, N., Razak, S. N. A., Kumar, V., Farhan, S. A., Adebanjo, I., Singh, A., Dixit, S., Singh, S., & Sergeevna, M. T. (2024). Development of performance-based models for green concrete using multiple linear regression and artificial neural network. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *18*(5), 2945–2956. https://doi.org/10.1007/s12008-023-01386-6

Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: a machine learning approach and visualization app for data science and management. *Data Science and Management*, *7*(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

Singh, S., Kayathwal, K., Wadhwa, H., & Dhama, G. (2021). *MeTGAN: Memory Efficient Tabular GAN for High Cardinality Categorical Datasets* (pp. 519–527). https://doi.org/10.1007/978-3-030-92310-5_60

Sireesha, P. V., & Thotakura, S. (2024). Wind power prediction using optimized MLP-NN machine learning forecasting model. *Electrical Engineering*, *106*(6), 7643–7666. https://doi.org/10.1007/s00202-024-02440-6

Skryjomski, P. aw, & Krawczyk, B. (2017). Influence of minority class instance types on SMOTE imbalanced data oversampling. *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 7–21.

Soltaninejad, M., Aghazadeh, R., Shaghaghi, S., & Zarei, M. (2024). Using Machine Learning Techniques to Forecast Mehram Company's Sales: A Case Study. *Journal of Business and Management Studies*, *6*(2), 42–53. https://doi.org/10.32996/jbms.2024.6.2.4

Stanly, H., S., M. S., & Paul, R. (2023). A review of generative and non-generative adversarial attack on context-rich images. *Engineering Applications of Artificial Intelligence*, *124*, 106595. https://doi.org/10.1016/j.engappai.2023.106595

Subrahmayam, K. V., Udupa, S. R., Kumar, K. K., Ramana, M. V., Srinivasulu, J., & Bothale, R. V. (2024). Prediction of Zonal Wind Using Machine Learning Algorithms: Implications to Future Projections of Indian Monsoon Jets. *Journal of the Indian Society of Remote Sensing*, *52*(2), 371–381. https://doi.org/10.1007/s12524-024-01817-1

Suman, N., Rao, A., & Ranjitha. (2024). *Forecasting air quality using random forest regression with hyperparameter optimization and LSTM network (RNN)*. 020046. https://doi.org/10.1063/5.0185778

Sun, C., van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, *143*, 104404. https://doi.org/10.1016/j.jbi.2023.104404

Sun, Z., Zhang, H., Bai, J., Liu, M., & Hu, Z. (2023). A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition. *Pattern Recognition*, *135*, 109157. https://doi.org/10.1016/j.patcog.2022.109157

Tan, B. T., Fam, P. S., Firdaus, R. B. R., Tan, M. L., & Gunaratne, M. S. (2021). Impact of Climate Change on Rice Yield in Malaysia: A Panel Data Analysis. *Agriculture*, *11*(6), 569. https://doi.org/10.3390/agriculture11060569

Tan, H., Chen, R., Qin, M., Tang, L., Wu, Z., Luo, Q., & Quan, Y. (2023). Tabular GAN-Based Oversampling of Imbalanced Time-to-Event Data for Survival Prediction. *2023 8th International Conference on Cloud Computing and Big Data*

Analytics (ICCCBDA), 376–380. https://doi.org/10.1109/ICCCBDA56900.2023.10154883

Tasneem, S., Ageeli, A. A., Alamier, W. M., Hasan, N., & Safaei, M. R. (2024). Organic catalysts for hydrogen production from noodle wastewater: Machine learning and deep learning-based analysis. *International Journal of Hydrogen Energy*, *52*, 599–616. https://doi.org/10.1016/j.ijhydene.2023.07.114

Toma, M. B., Belete, M. D., & Ulsido, M. D. (2023). Trends in climatic and hydrological parameters in the Ajora-Woybo watershed, Omo-Gibe River basin, Ethiopia. *SN Applied Sciences*, *5*(1), 45. https://doi.org/10.1007/s42452-022-05270-y

Torsoni, G. B., de Oliveira Aparecido, L. E., dos Santos, G. M., Chiquitto, A. G., da Silva Cabral Moraes, J. R., & de Souza Rolim, G. (2023). Soybean yield prediction by machine learning and climate. *Theoretical and Applied Climatology*, *151*(3–4), 1709–1725. https://doi.org/10.1007/s00704-022-04341-9

Tran, T. K., Nguyen, C. T., Phung, N. H., Truong, T. T. H., Tran, D. C., Nguyen, T. T., & Lex, T. M. P. (2022). Evaluation of Synthetic Data Generating Methods in Down Syndrome Prediction During the First Trimester Screening in Vietnam. *2022 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2022 - Proceedings*. https://doi.org/10.1109/MAPR56351.2022.9924885

Tyralis, H., & Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review*, *57*(4), 94. https://doi.org/10.1007/s10462-023-10698-8

Vaghefi, N., Shamsudin, M. N., Radam, A., & Rahim, K. A. (2013a). Impact of Climate Change on Rice Yield in the Main Rice Growing Areas of Peninsular Malaysia. *Research Journal of Environmental Sciences*, *7*(2), 59–67. https://doi.org/10.3923/rjes.2013.59.67

Vaghefi, N., Shamsudin, M. N., Radam, A., & Rahim, K. A. (2013b). Impact of Climate Change on Rice Yield in the Main Rice Growing Areas of Peninsular Malaysia. *Research Journal of Environmental Sciences*, *7*(2), 59–67. https://doi.org/10.3923/rjes.2013.59.67

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. In *Computers and Electronics in Agriculture* (Vol. 177). Elsevier B.V. https://doi.org/10.1016/j.compag.2020.105709

Vance, J., Rasheed, K., Missaoui, A., & Maier, F. W. (2022). Data Synthesis for Alfalfa Biomass Yield Estimation. *AI*, *4*(1), 1–15. https://doi.org/10.3390/ai4010001

Vignesh, K., Askarunisa, A., & Abirami, A. M. (2023). Optimized Deep Learning Methods for Crop Yield Prediction. *Computer Systems Science and Engineering*, *44*(2), 1051–1067. https://doi.org/10.32604/csse.2023.024475

Wang, A. X., Chukova, S. S., Sporle, A., Milne, B. J., Simpson, C. R., & Nguyen, B. P. (2024). Enhancing public research on citizen data: An empirical investigation of data synthesis using Statistics New Zealand's Integrated Data Infrastructure. *Information Processing and Management*, *61*(1). https://doi.org/10.1016/j.ipm.2023.103558

Wang, R., Ding, Y., Li, L., & Fan, C. (2020). One-Shot Voice Conversion Using Star-Gan. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7729–7733. https://doi.org/10.1109/ICASSP40776.2020.9053842

Wang, T., & Lin, Y. (2024). CycleGAN with Better Cycles. *Computer Vision and Pattern Recognition (Cs.CV)*.

Wang, X. (2023). Analysis and Application Research of InfoGAN. *Highlights in Science, Engineering and Technology*, *57*, 21–26. https://doi.org/10.54097/hset.v57i.9891

Wang, Y., Shi, W., & Wen, T. (2023). Prediction of winter wheat yield and dry matter in North China Plain using machine learning algorithms for optimal water and nitrogen application. *Agricultural Water Management*, *277*. https://doi.org/10.1016/j.agwat.2023.108140

Wen, B., Colon, L. O., Subbalakshmi, K. P., & Chandramouli, R. (2021). *Causal-TGAN: Generating Tabular Data Using Causal Generative Adversarial Networks*.

Wickramasinghe, L., Weliwatta, R., Ekanayake, P., & Jayasinghe, J. (2021). Modeling the Relationship between Rice Yield and Climate Variables Using Statistical and Machine Learning Techniques. *Journal of Mathematics*, *2021*, 1–9. https://doi.org/10.1155/2021/6646126

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, *32*, 7335–7345. https://doi.org/10.5555/3454287.345494

Yang, D., & Li, M. (2023). A Novel Data Statistical Analysis Method Based on Hyper-Erlang Distribution. *2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI)*, 66–71. https://doi.org/10.1109/ICCBD-AI62252.2023.00019

Yang, Y., & Li, J. (2022). Electric Vehicle Charging Anomaly Detection Method Based on Multivariate Gaussian Distribution Model. *2022 Asia Conference on Electrical, Power and Computer Engineering (EPCE 2022)*, 1–6. https://doi.org/10.1145/3529299.3531488

Yildirim, A., Bilgili, M., & Ozbek, A. (2023). One-hour-ahead solar radiation forecasting by MLP, LSTM, and ANFIS approaches. *Meteorology and Atmospheric Physics*, *135*(1), 10. https://doi.org/10.1007/s00703-022-00946-x

Yuan, S., Wu, P., & Chen, Y. (2023). Data-level Hybrid Strategy Selection for Disk Fault Prediction Model Based on Multivariate GAN. *International Journal on Cybernetics & Informatics*, *12*(6), 53–68. https://doi.org/10.5121/ijci.2023.120605

Yusri, N. H. M., Shafie, N. A., & Ghani, N. A. Md. (2022). Rice Price Prediction in Malaysia. *2022 IEEE International Conference on Computing (ICOCO)*, 249–252. https://doi.org/10.1109/ICOCO56118.2022.10031931

Zellinger, M. J., & Bühlmann, P. (2025). Natural Language-Based Synthetic Data Generation for Cluster Analysis. *Machine Learning (Cs.LG)*.

Zema, D. A., Parhizkar, M., Plaza-Alvarez, P. A., Xu, X., & Lucas-Borja, M. E. (2024). Using random forest and multiple-regression models to predict changes in surface runoff and soil erosion after prescribed fire. *Modeling Earth Systems and Environment*, *10*(1), 1215–1228. https://doi.org/10.1007/s40808-023-01838-8

Zhang, H., Huang, L., Wu, C. Q., & Li, Z. (2020). An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. *Computer Networks*, *177*, 107315. https://doi.org/10.1016/j.comnet.2020.107315

Zhang, L., & Jánošík, D. (2024). Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert Systems with Applications*, *241*, 122686. https://doi.org/10.1016/j.eswa.2023.122686

Zhang, X., Liu, J., & Zhu, Z. (2024). Learning Coefficient Heterogeneity over Networks: A Distributed Spanning-Tree-Based Fused-Lasso Regression. *Journal of the American Statistical Association*, *119*(545), 485–497. https://doi.org/10.1080/01621459.2022.2126363

Zhang, Y., Zhou, Z., Liu, J., & Yuan, J. (2022). Data augmentation for improving heating load prediction of heating substation based on TimeGAN. *Energy*, *260*, 124919. https://doi.org/10.1016/j.energy.2022.124919

Zhao, Z., Birke, R., Kunar, A., & Chen, L. Y. (2021). *Fed-TGAN: Federated Learning Framework for Synthesizing Tabular Data*.

Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. *Asian Conference on Machine Learning*, 97–112.

Zhao, Z., Kunar, A., Van der Scheer, H., Birke, R., & Chen, L. Y. (2021). *CTAB-GAN: Effective Table Data Synthesizing*.

Zheng, J., Xin, D., Cheng, Q., Tian, M., & Yang, L. (2024). *The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance*.

Zhou, C. (2018). Expressing the randomity of events – An analysis of random number generation with given distributions. *University of Ottawa Science Undergraduate Research Journal*, *1*, 47–54. https://doi.org/10.18192/osurj.v1i1.3702

Zhou, K., Gao, S., Cheng, J., Gu, Z., Fu, H., Tu, Z., Yang, J., Zhao, Y., & Liu, J. (2020). Sparse-Gan: Sparsity-Constrained Generative Adversarial Network for Anomaly Detection in Retinal OCT Image. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1227–1231. https://doi.org/10.1109/ISBI45749.2020.9098374

Zhou, S., Xu, L., & Chen, N. (2023). Rice Yield Prediction in Hubei Province Based on Deep Learning and the Effect of Spatial Heterogeneity. *Remote Sensing*, *15*(5), 1361. https://doi.org/10.3390/rs15051361

Zhu, J., Ai, Q., Chen, Y., & Wang, J. (2023). Single-Location and Multi-Locations Scenarios Generation for Wind Power Based On WGAN-GP. *Journal of Physics: Conference Series*, *2452*(1), 012022. https://doi.org/10.1088/1742-6596/2452/1/012022

Zhu, Y., Zhao, Z., Birke, R., & Chen, L. Y. (2022). Permutation-Invariant Tabular Data Synthesis. *2022 IEEE International Conference on Big Data (Big Data)*, 5855–5864. https://doi.org/10.1109/BigData55660.2022.10020639

Zulkifley, M. A., Abdul Aziz, N. A., & Akhyar, A. (2024). Gross Paddy Yield Estimation Using a Lightweight Deep Learning Model. *2024 International*

Conference on Decision Aid Sciences and Applications (DASA), 1–5. https://doi.org/10.1109/DASA63652.2024.10836375

Appendix A

CLIMATE DATA SAMPLE (TEMPERATURE) (^{O}C)

MM	YYYY	KEDAH	JOHOR	PPINANG	PERLIS	PERAK	KELANTAN	PAHANG	MELAKA	SELANGOR	SARAWAK	SABAH
Jan	2010	28.2	26.4	27.9	27.3	26.9	25.8	26.3	27.3	27.3	26.2	26.6
Feb	2010	29.4	27.6	29.2	28.7	28.3	26.8	27.6	28.5	28.3	27.2	27.5
Mar	2010	29.4	27.4	29.1	29.1	28.2	27.3	28.0	28.3	28.3	27.3	28.2
Apr	2010	29.3	27.4	29.1	28.7	28.3	28.2	28.2	28.2	28.7	27.4	28.1
May	2010	29.7	27.8	29.5	29.2	28.4	28.8	28.4	28.6	28.7	27.7	28.2
Jun	2010	28.3	26.9	28.2	27.6	27.5	27.6	27.5	28.0	28.3	27.1	27.6
Jul	2010	27.6	26.3	27.8	26.9	27.2	27.2	26.9	27.4	27.8	26.5	26.7
Aug	2010	28.0	26.2	28.3	27.5	27.4	27.3	27.0	27.3	28.2	26.8	27.2
Sep	2010	27.6	26.4	27.8	26.8	27.0	26.9	26.9	27.3	27.6	26.5	27.1
Oct	2010	28.1	26.7	28.3	27.0	27.5	27.0	27.0	27.9	28.3	26.8	27.0
Nov	2010	27.3	26.2	27.2	26.5	26.6	26.0	26.4	27.1	27.5	26.5	26.8
Dec	2010	26.8	25.6	26.5	25.8	25.8	25.1	25.4	26.4	26.8	26.1	26.4
Jan	2011	27.5	25.2	27.1	26.5	26.0	25.2	25.1	25.8	26.9	25.8	26.0
Feb	2011	28.1	26.3	27.9	27.7	27.0	26.0	26.3	27.3	28.0	26.0	26.0
Mar	2011	27.4	26.1	27.3	26.8	26.5	26.0	26.2	27.2	27.7	26.3	25.9
Apr	2011	28.5	26.9	28.3	28.1	27.1	27.3	27.4	27.8	28.1	26.7	26.9
May	2011	28.5	27.2	28.6	27.8	27.7	27.6	27.5	28.1	28.8	27.0	27.4
Jun	2011	28.0	27.0	28.6	27.5	27.8	27.4	27.4	27.9	28.8	27.1	27.4
Jul	2011	27.9	27.2	28.5	27.3	27.7	27.2	27.1	27.7	28.5	27.1	27.1
Aug	2011	27.2	26.8	27.6	26.8	26.8	27.0	27.2	27.4	28.1	27.0	27.1
Sep	2011	27.5	26.4	27.7	26.9	26.8	26.7	26.9	27.4	28.1	26.9	27.3
Oct	2011	27.7	26.2	27.4	27.1	26.7	26.0	26.2	27.1	27.6	26.6	27.2
Nov	2011	27.7	26.5	27.3	26.8	26.5	25.8	26.2	26.8	27.3	26.8	26.9
Dec	2011	27.7	26.0	27.5	26.6	26.6	25.5	25.6	26.3	27.4	26.4	26.8

Appendix B

CLIMATE DATA SAMPLE (HUMIDITY) (%)

MM	YYYY	KEDAH	JOHOR	PPINANG	PERLIS	PERAK	KELANTAN	PAHANG	MELAKA	SELANGOR	SARAWAK	SABAH
Jan	2010	70.9	82.0	75.9	73.5	83.1	83.7	83.8	76.3	79.4	86.2	83.6
Feb	2010	68.8	79.2	74.5	70.2	79.3	80.5	79.5	73.7	77.4	83.8	78.0
Mar	2010	72.1	81.3	76.3	70.1	79.6	79.2	79.4	76.9	78.4	84.2	76.8
Apr	2010	78.6	84.8	80.6	79.0	82.9	80.8	81.7	80.7	79.8	84.1	81.1
May	2010	79.8	84.3	79.7	79.1	83.4	80.5	82.5	81.2	81.9	84.5	82.4
Jun	2010	82.7	85.0	81.7	82.3	83.1	80.8	83.4	80.0	79.6	84.1	82.8
Jul	2010	83.8	85.7	80.6	83.4	83.7	82.1	84.3	81.0	80.1	85.0	83.9
Aug	2010	83.1	86.4	78.7	82.3	82.0	81.9	83.8	81.7	79.3	83.1	80.9
Sep	2010	83.9	85.3	81.6	83.5	83.1	83.3	83.5	81.6	82.4	84.3	82.3
Oct	2010	82.5	83.9	79.3	83.4	81.8	83.5	84.0	78.5	77.9	82.9	81.3
Nov	2010	84.1	85.9	84.9	84.5	85.2	87.7	85.1	81.6	81.5	84.5	83.8
Dec	2010	83.2	85.6	84.8	83.9	85.7	87.8	86.9	82.9	82.1	85.3	84.6
Jan	2011	73.8	85.2	77.3	76.2	83.0	83.0	85.5	82.8	82.9	86.5	85.8
Feb	2011	73.6	80.2	76.1	71.7	80.1	79.7	80.4	76.5	79.7	85.5	84.9
Mar	2011	81.1	85.3	82.8	81.1	84.0	85.4	85.9	80.2	81.5	84.9	86.3
Apr	2011	80.2	83.6	79.9	76.6	82.8	80.2	81.5	79.8	82.3	85.2	84.1
May	2011	83.6	86.3	79.9	81.8	82.4	80.3	83.8	81.1	79.7	84.1	83.6
Jun	2011	85.0	87.3	78.7	83.4	81.1	81.0	83.8	82.6	78.8	84.1	82.6
Jul	2011	83.5	84.3	78.5	81.2	78.7	82.8	83.5	81.8	78.2	83.0	81.2
Aug	2011	85.2	85.4	81.3	82.7	82.9	82.2	83.0	82.1	80.1	83.2	80.9
Sep	2011	84.9	86.8	82.0	83.1	83.2	82.7	84.3	82.4	78.9	83.9	79.4
Oct	2011	83.2	88.8	83.3	81.9	84.6	86.9	87.2	83.8	82.6	86.0	80.4
Nov	2011	80.0	87.6	83.0	83.4	86.5	88.1	87.5	85.3	85.8	85.9	83.5
Dec	2011	75.1	87.2	77.9	80.7	83.8	86.7	87.9	85.1	82.7	87.6	84.3

Appendix C

THE WHOLE RESULT OF EXPERIMENTS

Methods	Metrics	Regression Models					
		Linear Regression	Lasso Regression	Decision Tree Regressor	Random Forest Regressor	XGBRegressor	MLPRegressor
Baseline	**MAE**	12310.96808	12261.0922	10448.84906	7860.72434	7610.760659	16004.38525
	MSE	325585024.7	324807317.7	545096039.9	230653632.4	181473557.9	555202720.7
	RMSE	18043.97475	18022.41154	23347.29192	15187.28522	13471.21219	23562.74009
	R-Squared	0.982417403	0.982459402	0.970563131	0.987543992	0.990199868	0.970017339
SMOTE-ENC	**MAE**	12310.96808	12261.0922	10055.26415	7810.518302	7610.760659	15989.88538
	MSE	325585024.7	324807317.7	528982684.9	220377228.7	181473557.9	543063177.3
	RMSE	18043.97475	18022.41154	22999.62358	14845.1079	13471.21219	23303.71596
	R-Squared	0.982417403	0.982459402	0.971433302	0.988098949	0.990199868	0.970672912
CTAB-GAN	**MAE Avg.**	70580.01461	70586.15435	23162.28302	26743.04601	28972.57464	18487.46374
	MAE (Best)	18176.58204	18170.60843	1291.830189	5068.154528	3019.251773	16072.32481
	MAE (Worst)	130076.4915	130113.5029	108063.1698	82581.15585	78571.2688	20762.07224
	MSE Avg.	8452661263	8453019362	6603563031	2188448145	3037236293	659985079
	MSE (Best)	615486215	615306142.8	10397950.89	99133945.57	60589838.46	543612882.1
	MSE (Worst)	26528428733	26523250975	44588920356	9857590125	10332039887	761489195.6
	RMSE Avg.	82376.50059	82376.68395	57358.28903	37349.34995	44284.45136	25643.09216
	RMSE (Best)	24808.99464	24805.3652	3224.585382	9956.603114	7783.947486	23315.50733
	RMSE (Worst)	162875.5007	9956.603114	8082.604925	23315.50733	38111.63652	38099.57263
	R-Squared Avg.	0.543530192	0.543510854	0.643387206	0.881817043	0.835979862	0.964358768
	R-Squared (Best)	0.966761844	0.966771568	0.999438479	0.994646461	0.996727962	0.970643226
	R-Squared (Worst)	-0.432617064	-0.432337449	-1.407939378	0.467659696	0.442037943	0.958877232
SCEG	**MAE Avg.**	38063.60176	38066.0283	5180.73148	7256.923479	8640.924794	24977.15093
	MAE (Best)	16225.41066	16225.12504	659.307231	3870.579988	1671.848444	13856.73423
	MAE (Worst)	71529.2939	71533.3667	25455.25313	22263.44433	26943.33723	56497.73942
	MSE Avg.	3777738785	3778249000	515745152.3	209455250.1	305765351.6	1737206466
	MSE (Best)	467444366.8	467246026.6	8524055.056	43745361.59	14911136.07	429305052.6
	MSE (Worst)	10729336823	10730809977	3970728231	1216023867	1366453674	7436293174

Methods	Metrics	Regression Models					
		Linear Regression	Lasso Regression	Decision Tree Regressor	Random Forest Regressor	XGBRegressor	MLPRegressor
	RMSE Avg.	55169.33172	55172.39255	16393.90908	12409.24616	13605.93211	37802.67437
	RMSE (Best)	21620.46176	21615.87441	2919.598441	6614.027638	3861.494021	20719.67791
	RMSE (Worst)	103582.5122	6614.027638	5282.083705	22677.45903	26595.3457	26596.19504
	R-Squared Avg.	0.795990441	0.795962888	0.972148169	0.98868877	0.983487727	0.906185487
	R-Squared (Best)	0.974756561	0.974767272	0.999539675	0.997637615	0.999194753	0.976816202
	R-Squared (Worst)	0.420582682	0.420503127	0.785568415	0.934330956	0.926207282	0.598417207
Post-processing (Limit)	**MAE Avg.**	22395.98517	22395.76799	2981.794184	4836.314055	4199.970129	17926.73393
	MAE (Best)	12922.1565	12920.73223	1081.252847	3700.982127	2146.713711	13598.42465
	MAE (Worst)	43900.50621	43905.95695	4727.649579	6984.686209	9471.429357	26333.41503
	MSE Avg.	1097312531	1097384130	126008830.7	94717793.8	102831363.6	732488529.5
	MSE (Best)	348662444.1	348548905.6	12114893.7	36856084.18	23351572.94	422195732.6
	MSE (Worst)	3513089577	3513836704	365204319.2	209023546.6	242913967.6	1648038237
	RMSE Avg.	30561.72229	30561.74448	10198.46738	9353.905144	9324.490251	26440.13404
	RMSE (Best)	18672.50503	18669.46452	3480.645587	6070.921197	4832.346525	20547.40209
	RMSE (Worst)	59271.32171	6070.921197	4832.346525	20547.40209	18704.65852	18706.43225
	R-Squared Avg.	0.940741735	0.940737869	0.993195134	0.994884947	0.994446789	0.960443358
	R-Squared (Best)	0.981171151	0.981177283	0.999345758	0.998009658	0.998738943	0.977200127
	R-Squared (Worst)	0.810282315	0.810241968	0.980277839	0.988712083	0.986881896	0.911000847
Post-processing (Limit and IQR)	**MAE Avg.**	11385.06438	11377.94522	2844.836995	4808.330976	3029.444757	15182.25259
	MAE (Best)	11368.18027	11364.30796	2548.814291	4657.63065	2983.192954	14545.23963
	MAE (Worst)	11386.5993	11379.18497	3170.109395	5000.901683	3538.214595	15605.09258
	MSE Avg.	287457750.1	287054928.1	124891481.1	107659276.2	67870919.08	503173732.6
	MSE (Best)	287363071.8	286949063.3	89165551.41	96657052.04	66413790.7	469025867.1
	MSE (Worst)	288499212	288219440.7	164512271.4	120450891	83899331.27	528776400
	RMSE Avg.	16954.57651	16942.69226	11075.35873	10372.01364	8233.648348	22428.68787
	RMSE (Best)	16951.78668	16939.57093	9442.751263	9831.431841	8149.465669	21657.00503
	RMSE (Worst)	16985.26456	9831.431841	8149.465669	21657.00503	16951.78668	16939.57093
	R-Squared Avg.	0.984476394	0.984498148	0.993255474	0.994186067	0.996334761	0.972827065
	R-Squared (Best)	0.98448151	0.98450386	0.99518478	0.99478022	0.99641345	0.97467115
	R-Squared (Worst)	0.98442015	0.98443526	0.99111583	0.99349528	0.99546918	0.97144444

Appendix D

THE IMPROVEMENT OF PROPOSED METHOD

Methods	Data Size	Best MAE	Best MSE	Best RMSE	Best R-squared	MAE Improvement (%)	MSE Improvement (%)	RMSE Improvement (%)	R-squared Improvement (%)
Baseline	-	7610.76066	181473557.9	13471.2122	0.99019987	-	-	-	-
SMOTE-ENC	-	7610.76066	181473557.9	13471.2122	0.99019987	No improvement for the best model. but SMOTE-ENC has the better value in Decision Tree. Random Forest. and MLP Regression (see Table 4.2)			
CTAB-GAN	0.1	3019.251773	60589838.46	7783.947486	0.996727962	60.33%	66.61%	42.22%	0.65%
	0.2	4912.661561	173014572.8	13153.5004	0.990656679	35.45%	4.66%	2.36%	0.05%
	0.25	1291.830189	10397950.89	3224.585382	0.999438479	83.03%	94.27%	76.06%	0.92%
	0.5	8824.458679	283257533.4	16830.25649	0.984703219	-15.95%	-56.09%	-24.93%	-0.56%
	0.75	10523.37472	406005411.5	20149.57596	0.978074454	-38.27%	-123.73%	-49.58%	-1.24%
	1	6118.528302	188239439.9	13720.0379	0.989834489	19.61%	-3.73%	-1.85%	-0.04%
	1.5	12658.28302	574677856.2	23972.43951	0.968965621	-66.32%	-216.67%	-77.95%	-2.19%
	2	12722.67925	644348645.2	25384.02342	0.965203184	-67.17%	-255.06%	-88.43%	-2.59%
	3	18965.19492	668392899.7	25853.29572	0.96390472	-149.19%	-268.31%	-91.92%	-2.73%
	4	20476.72566	761489195.6	27595.09369	0.958877232	-169.05%	-319.61%	-104.84%	-3.27%
	5	17749.34069	584276026.2	24171.8023	0.96844729	-133.21%	-221.96%	-79.43%	-2.25%

Methods	Data Size	Best MAE	Best MSE	Best RMSE	Best R-squared	MAE Improvement (%)	MSE Improvement (%)	RMSE Improvement (%)	R-squared Improvement (%)
	10	20762.07224	736087422.2	27130.9311	0.960249006	-172.80%	-305.62%	-101.40%	-3.12%
SCEG (Before Post-processing)	0.1	1671.848444	14911136.07	3861.494021	0.999194753	78.03%	91.78%	71.34%	0.90%
	0.2	2234.618969	27900408.27	5282.083705	0.998493292	70.64%	84.63%	60.79%	0.83%
	0.25	2391.08445	32844299.11	5730.9946	0.998226306	68.58%	81.90%	57.46%	0.80%
	0.5	2491.220158	69752964.56	8351.824026	0.996233125	67.27%	61.56%	38.00%	0.61%
	0.75	3305.339377	31025918.05	5570.091386	0.998324505	56.57%	82.90%	58.65%	0.81%
	1	659.307231	8524055.056	2919.598441	0.999539675	91.34%	95.30%	78.33%	0.93%
	1.5	1882.940466	40427032.76	6358.225598	0.997816815	75.26%	77.72%	52.80%	0.76%
	2	2443.09718	83856240.99	9157.305334	0.995471504	67.90%	53.79%	32.02%	0.53%
	3	3658.882138	118240086.7	10873.82576	0.99361467	51.92%	34.84%	19.28%	0.34%
	4	1939.39822	17810912.15	4220.297638	0.999038156	74.52%	90.19%	68.67%	0.88%
	5	9137.01476	339127290.3	18415.40905	0.981686079	-20.05%	-86.87%	-36.70%	-0.87%
	10	22263.44433	1216023867	34871.53377	0.934330956	-192.53%	-570.08%	-158.86%	-5.98%
Post-pro (Limit)	0.1	3638.814863	85956564.59	9271.276319	0.99535808	52.19%	52.63%	31.18%	0.52%
	0.2	1789.525719	26076182.35	5106.484343	0.998591806	76.49%	85.63%	62.09%	0.84%
	0.25	2146.713711	23351572.94	4832.346525	0.998738943	71.79%	87.13%	64.13%	0.85%

Methods	Data Size	Best MAE	Best MSE	Best RMSE	Best R-squared	MAE Improvement (%)	MSE Improvement (%)	RMSE Improvement (%)	R-squared Improvement (%)
	0.5	4163.931709	81366587.21	9020.342965	0.995605953	45.29%	55.16%	33.04%	0.54%
	0.75	2883.678943	67946428.07	8242.962336	0.996330683	62.11%	62.56%	38.81%	0.62%
	1	2880.694672	58472227.08	7646.713482	0.996842319	62.15%	67.78%	43.24%	0.67%
	1.5	1984.027602	26402803.8	5138.365869	0.998574167	73.93%	85.45%	61.86%	0.84%
	2	2195.994689	28622661.84	5350.015125	0.998454288	71.15%	84.23%	60.29%	0.83%
	3	4326.812642	112891057.5	10625.02035	0.993903534	43.15%	37.79%	21.13%	0.37%
	4	1336.883934	23801147.61	4878.641985	0.998714665	82.43%	86.88%	63.78%	0.85%
	5	1081.252847	12114893.7	3480.645587	0.999345758	85.79%	93.32%	74.16%	0.92%
	10	4181.269137	156484368.1	12509.37121	0.991549361	45.06%	13.77%	7.14%	0.14%
Post-pro (Limit and IQR)	0.1	2983.192954	66413790.7	8149.465669	0.99641345	60.80%	63.40%	39.50%	0.62%
	0.2	2983.192954	66413790.7	8149.465669	0.99641345	60.80%	63.40%	39.50%	0.62%
	0.25	2578.133374	66413790.7	8149.465669	0.99641345	66.13%	63.40%	39.50%	0.62%
	0.5	2983.192954	66413790.7	8149.465669	0.99641345	60.80%	63.40%	39.50%	0.62%
	0.75	2740.279213	66413790.7	8149.465669	0.99641345	63.99%	63.40%	39.50%	0.62%
	1	2548.814291	66413790.7	8149.465669	0.99641345	66.51%	63.40%	39.50%	0.62%
	1.5	2591.374145	66413790.7	8149.465669	0.99641345	65.95%	63.40%	39.50%	0.62%

Methods	Data Size	Best MAE	Best MSE	Best RMSE	Best R-squared	MAE Improvement (%)	MSE Improvement (%)	RMSE Improvement (%)	R-squared Improvement (%)
	2	2598.79267	66413790.7	8149.465669	0.99641345	65.85%	63.40%	39.50%	0.62%
	3	2983.192954	66413790.7	8149.465669	0.99641345	60.80%	63.40%	39.50%	0.62%
	4	2695.772205	66413790.7	8149.465669	0.99641345	64.58%	63.40%	39.50%	0.62%
	5	2983.192954	66413790.7	8149.465669	0.99641345	60.80%	63.40%	39.50%	0.62%
	10	2735.406663	83899331.27	9159.657814	0.995469177	64.06%	53.77%	32.01%	0.53%

Appendix E

Published Article Based on This Study

Title: *Improving Rice Yield Prediction Accuracy Using Regression Models with Climate Data*
Authors: Mohamad Farhan Mohamad Mohsin, Muhammad Khalifa Umana, Mohamad Ghozali Hassan, Kamal Imran Mohd Sharif, Mohd Azril Ismail, Khazainani Salleh, Suhaili Mohd Zahari, Mimi Adilla Sarmani & Neil Gordon
Conference: *International Conference on Computing and Informatics*
Year: 2024
DOI: https://doi.org/10.1007/978-981-99-9592-9_20

Title: *Prediction of Rice Yields in a Changing Climate Using the Mobile Rice Yield Prediction Application*
Authors: Mohamad Farhan Mohamad Mohsin, Muhammad Khalifa Umana, Mohamad Ghozali Hassan, Kamal Imran Mohd Sharif, Mohd Azril Ismail, Khazainani Salleh, Suhaili Mohd Zahari, Mimi Adilla Sarmani & Neil Gordon
Journal: *Journal of Information and Communication Technology (JICT) Vol.24, No.2, April 2025*
Volume: 24, **Issue**: 2, **Pages**: 1-18
Year: 2025
DOI: https://doi.org/10.32890/jict2025.24.2.1

Title: *Malaysia Rice Yield Prediction Using Synthetic Data Generation of CTAB-GAN*
Authors: Muhammad Khalifa Umana, Mohamad Farhan Mohamad Mohsin, Mohamad Ghozali Hassan, Kamal Imran Mohd Sharif, Mohd Azril Ismail & Maslina Ismail
Conference: Knowledge Management International Conference (KMICe)
Volume: 49, **Pages**: 473–484
Year: 2025
DOI: https://doi.org/10.32890/jict2025.24.2.1