engineering proceedings

MDPI

Proceeding Paper

Prediction of Mechanical Properties of Austenitic Stainless Steels with the Use of Synthetic Data via Generative Adversarial Networks †

Desmarita Leni [1,*], Dytchia Septi Kesuma [2], Maimuzar [3], Haris [3] and Sicilia Afriyani [4]

[1] Department of Mechanical Engineering, Universitas Muhammadiyah Sumatera Barat, Padang 25172, Indonesia
[2] Department of Electronics Engineering, Universitas Muhammadiyah Sumatera Barat, Padang 25172, Indonesia; dytchia@gmail.com
[3] Department of Mechanical Engineering, Politeknik Negeri Padang, Padang 25164, Indonesia; maimuzar@ac.pnp.id (M.); haris@ac.pnp.id (H.)
[4] Department of Civil Engineering, Politeknik Negeri Padang, Padang 25164, Indonesia; sicilia@ac.pnp.id
* Correspondence: desmaritaleni@gmail.com
† Presented at the 7th Mechanical Engineering, Science and Technology International Conference, Surakarta, Indonesia, 21–22 December 2023.

Abstract: This study involves data augmentation modeling using Generative Adversarial Networks (GAN) on the tensile test data of austenitic stainless steel, which encompasses chemical compositions, heat treatments, and mechanical properties. The synthetic data generated by GAN is subsequently used as the training dataset for six different algorithm models. The best-performing algorithm is selected based on the best evaluation metric values. The results of the Kolmogorov–Smirnov (KS) test indicate that the distribution of synthetic data does not significantly differ from the distribution of experimental data. Furthermore, the training results of predictive models employing synthetic data with the six machine learning algorithms demonstrate that the gradient boosting model exhibits superior performance in predicting the mechanical properties of austenitic stainless steel.

Keywords: mechanical properties; austenitic stainless steels; synthetic data; generative adversarial networks

check for updates

Citation: Leni, D.; Kesuma, D.S.; Maimuzar; Haris; Afriyani, S. Prediction of Mechanical Properties of Austenitic Stainless Steels with the Use of Synthetic Data via Generative Adversarial Networks. *Eng. Proc.* **2024**, *63*, 4. https://doi.org/10.3390/engproc2024063004

Academic Editors: Waluyo Adi Siswanto, Sarjito, Supriyono, Agus Dwi Anggono, Tri Widodo Besar Riyadi and Taurista Perdana Syawitri

Published: 22 February 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Materials informatics is a novel approach in the field of materials science that integrates information technology and materials science to optimize the discovery of new materials more efficiently and innovatively [1,2]. In materials informatics, experimental data and simulations are combined with data-driven methods such as big data, data augmentation, and machine learning to gain a deeper understanding of material properties [3,4]. Data augmentation is a technique used to expand data by modifying existing data using machine learning algorithms [5]. Generative Adversarial Networks (GAN) are one type of artificial neural network architecture used to generate realistic synthetic data. GAN consists of two main models: the generator and the discriminator. The generator produces synthetic data that closely resembles real data, while the discriminator distinguishes between real and synthetic data. These two models compete during the training process, where the generator aims to produce synthetic data that increasingly resembles real data, while the discriminator strives to distinguish between real and synthetic data more accurately. The training process concludes when the generator generates synthetic data that closely resembles real data, making it indistinguishable from the discriminator [6].

Previous studies have examined the use of GANs in generating synthetic data in tabular form related to the multiaxial fatigue life of various types of materials, such as aluminum, brass, and stainless steel. Findings from these studies indicate that the use of

synthetic data can lead to improvements in the accuracy of machine learning models for predicting the fatigue life of these materials [7]. In a study [8] on predicting the strength of high-strength concrete, it was discovered that GANs could increase the number of original experimental data points from an initial 810 samples with 15 input features and one output, which is the compressive strength of concrete. The results of this research state that GANs can expand the volume of original data by eightfold, and this data can predict 810 experimental data points that were not present in the training data.

Based on the issues outlined above and the positive impact of previous research on the use of GANs as a modeling approach to generate synthetic material property data, this research aims to evaluate the modeling of predicting the mechanical properties of stainless steel using synthetic data generated by GANs. The results of this research are expected to contribute to the development of data augmentation techniques for material mechanical properties, particularly to support advancements in technology and a better understanding of the mechanical properties of austenitic stainless steel.

2. Research Method

This study is an experimental research type with the aim of generating synthetic data regarding alloy chemical elements, heat treatment temperature, heating time, cooling media, and the mechanical properties of austenitic stainless steel in tabular form. The synthetic data is produced using the Generative Adversarial Networks (GAN) method. Subsequently, the synthetic data generated by GAN is utilized as a dataset for training machine learning models to predict the mechanical properties of austenitic stainless steel. The data augmentation and machine learning modeling processes are conducted using the Python programming language with the TensorFlow library, executed within the Google Colab environment. The dataset used in this study consists of tensile test results for various types of austenitic stainless steel (ASS), such as SUS 304, SUS 316, SUS 321, SUS 347, and NCF 800H. This dataset comprises 1916 data points, including the mechanical properties of austenitic stainless steel, alloy chemical elements, heat treatment temperatures, and cooling methods. The data was obtained from the British Steelmakers Creep Committee [9]. This data was collected by the Material Algorithm Project (MAP), which is a project conducted by the University of Cambridge. An illustration of the research stages can be found in Figure 1 to provide a clearer overview of how this entire study was carried out.

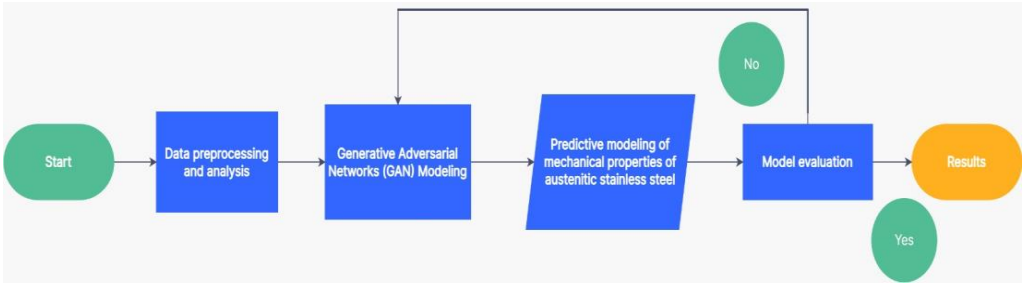


Figure 1. Research Scheme.

2.1. Data Preprocessing and Analysis

Data preprocessing is the stage of data preparation in research, which includes handling missing values, cleaning data from outliers, encoding categorical variables, and normalizing data to ensure it is ready for analysis [10]. In this study, data normalization is carried out using the MinMaxScaler method. Subsequently, data analysis is performed using descriptive statistical methods and Pearson correlation to understand the patterns, relationships, and characteristics of the austenitic stainless steel tensile test data.

2.2. Generative Adversarial Networks (GAN) Modeling

In this stage, generative adversarial networks (GAN) are employed to obtain synthetic data on the mechanical properties of stainless steel. GAN is a deep learning algorithm

consisting of two networks, namely the generator and the discriminator [6]. The mechanical properties of austenitic stainless steel, heat treatment temperature, and chemical elements are used as the training dataset for the model, where the generator aims to produce synthetic data that closely resembles the real data while the discriminator learns to distinguish between real and synthetic data. The architecture of GAN can be seen in Figure 2.

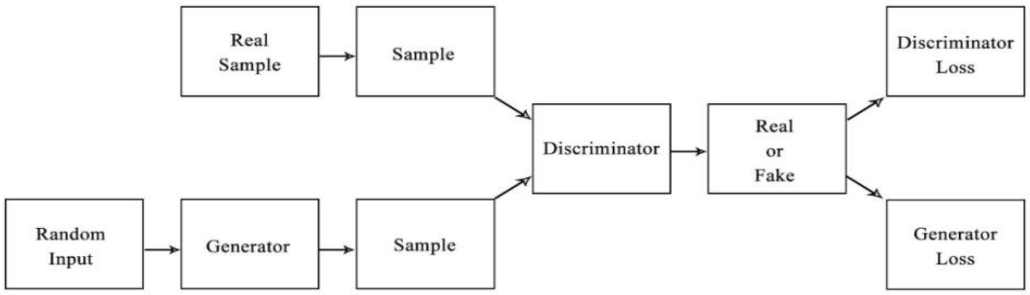


Figure 2. Generative Adversarial Networks (GAN) Architecture [11].

The synthetic data generated will be evaluated using descriptive statistics and the Kolmogorov–Smirnov (KS) test to assess how closely the synthetic data for austenitic stainless steel's tensile properties resemble the actual data. The Kolmogorov–Smirnov (KS) test can be computed using Equation (1).

$$D = \max\left(F1_{(x)} - F2_{(x)}\right) \quad (1)$$

where D is the KS Test, $F1_{(x)}$ is the empirical distribution function (ECDF) of the first sample, and $F2_{(x)}$ is the empirical distribution function (ECDF) of the second sample [12].

2.3. Modeling the Prediction of Austenitic Stainless Steel Mechanical Properties

In this stage, a comparison of six algorithms is conducted, including linear regression (LR), decision trees (DT), random forests (RF), K-nearest neighbors (KNN), gradient boosting (GB), and artificial neural networks (ANN). These algorithms are implemented using the default parameters provided by the scikit-learn library version 1.2.2 and Keras version 2.12.0 for ANN. The data used for training and testing the models is divided into two parts: 80% for training data and 20% for testing data. Each machine learning model is validated using cross-validation, a technique that allows the training data to be divided into several different subsets or folds. Iterations are performed on each subset to serve as testing data, while the other subsets are used as training data [13,14]. In the prediction modeling of austenitic stainless steel mechanical properties, cross-validation with a K value of 10 is used, where the data is divided into 10 different subsets or folds, and iterations are performed 10 times, selecting each subset alternately as the testing data and the others as the training data.

2.4. Model Evaluation

Model evaluation is conducted using a separate testing dataset from the training dataset to assess the predictive capability of the model on previously unseen data. In this study, the evaluation metrics employed are root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2) [15]. These evaluation metrics can be calculated using the following equations:

1. Mean Absolute Error (MAE)

$$MAE = \frac{i}{N}\sum|y_i - z_i| \quad (2)$$

where i is the index of data in the sample, N is the total number of samples, y_i is the actual value of the i-th data, and z_i is the model's predicted value for the i-th data.

2. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2} \quad (3)$$

where n is the number of data points used to test the model, $f(Xi)$ is the value predicted by the model for the i-th data point, and Y_i is the actual value for the i-th data point.

3. R-squared

$$R = \frac{\sum_{i}^{n} = 1\left(f(X_i) - f(\overline{X})\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(f(X_i) - f(\overline{X}\right))^2}\sqrt{\sum_{i=1}^{n} Y_i - \overline{Y})^2}} \quad (4)$$

3. Results and Discussion

3.1. Data Preprocessing and Analysis

The dataset of austenitic stainless steel comprises 2180 samples, but after data preprocessing, 1194 data samples were found to contain unnecessary information such as missing and invalid values, rendering them unusable for this research. The original database includes other features such as melting type, grain size, and product shape; however, these data were incomplete and had a very low correlation with the mechanical properties of austenitic stainless steel. As a result, only 986 data samples with complete and relevant information for this research were used. These samples consist of 20 input variables and 3 output variables. The input variables include chemical elements and heat treatment, while the output variables encompass the mechanical properties of austenitic stainless steel, including yield strength (YS), ultimate tensile strength (UTS), and elongation (EL).

3.2. Generative Adversarial Networks (GAN) Modeling

In this study, the synthetic data to be generated will match the number of original data samples, which is 986. This is done to facilitate the evaluation of synthetic data against real data. The GAN parameters used in this research can be seen in Table 1, and the loss function during model training is shown in Figure 3.

Table 1. GAN Parameters.

No.	Parameters	Value	No.	Parameters	Value
1	Epochs	1500	7	Gen Output Activation	11
2	Batch	128	8	Neuron Layer Dense Pert	256
3	Vektor Noise	100	9	Neuron Layer Dense Ked	128
4	Output Gen	11	10	Disc Output Activation	Sigmoid
5	Neuron Layer Dense Pert	128	11	Loss Function Disc	Binary Crossentropy
6	Neuron Layer Dense Ked	256	12	Loss Function GAN	Binary Crossentropy

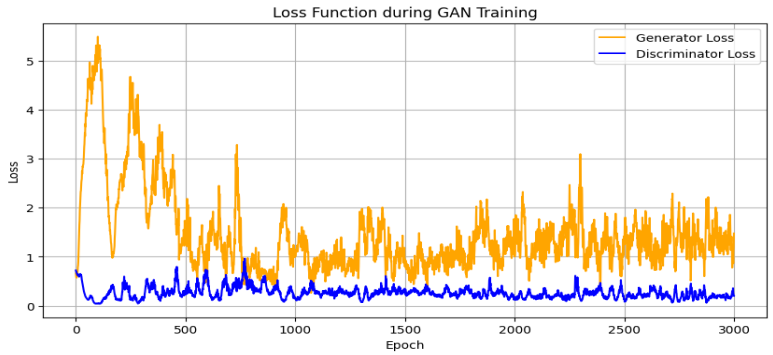


Figure 3. GAN Model Training Loss.

Based on Figure 3, it can be observed that the loss functions of the generator and discriminator change over time during training. At the beginning of training, the generator produces poor-quality data, as evident from the initially high generator loss on the graph. However, as training progresses, the generator learns to generate data that increasingly resembles the real data, resulting in a reduction in the generator's loss function. On the other hand, the discriminator's loss function is initially high as well but decreases as training continues because the discriminator must contend with an increasingly proficient generator. The ultimate goal of GAN training is to achieve a balance where the generator produces data that closely resembles real data and the discriminator struggles to differentiate between the two. This is indicated by both the generator and discriminator loss functions stabilizing at relatively low levels, which occurs at Epoch 3000, with a discriminator loss of 0.176 and a generator loss of 1.83.

The synthetic data generated using GAN is evaluated to determine the extent of its similarity to the real data. To facilitate the comparison between each chemical element and the mechanical properties of stainless steel, both datasets are visualized using kernel density, as seen in Figure 4. The evaluation results indicate that nearly every variable of the chemical elements and mechanical properties has a similar kernel density shape, except for the heating time. This suggests that each variable in the synthetic data has a distribution density that is nearly similar to the real data. The synthetic data produced by GAN can closely resemble real data because GAN learns from the distribution of real data. This learning process allows GAN to grasp the patterns and structure of the real data, resulting in synthetic data with similar characteristics [16].

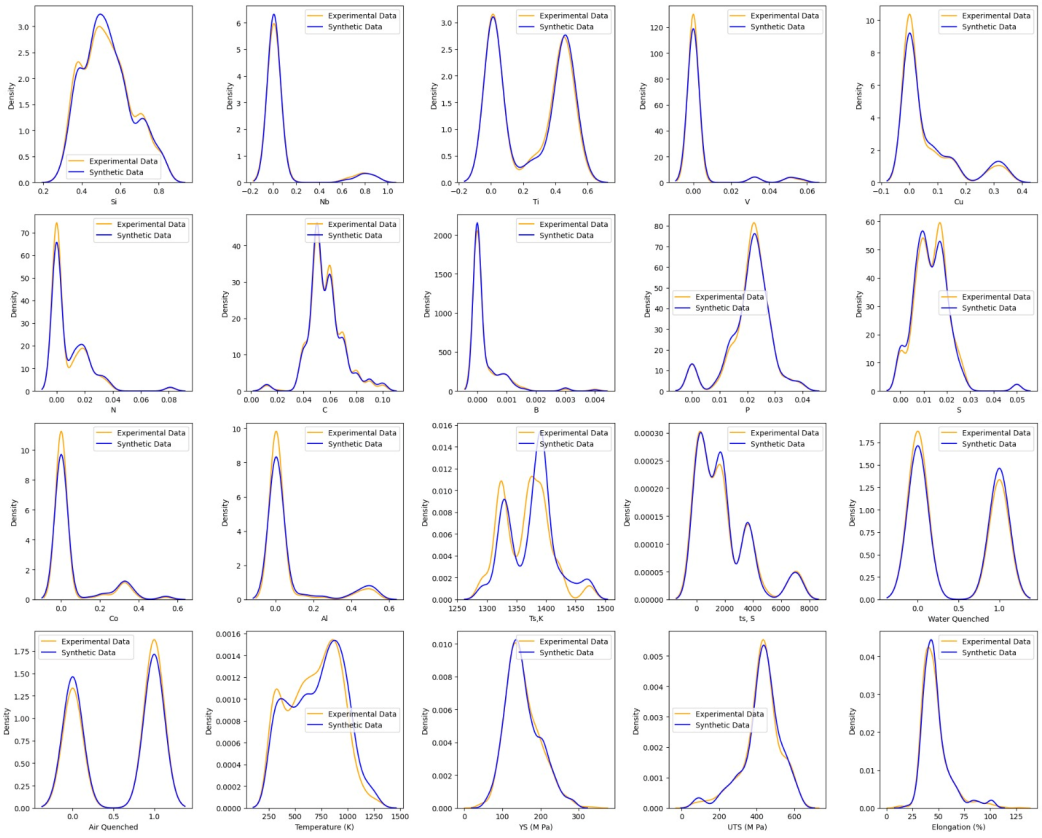


Figure 4. Comparison of Experimental Data with Synthetic Data.

In this study, the Kolmogorov–Smirnov (KS) test was also conducted, which is a statistical method used to test the similarity of distributions between two groups of data or samples [13]. In the context of this research, the KS test is employed to compare the distribution of synthetic data with the distribution of real data. The KS testing process involves comparing the two cumulative distribution functions (CDF) of the two data groups

to be compared, namely, the real data and synthetic data. This test produces two statistics: the KS statistic and the *p* value. The KS statistic measures the degree of deviation, or difference, between the two CDF distributions. The larger the KS statistic, the greater the difference between the two distributions. On the other hand, the *p*-value is an indicator of statistical significance. A low *p*-value reflects a significant difference between the two distributions, while a high *p*-value indicates that the difference is not significant [17]. The results of the KS test for synthetic data compared to real data can be seen in Table 2.

Table 2. Kolmogorov–Smirnov (KS) Test Results for Synthetic Data compared to Real Data.

No.	Variable	KS Statistic	*p*-Value	No.	Variable	KS Statistic	*p*-Value
1	Si	0.025	0.909	11	Co	0.042	0.362
2	Nb	0.019	0.993	12	Al	0.05	0.158
3	Ti	0.08	0.004	13	Ts, K	0.28	0
4	V	0.009	1	14	ts, S	0.032	0.677
5	Cu	0.052	0.143	15	Water Quenched	0.045	0.28
6	N	0.053	0.129	16	Air Quenched	0.045	0.28
7	C	0.03	0.752	17	Temperature (K)	0.03	0.4
8	B	0.011	1	18	YS (M Pa)	0.027	0.854
9	P	0.032	0.677	19	UTS (M Pa)	0.039	0.457
10	S	0.05	0.175	20	Elongation (%)	0.061	0.052

Based on Table 1, it can be observed that the KS statistic for all variables, including chemical elements, heat treatment, and the mechanical properties of austenitic stainless steel, falls within a relatively low range approaching 0, except for the heating time, which has a value of 0.28. These values indicate that the distribution of chemical elements, heat treatment, and the mechanical properties of austenitic stainless steel in synthetic data tends to be similar to the distribution in real data.

3.3. Modeling the Prediction of Austenitic Stainless Steel Mechanical Properties

The synthetic data of austenitic stainless steel, totaling 986 samples, is divided into 80% for training and 20% for testing. In this stage, six machine learning algorithm models are compared, with each model validated using cross-validation with a K value of 10, and then evaluated using three evaluation metrics: MAE, RMSE, and R-squared. The selection of the best model is based on the smallest MAE and RMSE values and the highest R-squared value. The results of the comparison of the six machine-learning algorithms can be seen in Figure 5.

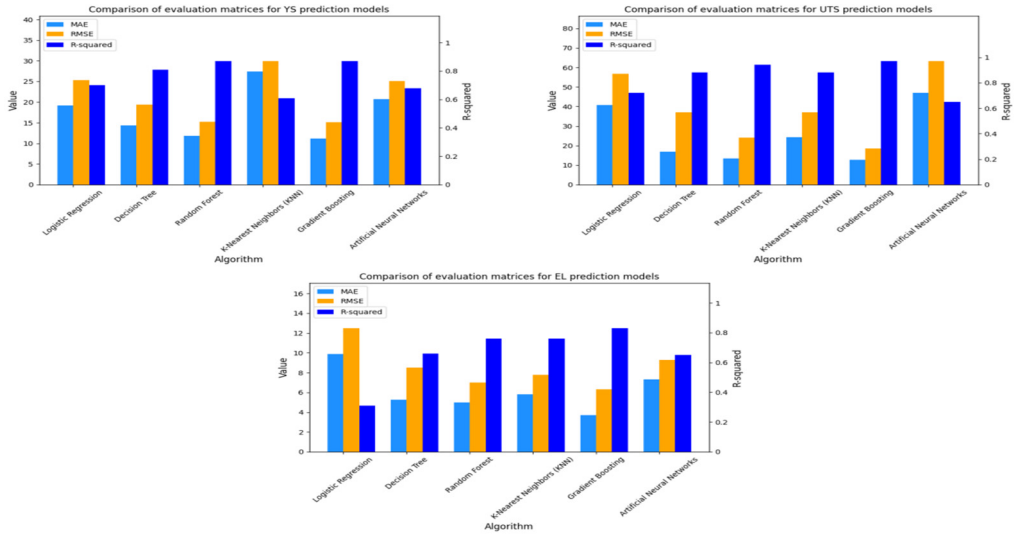


Figure 5. Comparison of Machine Learning Evaluation Metrics.

From the testing results of the three mechanical properties of austenitic stainless steel, the Random Forest and Gradient Boosting algorithms outperform the other four algorithms. Random Forest and Gradient Boosting exhibit lower MAE and RMSE values as well as higher R-squared values. This is evident in the MAE value of 11.21, RMSE of 15.09, and R-squared of 0.85 for YS. For UTS, the model achieves an MAE of 12.68, an RMSE of 18.59, and an R-squared of 0.95, while EL obtains an MAE of 3.71, an RMSE of 6.32, and an R-squared of 0.81. These results indicate low error rates and a strong fit between predictions and actual values for the three mechanical properties of austenitic stainless steel.

3.4. Model Evaluation

The predictive model of austenitic stainless steel mechanical properties using gradient boosting is tested with 200 experimental data points that were not included in the model's training dataset. The purpose of this testing is to assess the extent to which the model can provide accurate predictions for new data. Testing the model with new data also serves other benefits, such as helping identify weaknesses or issues that may arise when the model encounters previously unseen situations. If the model cannot provide accurate predictions or its performance deteriorates when faced with new data, it may indicate that the model struggles to generalize patterns or lacks flexibility in handling data variations.

The results of testing with new data indicate that the model's performance is not significantly different from when it was previously trained. The model can predict new data very effectively, as shown in Figure 6.

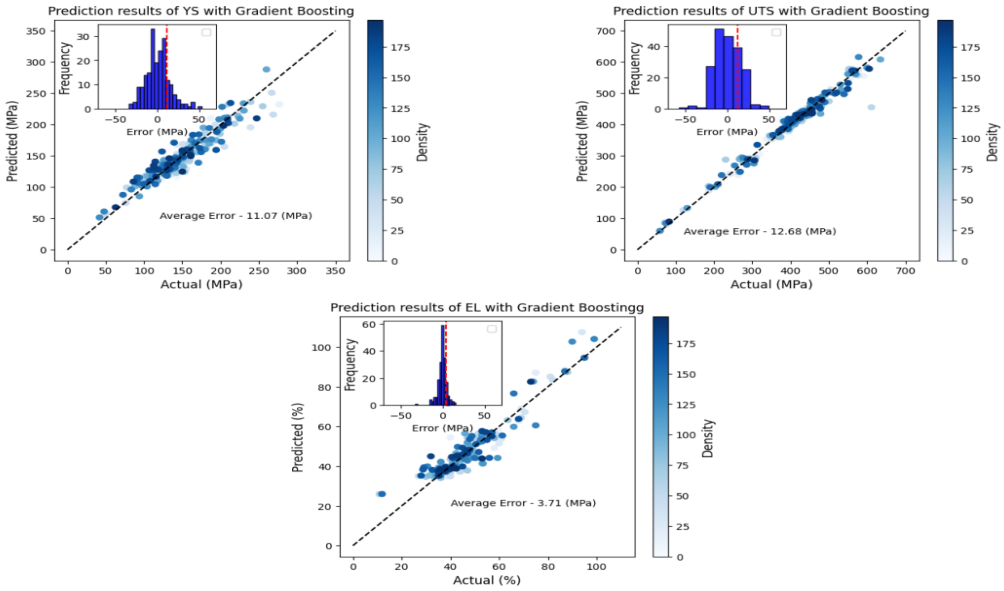


Figure 6. Comparison of Prediction Results with Experimental Data.

4. Conclusions

Based on the results of modeling mechanical properties data augmentation for stainless steel using Generative Adversarial Networks (GAN), it can be concluded that the modeling of mechanical properties data augmentation for stainless steel using Generative Adversarial Networks (GAN) has the potential to generate synthetic data that closely resembles real data. This can be observed through the Kolmogorov–Smirnov (KS) test, which shows that the distribution of synthetic data and experimental data exhibits significant similarity, indicating that synthetic data can effectively represent real data.

The results of predictive model training using synthetic data with six machine learning algorithms indicate that the gradient boosting model performs the best in predicting the mechanical properties of austenitic stainless steel. These findings have positive implications in the context of using GAN to generate synthetic data for material mechanical properties.

High-quality synthetic data that closely resembles real data can be used in various applications, such as modeling and analysis, without being reliant on limited real data availability. Therefore, the use of GAN in material data augmentation presents a promising alternative to support future material research and development.

Author Contributions: Conceptualization, D.L.; methodology, D.S.K. and M.; data analysis and validation, H. and S.A.; writing and proofreading the manuscript, D.L. and D.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data source code can be accessed at https://www.phase-trans.msm.cam.ac.uk/map/data/materials/austenitic.data.html (accessed on 21 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rajan, K. Materials Informatics: The Materials 'Gene' and Big Data. *Annu. Rev. Mater. Res.* **2015**, *45*, 153–169. [CrossRef]
2. Frydrych, K.; Karimi, K.; Pecelerowicz, M.; Alvarez, R.; Dominguez-Gutiérrez, F.J.; Rovaris, F.; Materials, S.P. Informatics for Mechanical Deformation: A Review of Applications and Challenges. *Materials* **2021**, *14*, 5764. [CrossRef]
3. Blaiszik, B.; Ward, L.; Schwarting, M.; Gaff, J.; Chard, R.; Pike, D.; Chard, K.; Foster, I. A data ecosystem to support machine learning in materials science. *MRS Commun.* **2019**, *9*, 1125–1133. [CrossRef]
4. Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the 'fourth paradigm' of science in materials science. *APL Mater.* **2016**, *4*, 053208. [CrossRef]
5. Yu, D.; Zhang, H.; Chen, W.; Yin, J.; Liu, T.-Y. How Does Data Augmentation Affect Privacy in Machine Learning? *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 10746–10753. [CrossRef]
6. Wang, K.; Gou, C.; Duan, Y.; Lin, Y.; Zheng, X.; Wang, F.-Y. Generative adversarial networks: Introduction and outlook. *IEEECAA J. Autom. Sin.* **2017**, *4*, 588–598. [CrossRef]
7. He, G.; Zhao, Y.; Yan, C. Application of tabular data synthesis using generative adversarial networks on machine learning-based multiaxial fatigue life prediction. *Int. J. Press. Vessel. Pip.* **2022**, *199*, 104779. [CrossRef]
8. Marani, A.; Jamali, A.; Nehdi, M.L. Predicting Ultra-High-Performance Concrete Compressive Strength Using Tabular Generative Adversarial Networks. *Materials* **2020**, *13*, 4757. [CrossRef] [PubMed]
9. Sourmail, I.S.T. Materials Algorithms Project Program Library. Available online: https://www.phase-trans.msm.cam.ac.uk/map/data/materials/austenitic.data.html (accessed on 12 July 2023).
10. Agrawal, A.; Deshpande, P.D.; Cecen, A.; Basavarsu, G.P.; Choudhary, A.N.; Kalidindi, S.R. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr. Mater. Manuf. Innov.* **2014**, *3*, 90–108. [CrossRef]
11. Li, D.-C.; Chen, S.-C.; Lin, Y.-S.; Huang, K.-C. A Generative Adversarial Network Structure for Learning with Small Numerical Data Sets. *Appl. Sci.* **2021**, *11*, 10823. [CrossRef]
12. Berger, V.W.; Zhou, Y. Kolmogorov–Smirnov Test: Overview. In *Wiley StatsRef: Statistics Reference Online*, 1st ed.; Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., Teugels, J.L., Eds.; Wiley: Hoboken, NJ, USA, 2014. [CrossRef]
13. Leni, D. Pemilihan Algoritma Machine Learning Yang Optimal Untuk Prediksi Sifat Mekanik Aluminium. *J. Engine Energi Manufaktur Dan Mater.* **2023**, *7*, 35–44. [CrossRef]
14. Leni, D.; Yermadona, H.; Berli, A.U.; Sumiati, R.; Haris, H. Pemodelan Machine Learning untuk Memprediksi Tensile Strength Aluminium Menggunakan Algoritma Artificial Neural Network (ANN). *J. Surya Tek.* **2023**, *10*, 625–632. [CrossRef]
15. Agrawal, A.; Choudhary, A. An online tool for predicting fatigue strength of steel alloys based on ensemble data mining. *Int. J. Fatigue* **2018**, *113*, 389–400. [CrossRef]
16. Fonseca, J.; Bacao, F. Tabular and latent space synthetic data generation: A literature review. *J. Big Data* **2023**, *10*, 115. [CrossRef]
17. Lall, A. Data streaming algorithms for the Kolmogorov-Smirnov test. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 95–104. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.