



PREDICTION OF STAINLESS STEEL MECHANICAL PROPERTIES USING SYNTHETIC DATA GENERATED BY WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS WITH GRADIENT PENALTY

Maimuzar¹, Desmarita Leni², Sir Anderson¹ and Yuda Perdana Kusuma³

¹Department of Mechanical Engineering, Politeknik Negeri Padang, Padang, Indonesia

²Department of Mechanical Engineering, Universiti Muhammadiyah Sumatera Barat, Padang, Indonesia

³Department of Mechanical Engineering, IJince Journal, Universiti Negeri Padang, Padang, Indonesia

E-Mail: maimuzar@pnp.ac.id

ABSTRACT

Machine learning has become one of the most widely developed methods for analyzing, predicting, and simulating the mechanical properties of materials. In practice, experimental data on stainless steel remain limited, especially when considering variations in heat treatment and chemical composition. This limitation poses a major challenge in developing robust predictive models. This study aims to address the scarcity of experimental data by applying a Generative Adversarial Network (GAN) model to generate synthetic data that exhibit characteristics similar to real data. Based on an analysis of 16 chemical variables using Pearson correlation, eight variables were identified as the most influential on mechanical properties, namely Cr, Ni, Mo, Si, Mn, P, N, and Cu. The evaluation results indicate that most variables achieve KS values below 0.10, reflecting a high level of similarity between the distributions of synthetic and real data. This study further evaluates the synthetic data using machine learning prediction models to estimate Yield Strength (YS) and Ultimate Tensile Strength (UTS), where 986 synthetic samples were used as the training set and 100 real experimental samples were randomly selected as the test set. The machine learning models employed include Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Gradient Boosting, and Artificial Neural Network. The results show that Gradient Boosting and K-Nearest Neighbors achieved the best performance, with R^2 values of 0.75 and 0.69, respectively. These findings indicate that the WGAN-GP-based approach can effectively overcome limitations in experimental data availability while accelerating the development and prediction of stainless steel mechanical properties.

Keywords: generative adversarial networks, synthetic data, mechanical properties, stainless steel, machine learning.

Manuscript Received 4 April 2026; Revised 19 April 2026; Published 10 July 2026

INTRODUCTION

The mechanical properties of stainless steel, such as yield strength and tensile strength, are critical parameters in the design and development of engineering components. This type of steel is widely used in applications requiring high corrosion resistance and mechanical durability [1] [2] [3]. Currently, testing of these mechanical properties is still predominantly performed manually using tensile testing machines. Such conventional approaches require considerable time, high operational costs, and significant material consumption. Repeated experimental testing on stainless steel samples with varying chemical compositions and heat treatments further increases the overall testing effort in terms of both cost and time [4] [5].

Alongside advances in materials science, data-driven approaches have begun to gain widespread adoption. This development has given rise to the field of materials informatics, which integrates experimental data, computational simulations, and artificial intelligence techniques such as machine learning. The goal is to accelerate material understanding and design by leveraging available data sources [6] [7] [8]. In practice, machine learning has demonstrated the ability to construct predictive models of material properties, including

stainless steel, with relatively high accuracy [9][10]. However, the effectiveness of these models is highly dependent on the availability of sufficiently large and representative datasets. In reality, experimental data for stainless steel remain limited, particularly when covering various heat treatments and chemical compositions [11] [12]. This limitation presents a major challenge in developing robust predictive models.

One promising solution is the use of synthetic data. Generative Adversarial Networks (GANs) allow the creation of synthetic datasets that exhibit characteristics similar to real data. Several studies have applied GANs in different material-related contexts, such as the design of metallic glass compositions [13], bearing damage prediction [14] [15] [16], and fatigue life estimation of materials [17] [18]. However, most of these studies employ synthetic data merely as supplementary augmentation to real datasets. Moreover, validation of model performance trained solely or predominantly on synthetic data has often not been conducted using comprehensive external evaluation.

Various studies have utilized GANs in material data processing to address the scarcity of experimental datasets. However, most of these studies still employ synthetic data merely as supplementary information rather



than as the sole source for training predictive models. Study [13], for example, used GANs to generate synthetic compositions of metallic glasses as a solution to data shortages in materials design. Although the generated data exhibited distinct and interesting statistical distributions compared to real data, the study was limited to producing synthetic samples and did not integrate them into predictive modeling.

Similarly, study [14] employed GAN-generated synthetic data to augment the training set for bearing damage prediction. The results showed improvements in model performance, yet the synthetic data functioned only as an additional source rather than the primary training dataset. Study [17] applied GAN-based synthetic data to support the prediction of multiaxial fatigue life across different materials. While the synthetic data improved predictive accuracy, they were still used as complementary inputs and were not evaluated as the sole training source. Moreover, external validation using independent experimental data was not conducted. Study [19] generated millions of synthetic chemical samples using GANs, achieving a validity rate of up to 84.5% according to chemical rules. However, these synthetic samples were used only for statistical validation and were not applied to train or test predictive models.

This study proposes a different and more comprehensive approach. The synthetic data generated by GANs are used as the sole training dataset for developing a predictive model of stainless steel mechanical properties. The resulting model is then evaluated using real experimental data as external validation. This approach is

expected to address data scarcity while demonstrating the effectiveness of synthetic data as a potential alternative to experimental data in the development of predictive models.

METHODOLOGY SECTION

This study adopts a Synthetic Data-Driven Machine Learning approach, in which synthetic data are utilized to train machine learning models to improve predictive accuracy under conditions of limited real experimental data. The research begins with collecting chemical composition data and mechanical properties of stainless steel, covering the key variables relevant to the material. These data are then used to train a Generative Adversarial Network (GAN) model with the objective of producing synthetic data that exhibit similar distributions and inter-variable correlations to those of the original dataset.

The synthetic data generated by the GAN are subsequently employed as the training dataset for machine learning models. These models are tested using new data, specifically experimental measurements of stainless steel mechanical properties, to evaluate their accuracy and reliability. Model performance is assessed using evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2), which provide a quantitative measure of predictive accuracy and model performance. Figure-1 illustrates the overall stages of the methodology employed in this research

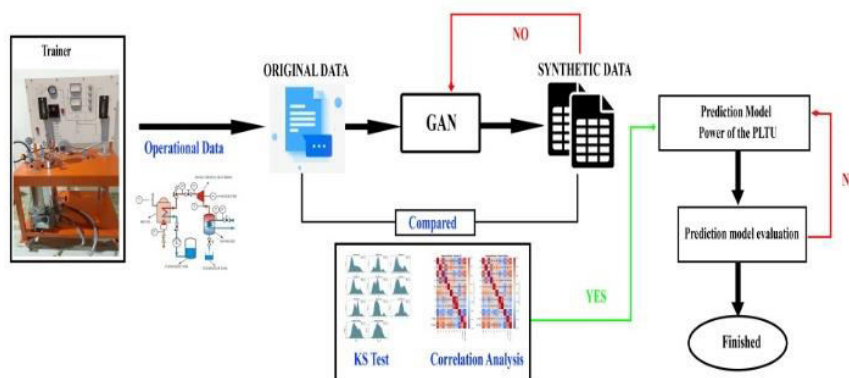


Figure-1. Research framework and methodology.

The data used in this study originate from two main sources. First, the training data were collected from various material databases, including the Open Quantum Materials database, Materials Project, Computational Materials Repository, Anorganic Crystal Structure Database, MatWeb, MakeItFrom, MatDat, as well as experimental datasets reported in scientific publications. These datasets include information on chemical composition, heat treatment conditions, and mechanical properties of stainless steel. Second, the experimental data were obtained directly through tensile testing and analysis

using a Scanning Electron Microscope equipped with Energy Dispersive X-ray Spectroscopy (SEM-EDX) on stainless steel grades 304, 304L, 316, and 316L.

RESULTS AND DISCUSSIONS

a) Data Collection

The dataset used in this study consists of tensile test results from several types of austenitic stainless steel (ASS), including SUS 304, SUS 316, SUS 321, SUS 347, and NCF 800H. The original dataset contains 2,180 entries



comprising mechanical properties, alloying element compositions, heat treatment temperatures, and cooling methods. However, after data preprocessing, 1,194 samples were excluded due to missing values, incomplete information, or invalid entries and therefore were not included in this research. The original database also contained additional features such as melting method, grain size, and product form. Nevertheless, these variables were removed because they were incomplete and showed very low correlation with the mechanical properties of austenitic stainless steel. Only 986 samples with complete and relevant information were used in this study, consisting of 16 input variables and 2 output variables.

b) Data Analysis

In this stage, a Pearson correlation analysis was performed to examine the linear relationship between numerical variables, specifically between the chemical

composition of stainless steel and its mechanical properties, namely Yield Strength (YS) and Ultimate Tensile Strength (UTS). The purpose of this analysis is to identify whether changes in the concentration of specific alloying elements correspond to linear changes in mechanical properties [4]. The Pearson correlation coefficient ranges from -1 to +1 [5]. A positive coefficient indicates a direct relationship, meaning that an increase in the concentration of an alloying element tends to be followed by an increase in mechanical properties. Conversely, a negative coefficient indicates an inverse relationship, where increasing the concentration of a particular element results in a decrease in mechanical strength. A coefficient close to zero suggests a very weak or nonexistent linear relationship between the variables. The results of the correlation analysis between chemical elements and the mechanical properties of stainless steel are shown in Figure-2.

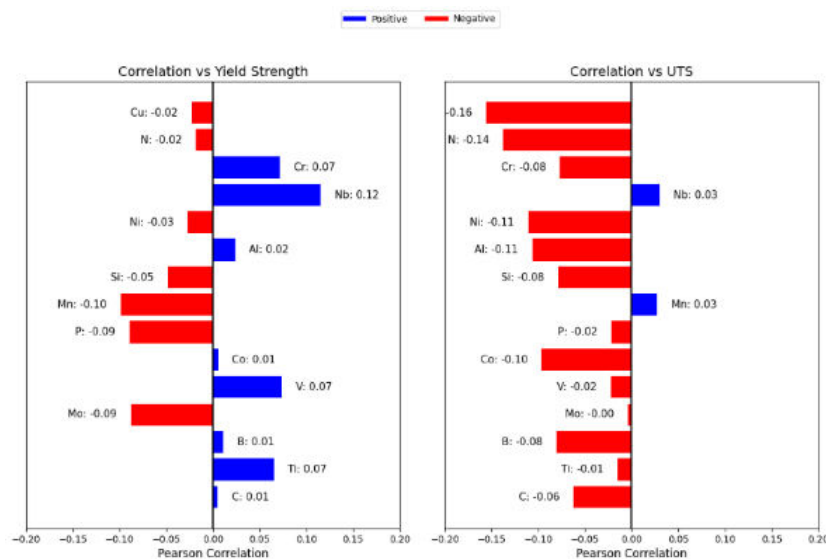


Figure-2. Pearson correlation analysis of alloying elements with mechanical properties.

Overall, the Pearson correlation values for all chemical elements fall around 0.2, indicating that the linear relationship between chemical composition and mechanical properties of stainless steel is relatively weak. This suggests that the influence of alloying elements does not act independently but is instead multivariate and nonlinear due to complex interactions among alloying constituents, phase distribution changes, and thermomechanical effects during material processing. This observation is consistent with previous studies [6], which emphasize that the relationship between chemical composition and mechanical properties in stainless steel cannot be adequately represented through simple linear correlation. The interactions between alloying elements can alter the ferrite-austenite phase balance, modify solid-solution hardening mechanisms, and influence precipitation kinetics during heat treatment. Given these conditions, machine learning-based predictive models

become highly relevant. Methods such as Random Forest and Gradient Boosting are better suited to capture nonlinear interactions and multivariate patterns that cannot be identified through conventional statistical analysis. These models can utilize element interactions to form more complex decision structures, resulting in higher predictive accuracy for materials systems such as stainless steel.

At this stage, feature selection was also conducted to identify the chemical elements that have the most significant influence on mechanical properties, particularly Yield Strength (YS) and Ultimate Tensile Strength (UTS). Feature selection is essential because stainless steel is a multicomponent alloy in which each element contributes differently to strength, toughness, phase stability, and mechanical behavior [20] [21]. Identifying the most influential elements enables the research to focus on the most relevant variables, making



predictive models more efficient, less prone to overfitting, and easier to interpret. Additionally, feature selection ensures that the synthetic data generated by GAN accurately represents the chemical characteristics typically found in industrial stainless steel, making it suitable for use as training data. In this study, feature selection was performed using the feature importance approach from the Random Forest algorithm. The results of the feature selection process are presented in Figure-3.

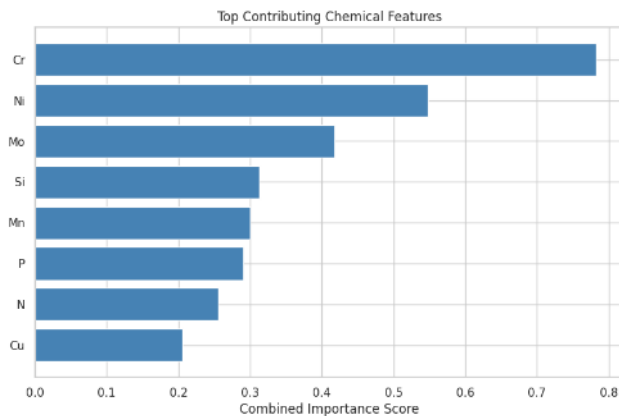
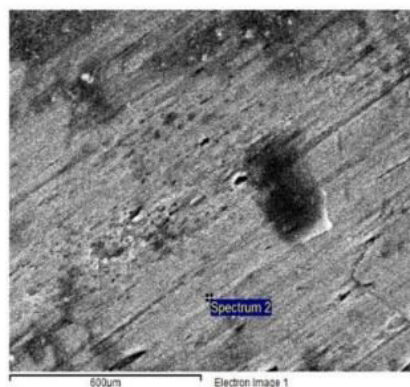
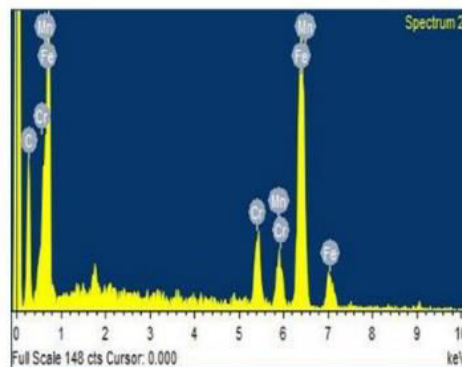


Figure-3. Feature importance results.

The analysis indicates that eight chemical elements exhibit the highest contribution scores to both mechanical properties, namely Cr, Ni, Mo, Si, Mn, P, N, and Cu. These eight elements were selected because they consistently showed high importance values ranging from 0.25 to 0.80, as illustrated in the Top Contributing Chemical Features graph. The feature importance results suggest that Chromium, Nickel, and Molybdenum are the most influential elements affecting the mechanical strength of stainless steel, followed by Silicon, Manganese, Phosphorus, Nitrogen, and Copper. The SEM-EDX analysis presented in Figure 4 further supports these findings by revealing the presence of key alloying elements such as Fe, Cr, Ni, Mn, and trace amounts of Mo. These elements are characteristic of stainless steel in the 300-series, particularly grades 304 and 316, which are widely known for their combination of high corrosion resistance and good mechanical strength. The dominant peaks of Cr and Ni in the EDX spectrum confirm that the tested material belongs to the austenitic stainless steel family, where Cr plays a critical role in forming the passive layer responsible for corrosion protection, while Ni enhances austenite stability and improves ductility.



(a)



(b)

Figure-4. (a) SEM image of ASS 201 and (b) EDX spectrum at the analysis point.

These findings are consistent with the results of the machine learning-based feature selection, which identified Cr, Ni, Mo, Si, Mn, P, N, and Cu as the eight most influential chemical features affecting the mechanical properties. Interestingly, most of the elements detected in the EDX analysis also belong to the group of features with the highest importance scores in the model.

The selection of these eight elements forms a critical foundation for the development of the predictive model, as they represent the combination of chemical factors most relevant to variations in the mechanical behavior of the material. The distribution of these eight selected chemical elements is presented in Figure-5.

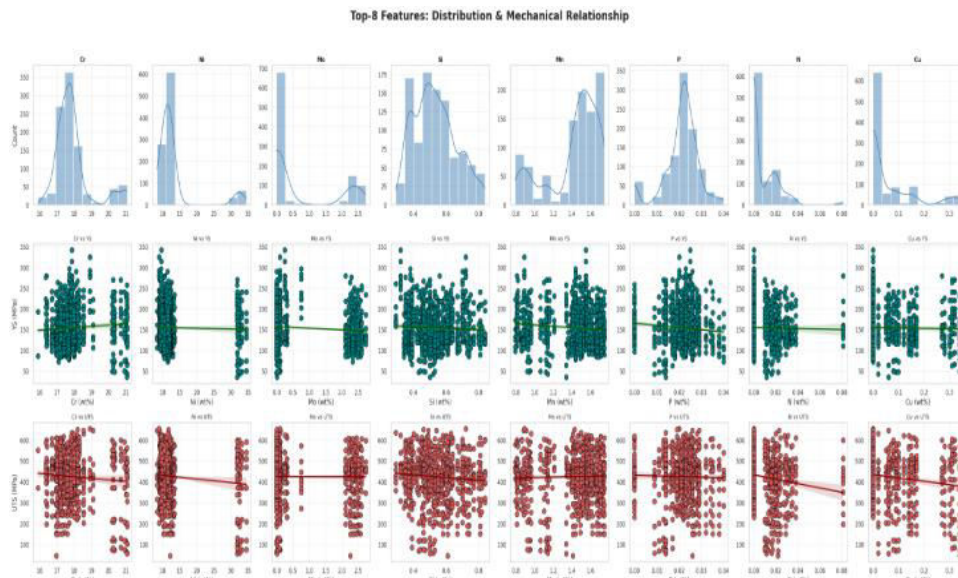


Figure-5. Distribution of chemical elements selected through feature importance analysis.

c) Synthetic Data Generation Using GAN

In this stage, the dataset that has undergone feature selection based on feature importance analysis is used as the primary input for generating synthetic data using a Generative Adversarial Network (GAN). The model employed in this study is specifically the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP), a more stable and accurate variant of GAN. WGAN-GP incorporates the Wasserstein distance and gradient penalty, enabling the synthetic data to follow the real data distribution more effectively. The chemical elements used as input variables consist of Cr, Ni, Mo, Si, Mn, P, N, and Cu, based on previous evaluations showing that these eight elements have the most significant influence on the mechanical properties of

stainless steel, particularly Yield Strength (YS) and Ultimate Tensile Strength (UTS).

WGAN-GP was selected due to its superior capability compared with conventional GANs in producing stable and realistic synthetic data. This approach addresses key limitations of classical GANs, such as training instability and mode collapse, by introducing a Wasserstein distance-based loss function and a gradient penalty term to maintain balance between the generator and the critic during training [22], [23]. By applying WGAN-GP, the model is expected to more effectively learn multivariate distribution patterns across chemical compositions and mechanical properties, allowing the generated synthetic data to not only match the mean and variance of experimental data but also preserve the underlying relationships between variables.

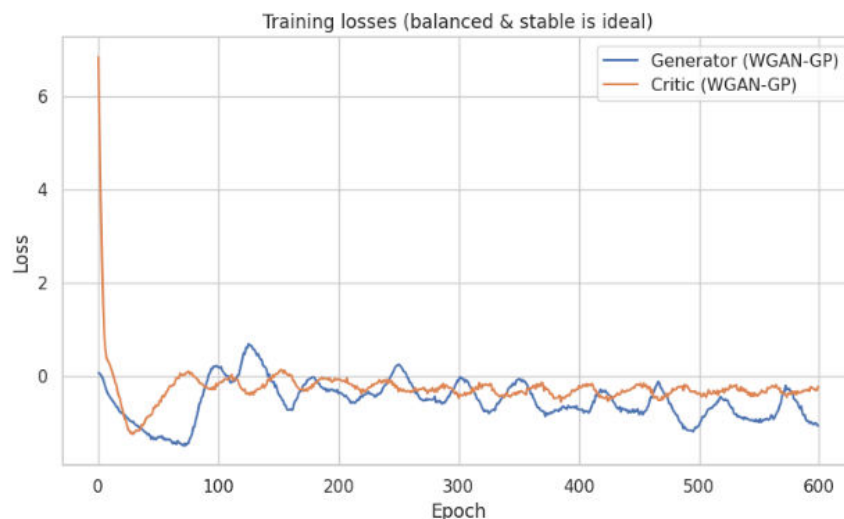


Figure-6. Training losses during synthetic data generation using WGAN-GP.



Based on Figure-6, the WGAN-GP model was trained for 600 epochs using two main components, the Generator and the Critic. The training loss curves show that, at the beginning, the critic loss is relatively high, indicating that the discriminator can clearly differentiate between real and synthetic data. However, as training progresses, the losses of both components decrease and begin to fluctuate around zero. This pattern indicates that the model has reached a stable equilibrium between the generator and the critic, meaning that the training process is stable. Such behavior also suggests that the generator has successfully learned the real data distribution. This result is consistent with previous findings reporting that WGAN-GP provides more stable training dynamics than conventional GAN due to the inclusion of a gradient penalty [23]. Overall, the balanced loss curves between the generator and critic demonstrate that the WGAN-GP model has been properly trained and is ready to generate synthetic data with distribution characteristics similar to real experimental data.

d) Synthetic Data Evaluation

After the WGAN-GP training process was completed, the next step was to evaluate whether the generated synthetic data truly resembled the characteristics of the original dataset. This stage is crucial because the quality of synthetic data determines whether it can serve as a reliable substitute or complement to experimental data in

building predictive models. The evaluation was conducted through two main approaches, namely distribution analysis and the Kolmogorov-Smirnov (KS) statistical test for each variable [24]. Overall, this evaluation process ensures that the synthetic data not only approximates the distribution of experimental data but also preserves consistent statistical structures. This validation step is essential before using synthetic data for model training because the quality of the input data directly influences the accuracy and generalization capability of the machine learning models developed.

i) Distribution Analysis

Distribution analysis was performed to compare the shape and behavior of the synthetic and experimental datasets through histogram and kernel density estimation (KDE) visualizations. This approach is commonly used in synthetic data research because it reveals similarities in distribution patterns, probability density profiles, and the presence of outliers [25], [26]. When the KDE curve of the synthetic data closely follows that of the real dataset, the model is considered successful in capturing the underlying distribution of the features. In addition to visual inspection, a quantitative evaluation was conducted using the KS test for each chemical variable and mechanical property to assess statistical similarity between the two datasets.

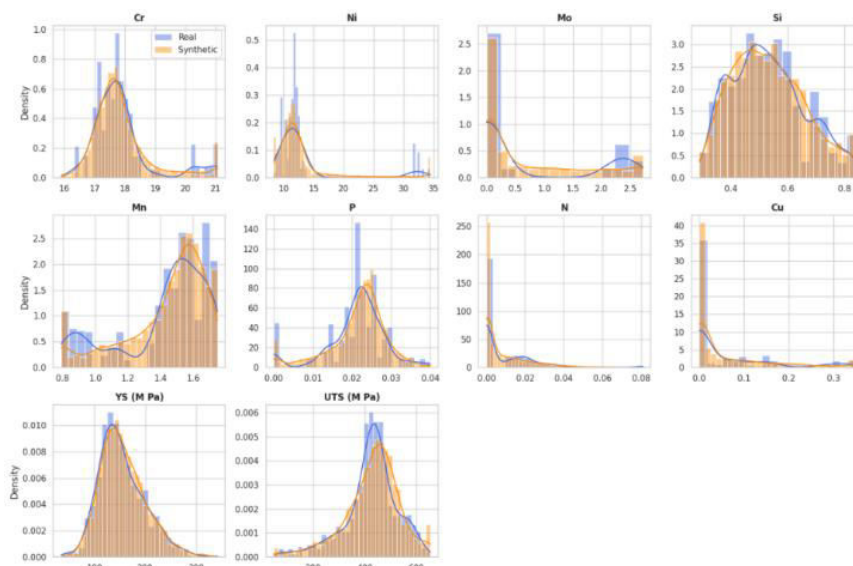


Figure-7. Comparison of synthetic and experimental data distributions.

Figure-6 presents a comparison of the distribution between real and synthetic data for eight chemical elements (Cr, Ni, Mo, Si, Mn, P, N, and Cu) as well as two mechanical properties (YS and UTS). Overall, the orange curves (synthetic data) closely follow the distribution patterns of the blue curves (real data), indicating a good level of agreement. Elements such as Cr, Si, Mn, P, and Mo exhibit strong similarity in distribution shape,

suggesting that the WGAN-GP model successfully learned the key statistical characteristics of the experimental dataset. For N and Cu, slight deviations are observed, particularly in the lower value regions, implying that the natural variability of these minor elements has not been fully replicated by the model. Meanwhile, for YS and UTS, the distributions of synthetic and real data nearly overlap, demonstrating that the model effectively captured



the relationships between chemical composition and mechanical properties in a consistent manner.

ii) Kolmogorov-Smirnov (KS) Test

The KS test was selected because it is non-parametric and effective for assessing the similarity between two distributions without assuming any specific distributional form, making it widely used for synthetic data validation in materials engineering and manufacturing

domains [27]. A smaller KS value indicates that the maximum difference between the real and synthetic distributions is minimal, implying that the generated data exhibit good similarity to the experimental data [28]. Therefore, the combination of visual inspection and statistical testing provides a comprehensive assessment of the WGAN-GP model's ability to replicate the characteristics of real data.

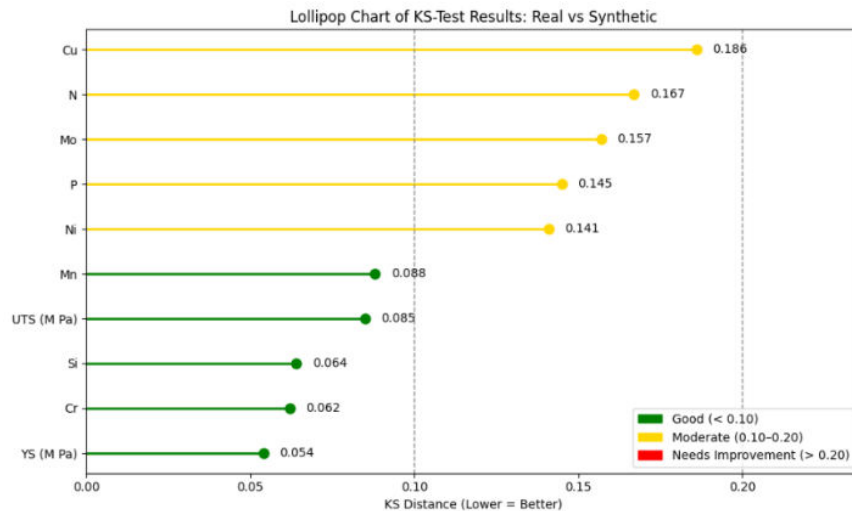


Figure-8. Results of the KS Test.

Based on Figure-8, the KS test results show varying levels of similarity between real and synthetic distributions. Variables with $KS < 0.10$ fall into the good similarity category, including Cr, Si, Mn, YS, and UTS. KS values between 0.10 and 0.20, as observed for Ni and P, are classified as moderate, indicating minor differences in distribution shape that are still within acceptable ranges. Meanwhile, variables such as Mo, N, and Cu exhibit KS values greater than 0.20, indicating the need for improvement. This suggests that the model requires further refinement to better capture the natural variability of these minor alloying elements. Both nitrogen (N) and copper (Cu) are minor elements in stainless steel with relatively narrow concentration ranges, making their

distributions more challenging to represent accurately using generative models.

A previous study [29] reported that nitrogen in austenitic stainless steel typically ranges from 0.03 to 0.25 percent depending on the grade, and plays a critical role in strengthening through solid-solution mechanisms. However, because the range is narrow, its distribution is often multimodal or deviates from normality, which generative algorithms such as GANs tend to struggle to reproduce. This aligns with findings from [30], which highlight that GAN models often face difficulties when reconstructing low-frequency variables or features that occupy narrow value ranges. As a result, the synthetic distributions for these elements tend to show slight shifts relative to the real data.

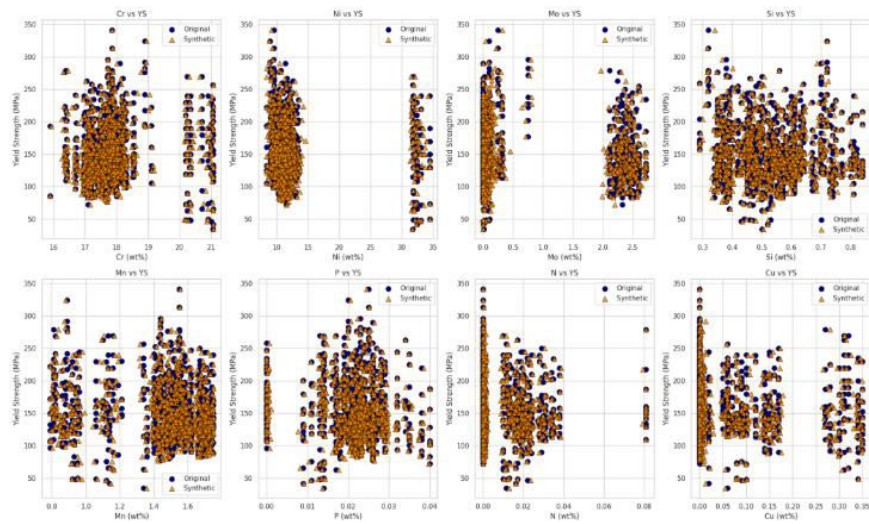


Figure-9. Relationship between chemical variables and Yield Strength (YS) in real and synthetic data.

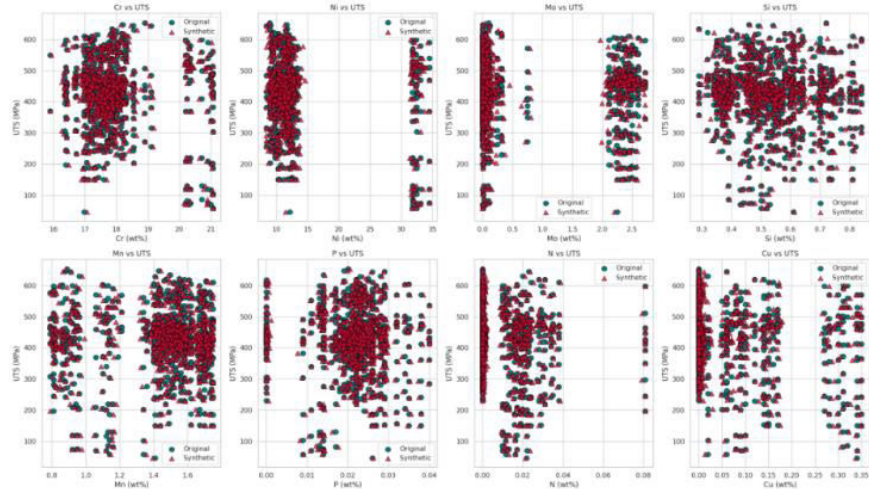


Figure-10. Relationship between chemical variables and Ultimate Tensile Strength (UTS) in real and synthetic data.

These results are further supported by Figures 9 and 10, where the visual scatter patterns of the synthetic data closely follow those of the real data across almost all variables. This demonstrates that the WGAN-GP model successfully learned the nonlinear relationships between chemical composition and mechanical properties of stainless steel. The clustering of synthetic samples around regions of dense real data indicates that the model does not merely replicate global distribution shapes but also captures local correlation structures, particularly for variables such as Cr, Ni, Si, and Mn, which are known to exert significant influence on YS and UTS.

e) Integration of Synthetic Data into the Predictive Model

In this stage, the synthetic data generated by the GAN model were used as the training dataset for constructing predictive models of mechanical properties.

This approach was implemented to evaluate the extent to which synthetic data can effectively replace experimental data in machine learning workflows. A total of 986 synthetic samples were used as the training set, while 100 experimental samples were randomly selected from the dataset and assigned as the test set. This separation ensures that the model is evaluated using completely unseen data, allowing the resulting predictive performance to more accurately represent the model's generalization capability. Such an evaluation scheme is widely adopted in materials informatics studies [31][32], particularly when experimental data are scarce or costly to obtain.

In this phase, six machine learning algorithms were compared, including Linear Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Artificial Neural Network (ANN). These algorithms were implemented using default parameters from scikit-learn



version 1.2.2 and Keras version 2.12.0 for the ANN model. The available data were divided into an 80% training set and a 20% testing set. Each machine learning model was validated using cross-validation, a technique that divides the training data into multiple subsets or folds and iteratively trains and tests the model on each fold [10][11].

For the predictive modeling of the mechanical properties of stainless steel, a 10-fold cross-validation scheme was utilized. In this approach, the data were split into 10 different folds, and the model was iteratively trained and evaluated 10 times, with each fold serving once as the validation data while the remaining folds were used for training. The performance comparison of all models is summarized in Table-1.

Table-1. Comparison of prediction models for mechanical properties of stainless steel.

Model	Target	MAE	RMSE	R ²
Linear Regression (LR)	YS (M Pa)	31.646	40.985	0.117
	UTS (M Pa)	69.583	95.742	0.134
Decision Tree (DT)	YS (M Pa)	26.738	35.229	0.348
	UTS (M Pa)	51.106	74.053	0.482
Random Forest (RF)	YS (M Pa)	15.724	23.073	0.720
	UTS (M Pa)	34.980	57.553	0.687
K-Nearest Neighbors (KNN)	YS (M Pa)	12.329	21.918	0.748
	UTS (M Pa)	32.193	58.137	0.681
Gradient Boosting Tree (GBT)	YS (M Pa)	12.325	21.548	0.756
	UTS (M Pa)	32.005	56.905	0.694
Artificial Neural Network (ANN)	YS (M Pa)	13.646	22.381	0.737
	UTS (M Pa)	33.929	57.156	0.692

The evaluation results indicate that the Linear Regression model exhibits the lowest performance, with R² values of only 0.117 for YS and 0.134 for UTS. This suggests that the relationship between chemical composition and mechanical properties is highly nonlinear, making linear models unable to capture the underlying complexity of the interactions among variables. Performance improves significantly with the Decision Tree model, which achieves R² values of 0.348 for YS and 0.482 for UTS, demonstrating its ability to model basic nonlinear patterns. The best performance is obtained from the group of models based on ensemble learning and nonlinear regression, namely RF, KNN, GBT, and ANN.

Both KNN and ANN show strong and stable predictive performance, with R² values exceeding 0.68 for both targets, indicating that nonlinear data-driven approaches are highly effective for complex material systems such as stainless steel. Among all models, Gradient Boosting Tree (GBT) provides the most optimal results, achieving R² values of 0.756 for YS and 0.694 for UTS, along with the lowest MAE and RMSE scores. This confirms that GBT is capable of capturing intricate interactions among chemical elements and yields superior generalization on unseen testing data. The comparison between actual and predicted values using the GBT model is presented in Figure-11.

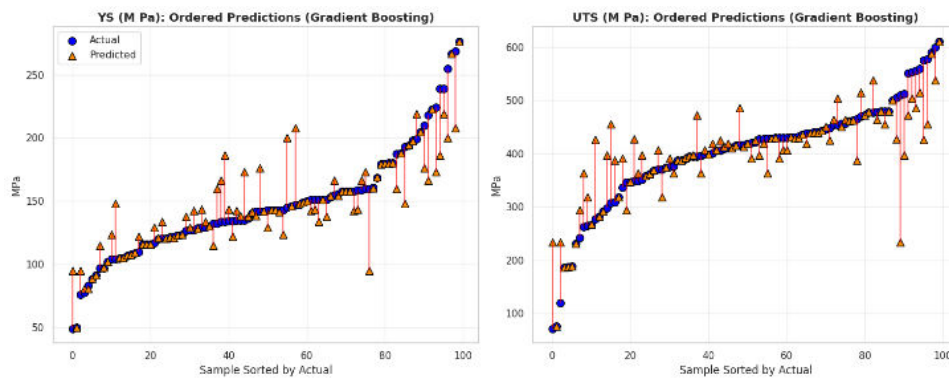


Figure-11. Comparison between actual and predicted values.

These findings are consistent with metallurgical studies that state chromium, nickel, and molybdenum exhibit synergistic effects on phase stability and strengthening mechanisms, whereby even small variations in composition may lead to disproportionately large changes in mechanical response [33]. Study [34] further explains that alloying elements in stainless steel often exhibit cross-effects, meaning that simple linear models are insufficient to capture the multivariate and highly interactive relationships among variables. Model performance begins to improve significantly when nonlinear approaches are applied, indicating that these models are capable of capturing the fundamental nonlinear patterns between chemical composition and mechanical properties. The most substantial improvement appears in ensemble learning methods, which are specifically designed to handle complex multivariate data, including nonlinear effects and cross-interactions among alloying elements. This observation is in line with study [35], which emphasizes that ensemble-based models and neural networks are far more adaptive in learning nonlinear behavior compared to linear models or single decision trees.

This research provides meaningful contributions to the field of materials engineering, particularly in the utilization of synthetic data to enhance the analysis and prediction of stainless steel mechanical properties. The results confirm that WGAN-GP is capable of generating synthetic data with distributional characteristics highly similar to experimental data, making it suitable as a replacement or complement when real datasets are limited. This is especially relevant considering that physical testing such as tensile tests or SEM-EDX characterization is often constrained by cost, time, and limited laboratory access. With access to representative synthetic datasets, material development can be conducted more efficiently without the need for repeated destructive or non-destructive testing. More broadly, this study demonstrates that the use of GANs is not merely a tool to address data scarcity but also a strategic approach to accelerate alloy design and material optimization. When supported by robust machine learning models, this approach paves the way for faster, more cost-effective, and highly adaptive predictive systems for material development in both laboratory and

industrial settings. Thus, this work not only provides methodological contributions but also opens new opportunities for broader applications in stainless steel development and other engineering materials.

Despite the promising results, several limitations should be acknowledged. The limited amount of experimental data constrains the generative model, particularly for minor alloying elements such as N and Cu, which naturally occur in narrow ranges in stainless steel. This contributes to the relatively higher KS values observed for these variables. Future research could expand the diversity of experimental data to provide the generative model with a richer representation and improve the synthetic distribution for minor elements. The adoption of more advanced generative models such as Diffusion Models or VAE-GAN may further enhance the stability and quality of synthetic data. Additional validation using microstructural information such as SEM images or fractography analysis could also strengthen the linkage between chemical composition, microstructure, and mechanical properties. Furthermore, predictive modeling performance may be improved through systematic hyperparameter tuning to enhance consistency across scenarios. Overall, this synthetic-data-driven approach holds significant potential for application across other material systems and opens new pathways for accelerating material design in a more efficient and economically viable manner.

CONCLUSIONS

The results of this study show that the integration of Generative Adversarial Networks, particularly the WGAN-GP model, is capable of generating synthetic data that closely resembles experimental data for predicting the mechanical properties of stainless steel. From the initial set of 16 chemical variables, feature selection based on Pearson correlation and feature importance identified eight key elements, namely Cr, Ni, Mo, Si, Mn, P, N, and Cu, as the most influential variables affecting yield strength and ultimate tensile strength. The distribution similarity evaluation using the Kolmogorov-Smirnov test indicates that several variables, including Cr, Si, Mn, YS, and UTS, obtained KS values below 0.10, which reflects a high level of similarity between synthetic and experimental data. The



synthetic data were then used as training input for machine learning models, and the results show that nonlinear models perform much better than linear approaches. Linear Regression produced low R-squared values of 0.117 for YS and 0.134 for UTS, which confirms that the relationship between chemical composition and mechanical properties is nonlinear and cannot be captured by linear models. Tree-based and ensemble models provided consistent improvements. Gradient Boosting achieved the best performance with R-squared values of 0.756 for YS and 0.694 for UTS, along with the lowest MAE and RMSE values. Good performance was also recorded for ANN and KNN, both obtaining R-squared values above 0.68, which supports the effectiveness of nonlinear data-driven models for analyzing the complex behavior of stainless steel. Overall, this study shows that the use of WGAN-GP is a promising solution to overcome the limited availability of experimental data, while also improving the predictive performance of models used to estimate the mechanical properties of stainless steel.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to the Directorate of Research and Community Service, Directorate General of Research and Development, for the financial support provided through the DPPM Funding Program 2025. Sincere appreciation is also extended to P3M Politeknik Negeri Padang for facilitating this research under grant number 829/PL9.15/AL.04/2025, as well as to all parties who have contributed to the successful completion of this study.

REFERENCES

- [1] Kim K. T. 2022. Mechanical performance of additively manufactured austenitic 316L stainless steel. *Nuclear Engineering and Technology*. 54(1): 244-254.
- [2] Théodore J., Girault B., Couturier L., Gloaguen D., Bertrand E., Paillard P. and Courant B. 2025. Investigation of the residual stresses and strains, microstructure and mechanical properties of stainless steel multipass welds produced using dual wire-tungsten inert gas process. *Materials*. 40, 102403.
- [3] Zheng X., Zhao K., Jing J. and Li H. 2025. Microstructure, mechanical properties and corrosion resistance of tungsten inert gas welded joints of selective laser melting 304L stainless steel. *Materials Today Communications*. 46, 112533.
- [4] Xie Q., Suvarna M., Li J., Zhu X., Cai J. and Wang X. 2021. Online prediction of mechanical properties of hot rolled steel plate using machine learning. *Materials & Design*. 197, 109201.
- [5] Zhou D., Xu K., Lv Z., Yang J., Li M., He F. and Xu G. 2022. Intelligent manufacturing technology in the steel industry of China: a review. *Sensors*, 22(21): 8194.
- [6] Agrawal Ankit, Choudhary Alok. 2016. Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *Apl. Materials*. 4.5: 053208.
- [7] Frydrych K., Karimi K., Pecelerowicz M., Alvarez R., Dominguez-Gutiérrez F. J., Rovaris F. and Papanikolaou S. (2021). Materials informatics for mechanical deformation: A review of applications and challenges. *Materials*. 14(19): 5764.
- [8] Wu S., Otake Y., Higo Y. and Yoshida I. 2023. Pathway to fully data-driven geotechnics: Lessons from materials informatics. *arXiv preprint arXiv:2312.00581*.
- [9] Wei J., Chu X., Sun X. Y., Xu K., Deng H. X., Chen J. and Lei M. 2019. Machine learning in materials science. *Info Mat*. 1(3): 338-358.
- [10] Wang Z., Yang C., Gao N., Wu X. and Yao Z. 2024. Machine learning-assisted prediction of yield strength in irradiated type 316 stainless steels. *Fusion Engineering and Design*. 208, 114691.
- [11] Maimuzar M., Hendra H., Khan S., Leni D. and Islahuddin I. 2024. Optimizing prediction of stainless steel mechanical properties with random forest: a comparison of feature selection methods. *Jurnal Polimesin*. 22(5): 498-507
- [12] Leni D., Karudin A., Abbas M. R., Sharma J. K. and Adriansyah A. 2024. Optimizing stainless steel tensile strength analysis: through data exploration and machine learning design with Streamlit. *EUREKA: Physics and Engineering*. (5): 73-88.
- [13] Xu X. and Hu, J. 2023. A Generative Adversarial Network (GAN) based efficient sampling method for inverse design of metallic glasses. *Journal of Non-Crystalline Solids*. 613, 122378.
- [14] Bhavsar K., Vakharia V., Chaudhari R., Vora J., Pimenov D. Y. and Giasin K. 2022. A comparative study to predict bearing degradation using discrete wavelet transform (DWT), tabular generative adversarial networks (TGAN), and machine learning models. *Machines*. 10(3): 176.



- [15] Hei Z., Sun W., Yang H., Zhong M., Li Y., Kumar A. and Zhou Y. 2025. Novel domain-adaptive Wasserstein generative adversarial networks for early bearing fault diagnosis under various conditions. *Reliability Engineering & System Safety*. 110847.
- [16] Chen, X. and Chen, Z. 2025. Cross-domain fault diagnosis of induction motor based on an improved unsupervised generative adversarial network and fine-tuning under limited labeled data. *Measurement*, 116988.
- [17] He G., Zhao Y., and Yan C. 2022. Application of tabular data synthesis using generative adversarial networks on machine learning-based multiaxial fatigue life prediction. *International Journal of Pressure Vessels and Piping*. 199, 104779.
- [18] He G., Zhao Y. and Yan C. 2024, August. A data augmentation method for multiaxial fatigue life prediction based on physics-informed tabular generative adversarial network. In *Journal of Physics: Conference Series*. (2816(1): 012048). IOP Publishing.
- [19] Dan Y., Zhao Y., Li X., Li S., Hu M. and Hu J. 2020. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*. 6(1): 84.
- [20] Leni D., Utami L. P., Berli A. U., Adriansyah and Haris. 2025, January. Feature selection with information gain in machine learning modeling for predicting low-alloy steel mechanical properties. In *AIP Conference Proceedings* (3223(1): 020023). AIP Publishing LLC.
- [21] Karudin A., Leni D., Lapisa R. and Kusuma Y. P. 2025. Utilizing Exploratory Data Analysis and Machine Learning to Enhance Learning Quality in Thermodynamics. *Journal of Theoretical and Applied Information Technology*. 103(1).
- [22] Singh Walia M. 2020. Synthetic Data Generation Using Wasserstein Conditional Gans with Gradient Penalty (WCGANS-GP).
- [23] Duran H. A. 2021. Wasserstein generative adversarial active learning for anomaly detection with gradient penalty (Master's thesis, Middle East Technical University (Turkey)).
- [24] Xu L. and Veeramachaneni K. 2018. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*.
- [25] Leni D., Kesuma D. S., Maimuzar Haris and Afriyani S. 2024. Prediction of mechanical properties of austenitic stainless steels with the use of synthetic data via generative adversarial networks. *Engineering Proceedings*. 63(1): 4.
- [26] Endres M., Mannarapotta Venugopal A. and Tran, T. S. 2022, August. Synthetic data generation: A comparative study. In *Proceedings of the 26th international database engineered applications symposium*. pp. 94-102.
- [27] Jiang Y., Li J., Yang X. and Yuan R. 2024. Applications of generative adversarial networks in materials science. *Materials Genome Engineering Advances*. 2(1): e30.
- [28] Falkiewicz M., Takeishi N. and Kalousis A. 2024. Kolmogorov-Smirnov GAN. *arXiv preprint arXiv:2406.19948*.
- [29] Speidel M. O. 2006. Nitrogen containing austenitic stainless steels. *Materialwissenschaft und Werkstofftechnik: Entwicklung, Fertigung, Prüfung, Eigenschaften und Anwendungen technischer Werkstoffe*. 37(10): 875-880.
- [30] Schwarz K., Liao Y. and Geiger A. 2021. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*. 34, 18126-18136.
- [31] Hittmeir M., Ekelhart A. and Mayer R. 2019, August. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In *Proceedings of the 14th international conference on availability, reliability and security*. pp. 1-6.
- [32] Cabrera M., Ninic J. and Tizani W. 2023. Fusion of experimental and synthetic data for reliable prediction of steel connection behaviour using machine learning. *Engineering with Computers*. 39(6): 3993-4011.
- [33] Niu M. C., Yin L. C., Yang K., Luan J. H., Wang W. and Jiao Z. B. 2021. Synergistic alloying effects on nanoscale precipitation and mechanical properties of ultrahigh-strength steels strengthened by Ni3Ti, Mo-enriched, and Cr-rich co-precipitates. *Acta Materialia*. 209, 116788.



- [34] Mohrbacher H. 2010. Principal effects of Mo in HSLA steels and cross effects with microalloying elements. Central Iron and Steel Research Institute (CISRI). 75-96.
- [35] Natras R., Soja B. and Schmidt M. 2022. Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. Remote Sensing. 14(15): 3547.