Jurnal Polimesin
Department of Mechanical Engineering
State Polytechnic of Lhokseumawe
http://e-jurnal.pnl.ac.id/polimesin
e-ISSN : 2549-1999 No. : 2 Month : April
p-ISSN : 1693-5462 Volume : 23 Year : 2025

Processing dates: received on 2024-11-13, reviewed on 2025-02-27, Accepted on 2025-02-28 and online availability on 2025-04-01

Prediction of aluminum alloy mechanical properties using synthetic data generated by generative adversarial networks

Lega Putri Utami[1], Armijal[2], Desmarita Leni[3,*], Andre Febrian Kasmar[4]

[1]Departement of Mechanical Engineering, University of Andalas, Padang 25175, Indonesia
[2]Department of Industrial Engineering, University of Andalas, Padang 25175, Indonesia
[3]Departement of Mechanical Engineering, University of Muhammadiyah Sumatera Barat, Padang 25586, Indonesia
[4]Department of Software Engineering Technology, State Polytechnic of Padang, Padang 25164, Indonesia
*Corresponding author: desmaritaleni@gmail.com

Abstract

Machine learning models are widely used to predict the mechanical properties of aluminum alloys. However, their accuracy is often hindered by the scarcity of high-quality tensile test data, as experimental data collection is costly and time-consuming. To address this limitation, this study employs Generative Adversarial Networks (GANs) to generate synthetic tensile test data for aluminum alloys, improving the accuracy of predictive models. The dataset consists of 200 real samples containing the compositions of nine chemical elements and two mechanical properties—Yield Strength (YS) and Ultimate Tensile Strength (UTS). A trained GAN model was used to generate 1,000 synthetic samples, whose statistical similarity to the original dataset was validated using the Kolmogorov-Smirnov (KS) test and Pearson correlation analysis. The results confirmed that all synthetic variables retained similar distributions and correlation patterns to the original dataset. To evaluate the impact of synthetic data on predictive accuracy, three machine learning algorithms—Random Forest Regressor (RF), Gradient Boosting Regressor (GBR), and Ada Boost Regressor (ABR)—were tested under two training schemes: (1) synthetic data for training and real data for testing and (2) real data for both training and testing. The RF model showed the highest improvement in UTS prediction, with reductions of 38.3% and 46.3% in Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), respectively. The GBR model exhibited notable enhancements in YS prediction, with MAE and RMSE reductions of 22.5% and 28.3%. These results demonstrate that GAN-generated synthetic data is highly effective in improving machine learning predictions of aluminum alloy properties, particularly when experimental data is limited.

Keywords:

Aluminum mechanical properties, machine learning, synthetic data, Generative Adversarial Network

1 Introduction

Aluminum is a lightweight metal with unique and highly versatile properties, making it one of the most widely used materials in modern industry. Aluminum is known for its resistance to corrosion, good thermal and electrical conductivity, and lightweight, with a density of only about one-third that of steel [1]. These properties make aluminum a preferred material in various applications, from transportation, construction, and electronics to household goods manufacturing. In the transportation industry, aluminum plays a crucial role in the production of vehicles such as cars, airplanes, and ships [2]. Its light weight helps reduce fuel consumption and emissions, making it an ideal material for efforts to enhance energy efficiency and sustainability. In the construction sector, aluminum is used in products such as wall panels, window frames, and other building structures due to its anti-corrosion properties, which make it durable even in extreme environments [3].

The mechanical properties of aluminum, such as Yield Strength (YS) and Ultimate Tensile Strength (UTS), are essential parameters that determine this material's ability to withstand loads without permanent deformation or damage. Yield Strength indicates the maximum stress a material can endure before beginning to undergo plastic deformation, while Ultimate Tensile Strength indicates the maximum stress it can withstand before breaking [4]. Understanding these mechanical properties is critical as they determine the applications and limitations of aluminum in construction design. In the aviation industry, aluminum with high tensile strength is required for parts that must withstand significant loads, while applications requiring flexible materials may need aluminum with softer mechanical properties. In the industry, testing the mechanical properties of aluminum requires significant resources, both in terms of time and cost. By using a synthetic data modeling approach with Tabular Generative Adversarial Networks (TGAN), the industry can reduce the number of physical experiments needed, thereby accelerating the development process of new materials.

Testing and modeling the mechanical properties of aluminum are crucial to ensure that this material meets the technical and safety requirements for its applications. A good understanding of aluminum's mechanical properties enables engineers and materials scientists to select, design, and modify materials according to the specific needs of an application [5]. Generally, the mechanical properties of aluminum are tested using tensile testing. In a tensile test, an aluminum sample is subjected to stress until it reaches the breaking point to determine the YS and UTS values. Although this method can provide accurate results, there are some challenges often encountered in the field. Conventional testing is highly expensive, requiring costly laboratory equipment and trained specialists to ensure accurate results. Additionally, this testing process is time-consuming, especially when repeated tests are required.

Studies [6] [7] mention that the percentage and composition of chemical elements in aluminum can influence the YS and UTS values. Each chemical element in an aluminum alloy, such as magnesium (Mg), zinc (Zn), copper (Cu), and silicon (Si), can affect the microstructure, which in turn can alter the material's characteristics. In the study [8], the addition of magnesium was found to increase strength through solid-solution strengthening. At the same time, zinc and copper contribute to precipitation hardening, enhancing strength but often reducing ductility or corrosion resistance. Consequently, any change in alloy composition necessitates retesting of mechanical properties, such as tensile tests, to ensure the material possesses characteristics that meet the intended application. This leads to increased production costs and slows material innovation due to repetitive tensile testing.

Machine learning approaches have become one of the promising methods for predicting the mechanical properties of materials and aiding in the design of new materials. Machine learning enables the processing and analysis of large volumes of data to discover complex patterns linking chemical composition, microstructure, and mechanical properties of materials [9]. By using algorithms that can learn from data, machine learning models can accurately predict mechanical properties based on input variables such as chemical composition. Previous research has demonstrated the successful use of machine learning in predicting the mechanical properties of materials. A study by Kulina [10] showed that ensemble models such as Random Forest can be used to predict the tensile strength of various metal alloys with high accuracy based on chemical

composition and heat treatment. Another study by Qian [11] utilized neural networks to design new composite materials with high strength and toughness. In that study, researchers could explore optimal element combinations to achieve desired mechanical properties.

However, one of the main challenges in using machine learning to predict the mechanical properties of materials is data limitations. Machine learning requires a sufficiently large and diverse dataset to train the model-train the model train the model effectively. Unfortunately, mechanical property data is often limited due to expensive testing processes, long testing times, and the destructive nature of tests such as tensile testing. Limited data can cause machine learning models to experience overfitting, a condition in which the model performs well on training data but is less accurate on new data [12]. This reduces the reliability of the model in predicting the properties of untested materials. To address data limitations in machine learning, the use of synthetic data generated through Generative Adversarial Networks (GAN) is becoming an increasingly popular solution. GAN is a type of machine learning model that uses two neural networks (a generator and a discriminator) that compete to generate new data that closely resembles real data. With GANs, additional synthetic datasets can be created, covering various combinations of chemical composition and other material characteristics, thereby enriching the training data and improving the accuracy of the predictive model.

Several studies have successfully utilized GAN to generate synthetic data for material applications. Study [13] developed synthetic data to predict the mechanical properties of polymer materials using Variational Autoencoders (VAE) and GAN. By leveraging the generated synthetic data, they were able to improve predictions of polymer material strength and toughness, which were previously limited due to insufficient experimental data. This model helped accelerate the design process for strong and durable polymer materials. The study [14] used GAN to generate synthetic data representing fatigue testing on metals. Fatigue testing usually takes considerable time and cost due to its repetitive nature. With synthetic data from GAN, machine learning models could predict a material's resistance to cyclic loading without requiring repeated physical testing. Another study [12] used synthetic data from GAN to enhance datasets in steel alloy optimization research. Their model combined GAN with optimization algorithms to find element combinations that result in optimal tensile strength. This study demonstrated that synthetic data could support deeper exploration of alloy composition variations that may not be achievable with only original data.

Based on the issues outlined above and the capability of GAN to create synthetic data closely resembling real data, this research aims to improve the prediction of aluminum's mechanical properties using synthetic data generated with GAN. In this study, synthetic data is created based on aluminum tensile test results, including chemical composition and mechanical properties of aluminum, such as Yield Strength (YS) and UTS. With synthetic data generated by GAN, machine learning can utilize a larger and more diverse dataset, enabling the model to learn from a wider range of chemical composition combinations. This not only enhances predictive accuracy but also saves the time and cost needed for repeated laboratory experiments. Thus, the combination of machine learning and synthetic data from GAN offers an efficient and economical approach to designing and evaluating new materials, particularly for the mechanical properties of aluminum.

2 Research methods/ materials

This study applies a Synthetic Data-Driven Machine Learning approach, where synthetic data is utilized to train machine learning models to improve prediction accuracy under conditions of limited original data. The research begins by collecting data on the chemical composition and mechanical properties of aluminum, which includes the key variables of the material. This data is then used to train a GAN model aimed at generating synthetic data with a distribution and inter-variable correlations similar to the original data. The synthetic data produced by GAN is subsequently used as training data for the machine learning model. The model is tested using new data, namely experimental data on the mechanical properties of aluminum to evaluate its accuracy and reliability. The performance of the machine learning model is assessed using evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2), providing insights into the model's prediction accuracy and precision.

2.1 Data Collection

In this study, data was collected from various material databases, such as MakeItFrom™ and Matmatch™, which provide comprehensive information on material properties. The obtained data includes 9 chemical element variables: Aluminum (Al), Magnesium (Mg), Zinc (Zn), Titanium (Ti), Copper (Cu), Manganese (Mn), Chromium (Cr), Iron (Fe), and Silicon (Si). Additionally, this data includes two heat treatment conditions for aluminum alloys: O (Annealed), referring to the material condition after undergoing a process to relieve internal stresses, and H (Strain Hardened), which refers to the material condition that has undergone strain hardening to increase its strength. The chemical element and heat treatment data are then linked to two main mechanical properties: YS and Ultimate UTS. Yield Strength is the maximum stress the material can withstand before experiencing plastic deformation, while Ultimate Tensile Strength is the maximum stress the material can endure before fracturing or total failure [15]. This information allows researchers to understand the influence of chemical composition and heat treatment on the mechanical properties of aluminum.

This study applies descriptive analysis to the distribution of chemical composition and mechanical properties of aluminum to address potential biases in the original dataset. The analysis is conducted in two main stages: (1) identifying the distribution of chemical elements to detect imbalances in composition and (2) analyzing the correlation between chemical composition and mechanical properties to evaluate potential biases in variable relationships.

With this approach, the study ensures that the data used has a more balanced distribution and unbiased variable relationships, resulting in more accurate and generalizable predictions.

2.2 Synthetic data modeling using GAN

After the data is collected, the next step is to build a synthetic data model using GAN. GAN is a machine learning technique consisting of two neural networks: a generator and a discriminator, which compete with each other in the training process to produce new data resembling the original data. Fig. 1 illustrates the GAN architecture. In this study, GAN is used to create synthetic data with patterns and distributions similar to the original data, thus expanding the number of samples available for predictive analysis.

The GAN model is designed with a generator network that aims to generate new data based on the distribution of the original data, such as the chemical composition and mechanical properties of aluminum. On the other hand, the discriminator network functions to distinguish between the original data and the data generated by the generator. Through an iterative training process, the generator continuously improves its ability to generate data increasingly similar to the original data, while the discriminator becomes more skilled at distinguishing between the two types of data.

The model training was conducted with a batch size of 64, meaning that each training iteration used 64 samples before updating the network weights. The latent dimension was set to 10, representing the number of latent random variables used as input for the generator to produce a realistic data distribution. The model was trained for 1000 epochs to ensure that the generator and discriminator reached equilibrium in generating and evaluating synthetic data.

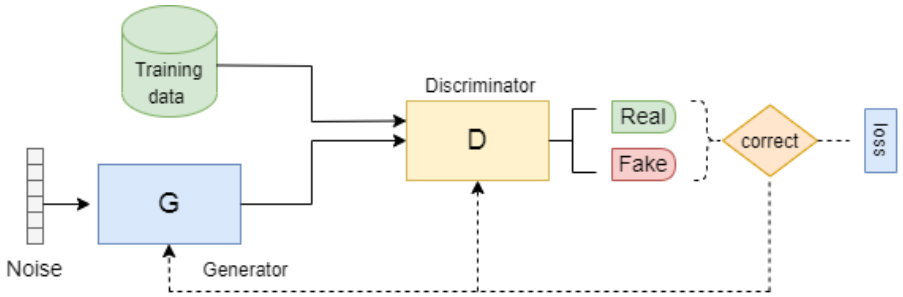


Fig. 1. Generative Adversarial Networks (GAN) Architecture ([12])

A learning rate of 0.0002 was chosen to maintain stable weight updates without causing oscillations or excessively slow convergence. With this configuration, the GAN model can generate synthetic data that more accurately represents the original data distribution, thereby improving the accuracy of mechanical property predictions for aluminum in this study.

The generated synthetic data is then validated to ensure that statistical patterns, such as value distribution and relationships between variables, remain consistent with the original data. In this study, the Kolmogorov-Smirnov (KS) test and Pearson correlation are used, which can be calculated using Eqs (1 and 2). Where D is the KS test statistic, F1(x) is the Empirical Cumulative Distribution Function (ECDF) of the first sample, and F2(x) is the ECDF of the second sample [16]. The larger the value of D, the more different the two data distributions

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

$$r_{xy} = \frac{\sum xy}{(n-1)s_x.s_y} \quad (2)$$

Where r_xy represents the Pearson correlation coefficient, Σxy denotes the sum of the products of x and y, n indicates the sample size, x stands for the independent variable, y represents the dependent variable, and S signifies the standard deviation [17]. The correlation coefficient ranges from -1 to 1. A value of -1 indicates a strong negative correlation between the two variables, a value of 0 indicates no correlation, and a value of 1 indicates a strong positive correlation.

The utilization of synthetic data enables stronger and more accurate predictive model testing, reducing the risk of overfitting that often occurs with limited datasets. Thus, synthetic data modeling using GAN is expected to improve the quality and robustness of predictive models in understanding the influence of chemical composition and heat treatment on the mechanical properties of aluminum.

2.3 Predictive Modeling Using Machine Learning

At this stage, predictions for YS and Ultimate UTS are made using various ensemble-based machine-learning algorithms. Ensemble models are chosen for their ability to enhance prediction accuracy and stability by combining multiple base models (weak learners) into a single strong model [18]. The algorithms applied include Random Forest Regressor (RF), Gradient Boosting Regressor (GBR), and Ada Boost Regressor (ABR). Each of these algorithms offers a unique approach to managing variance and bias in the data, potentially leading to more accurate predictions. To evaluate model performance, two testing schemes are implemented:

1. Synthetic training - real testing

In this scheme, synthetic data generated by GAN is used to train the model, while real data is used as test data. This scheme aims to test the model's generalization ability on real data after being trained with synthetically expanded data, which is expected to enrich the variation in the training process.

2. Real training - real testing

In this scheme, the model is trained and tested using real data. This scheme serves as a baseline to compare the performance of the model trained with synthetic data. Through this scheme, evaluations are conducted to determine whether the generated synthetic data can improve prediction accuracy or at least maintain performance comparable to training with real data alone.

To ensure model performance and reduce the risk of overfitting, this study applies the K-Fold Cross-Validation method with K=10. With a total of 1000 data samples, the dataset is divided into 10 subsets (folds), where each subset is alternately used for training and validation. The choice of K=10 is based on the balance between model bias and variance. A sufficiently large K value allows the model to be tested on various data combinations, reducing the risk of the model recognizing specific patterns in the dataset rather than understanding more general relationships. With this approach, the model is thoroughly tested against variations in the dataset, minimizing overfitting and ensuring reliable performance that can be well-generalized.

Each model is evaluated based on performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2), to assess how well the model predicts the YS and UTS of aluminum alloys. The comparison results from these two schemes provide insight into the effectiveness of synthetic data in enhancing or maintaining model prediction accuracy and highlight the advantages of each algorithm in modeling the mechanical properties of aluminum alloys. These three evaluation metrics can be calculated using Eqs (3-5) [19].

$$MAE = \frac{i}{N}\sum|y_i - z_i| \quad (3)$$

Where i is the index of the data sample, N is the total number of samples, y_i is the actual value of the i-th data point, and z_i is the predicted value by the model for the i-th data point.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2} \quad (4)$$

Where n is the number of data points used to test the model, f(X_i) 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.

$$R = \frac{\sum_{i}^{n}=1(f(X_i)-\bar{f(X)})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(f(X_i)-\bar{f(X)})^2}\sqrt{\sum_{i=1}^{n}Y_i-\bar{Y})^2}} \quad (5)$$

Where f(Xi) is the predicted value of the dependent variable (Y) based on the independent variable (X) for the i-th observation, $\bar{f(X)}$ is the average of all predicted values f(Xi) across all observations, Yi is the actual observed value of the dependent variable for the i-th observation, $\bar{Y}$ is the average of all observed values Yi across all observations, and n is the total number of observations.

Where f (Xi) is the predicted value of the dependent variable (Y) based on the independent variable (X) for the i-th observation, f ($\bar{X}$) is the average of all predicted values f(Xi) across all observations, Yi is the actual observed value of the dependent variable for the i-th observation, $\bar{Y}$ is the average of all observed values Yi across all observations, and n is the total number of observations.

3 Results and discussion

3.1 Data collection

The mechanical properties data for aluminum in this study were sourced from material databases such as Matmatch and MakeItFrom, which are open-access online material libraries providing thousands of experimental material datasets from various sources. These databases exemplify a new trend in materials informatics, which integrates experimental data with digital applications. However, most of the data on these sites remains in raw form, requiring further data processing to obtain the necessary information. Once the data is collected, the next step is data cleaning and filtering, which involves removing duplicate, irrelevant, and inconsistent data with the specified chemical element variables.

This study selects nine chemical element variables: Magnesium (Mg), Zinc (Zn), Titanium (Ti), Copper (Cu), Manganese (Mn), Chromium (Cr), Iron (Fe), Silicon (Si), and Aluminum (Al) as the main variables in the dataset because these elements significantly influence the mechanical properties of aluminum alloys, particularly YS and UTS. The selection of these elements is based on previous studies [20],[21] showing that variations in the composition of these elements play a crucial role in shaping the mechanical properties of aluminum. After cleaning the data, 200 samples were obtained, consisting of 9 chemical element variables and 2 mechanical properties (YS and UTS).

This element selection process aligns with trends in materials informatics, where experimental data is systematically organized and processed for use in data-driven predictive model development. The data collected from various material databases allows researchers to obtain a broader and more representative dataset without conducting all experiments manually, which would require considerable time and expense. Table 1 summarizes the basic statistics of the aluminum dataset.

Table 1. Statistics of Aluminum dataset

Variable	Min	Max	Mean
Mg (%)	0	3.1	0.978
Zn (%)	0.05	0.5	0.207
Ti (%)	0	0.25	0.069
Cu (%)	0	0.5	0.13
Mn (%)	0.02	1	0.404
Cr (%)	0	0.3	0.11
Fe (%)	0.15	1	0.576
Si (%)	0.15	1	0.436
Al (%)	94.65	99.58	97.089
Yield strength	55	230	112.408
Tensile strength	90	375	232.95

It including the minimum, maximum, and mean values of each variable. The variability in this data shows that the available dataset covers a wide range of chemical compositions and mechanical properties, providing a strong foundation for developing synthetic data that closely resembles the original data in this study.

3.2 Synthetic data modeling using GAN

In this study, a GAN is used to generate synthetic data that can mimic the characteristics of the original tensile test data for aluminum. The original data includes chemical elements such as Mg, Zn, Ti, Cu, Mn, Cr, Fe, Si, and Al, as well as the mechanical properties of Yield Strength and Ultimate Tensile Strength. The purpose of using synthetic data is to expand the variation of available data, which is ultimately expected to improve the accuracy of the predictive model for aluminum's mechanical properties. The parameters used in the GAN model can be seen in Table 2.

Table 2. GAN parameters

Hyperparameter	Value
batch_size	64
latent_dim	10
num_epochs	1000
learning_rate	0.0002

An important component of this process is monitoring the loss values of the discriminator and generator during GAN training, which provides information to ensure that the model effectively generates realistic data [22]. Training results indicate that, in the early stages, the discriminator loss tends to fluctuate with high values. This occurs because the discriminator can easily distinguish between real data and synthetic data, which is still not realistic from the generator. On the other hand, the generator loss decreases more steadily, indicating that the generator is beginning to learn to produce data closer to the real data. In many cases of synthetic data generation using GANs, the discriminator has a relatively easier task compared to the generator. The discriminator only needs to distinguish between two types of data (real and fake), while the generator must learn the complex distribution of the real data to generate realistic data [23]. Because the discriminator's task is simpler, it often achieves a higher level of accuracy first, resulting in a lower discriminator loss compared to the generator loss. The GAN training graph can be seen in Fig. 2.

As training iterations increase, the loss values show stable fluctuations between the two losses. The generator and discriminator appear to reach a dynamic balance, where the generator begins producing sufficiently realistic data, making it increasingly difficult for the discriminator to differentiate between real and synthetic data.

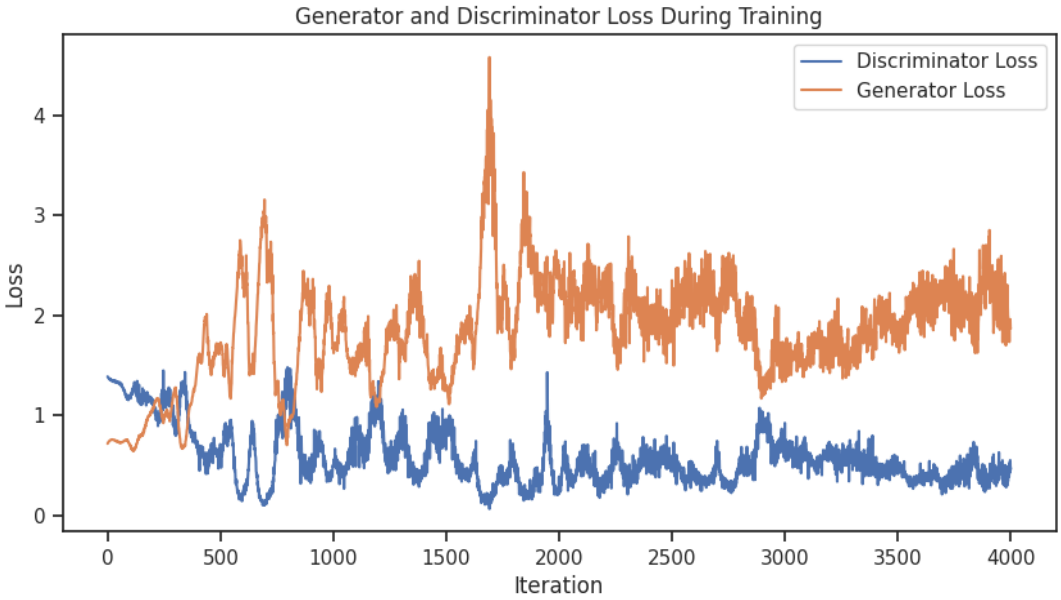


Fig. 2. Comparison of loss descriptor and generator

The fluctuations observed in both losses reflect a healthy competition between the generator and discriminator, indicating that the GAN has reached the desired level of stability. The fluctuating loss values seen in the generator and discriminator on the graph are an important indication that mode collapse has been successfully avoided. Mode collapse is a condition where the generator produces very uniform data, drastically reducing variation in synthetic data [24]. To ensure that the synthetic data generated by GAN has characteristics similar to the original data, it is necessary to evaluate it using the KS test and Pearson correlation. Table 3 shows the results of the KS test between the original and synthetic data, while Fig. 3 provides a visualization of the distribution for each variable.

Based on Table 3, the p-value for all variables is greater than 0.05, meaning the null hypothesis cannot be rejected. This indicates that the distribution of synthetic data is statistically similar to the distribution of the original data for all tested variables. The KS Statistic for all variables is also relatively low, indicating a small maximum difference between the cumulative distributions of the original and synthetic data.

Table 3. KS test results

Variable	KS Statistic	p-value	Same Distribution
Mg	0.057	0.591	Yes
Zn	0.101	0.051	Yes
Ti	0.069	0.369	Yes
Cu	0.094	0.082	Yes
Mn	0.059	0.555	Yes
Cr	0.080	0.204	Yes
Fe	0.063	0.478	Yes
Si	0.050	0.747	Yes
Al	0.068	0.367	Yes
YS (Mpa)	0.046	0.840	Yes
UTS (Mpa)	0.064	0.456	Yes

These results suggest that the synthetic data generated using GAN successfully replicates the distribution of the original data for variables such as Mg, Zn, Ti, Cu, Mn, Cr, Fe, Si, Al, as well as the mechanical properties of YS and UTS. This demonstrates that the GAN used in this study is effective in producing synthetic data with statistical characteristics similar to the original data.

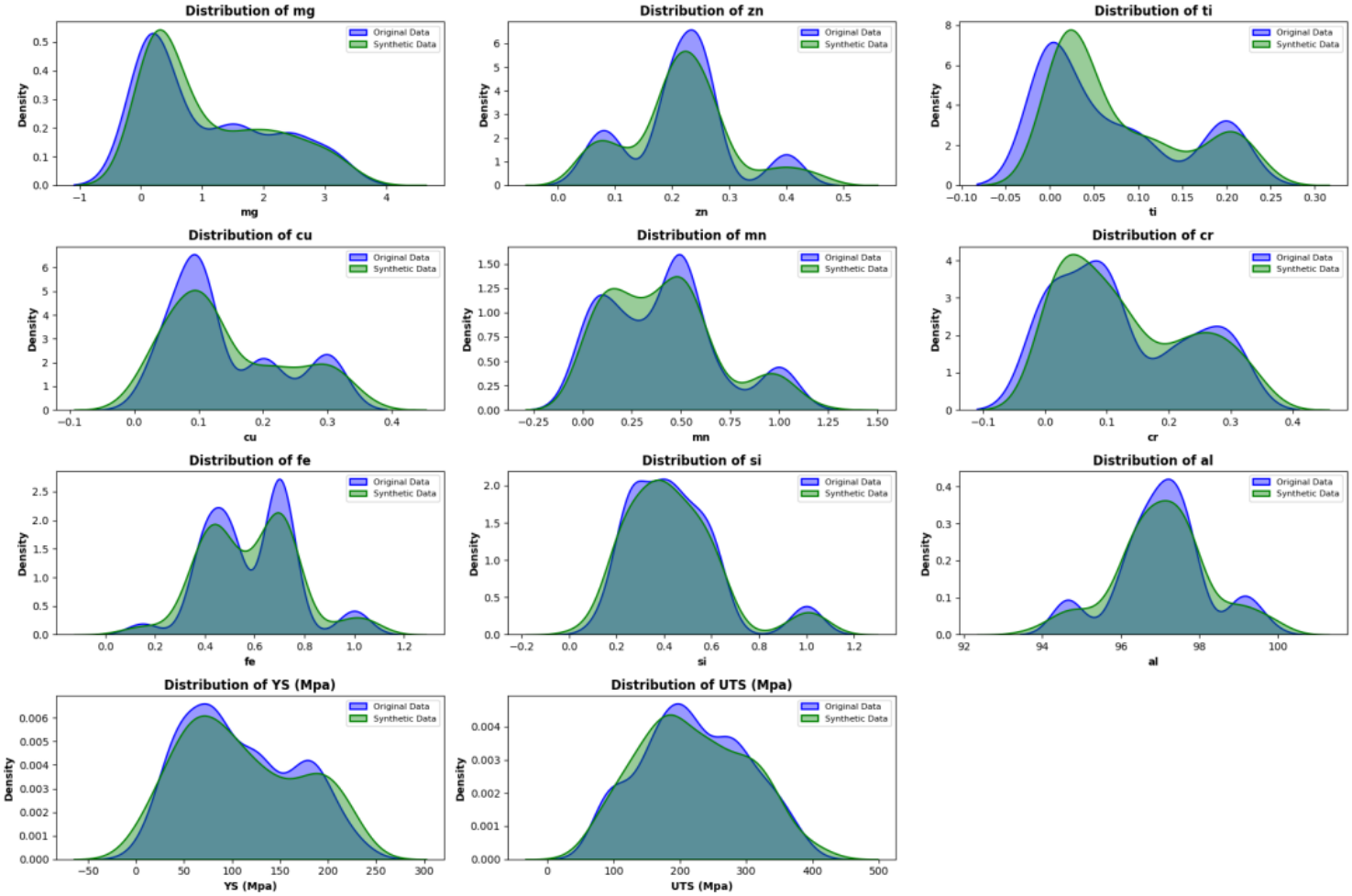
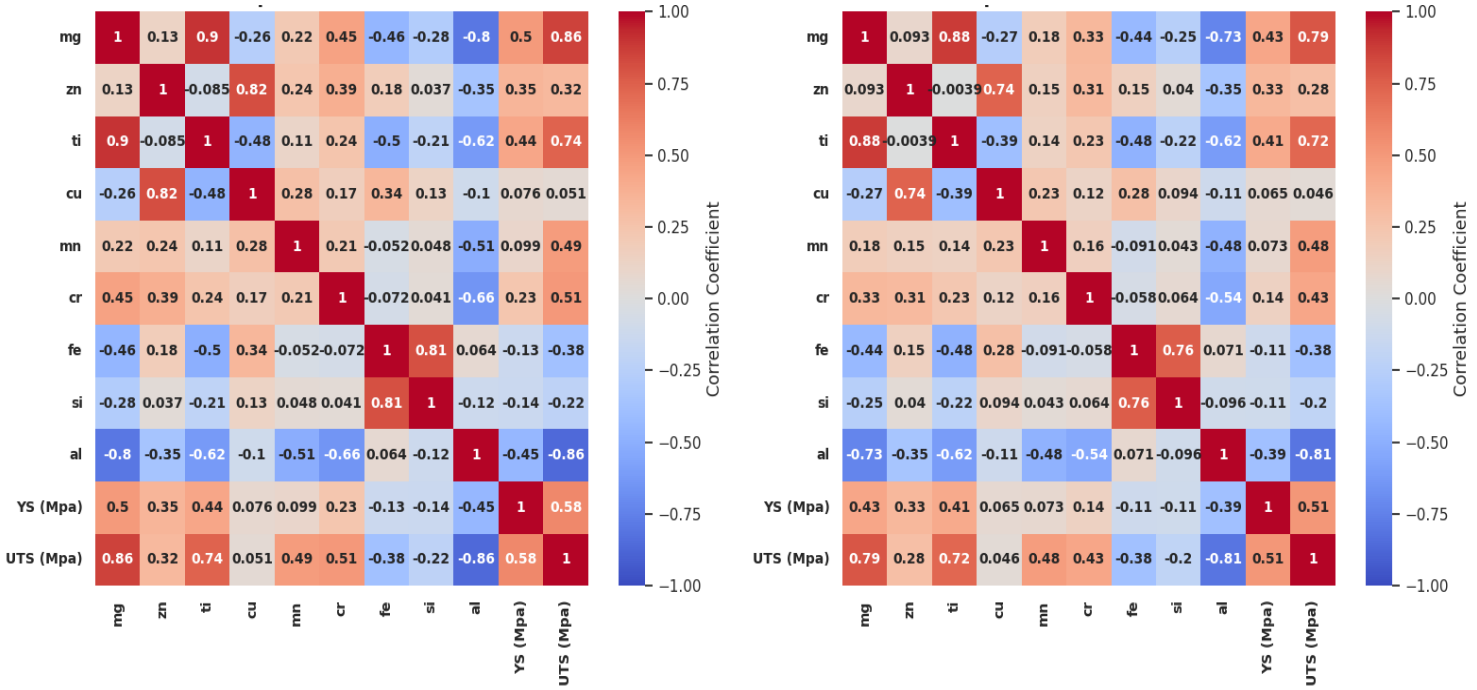


Fig. 3. Comparison of the distribution of original data with synthetic data.

This is further supported by the comparison of distributions for each variable, as shown in Fig. 4. The distribution of synthetic data (green line) and original data (blue line) largely overlap for most variables, indicating that the GAN has successfully mimicked the distribution pattern of the original data. Although most variables show similar distributions, there are slight differences in the Ti and Cu variables, where the synthetic data distribution does not completely overlap with the original data. These differences are due to the variability in the original data, which is more challenging for GAN to capture, or due to unique characteristics in the distribution of these variables [25]. However, the KS test p-value for these variables remains above 0.05, indicating that these differences are still within statistically acceptable limits.

In an effort to ensure that the data generated by GAN truly aligns with the original data, this study also conducts a correlation analysis comparison to confirm that the synthetic data not only mimics the univariate distribution of each variable but also preserves the relationships between variables as found in the original data. Preserving the correlation structure between variables is key to maintaining a realistic representation in synthetic data, especially in the context of materials research like this, where interactions between chemical elements significantly influence the mechanical properties of aluminum [26]. Therefore, comparing the correlations between the original and synthetic data is a crucial evaluation step. The results of the correlation comparison between the two datasets can be seen in Fig. 4.



(a) (b)

Fig. 4. Comparison of correlations between (a) original data and (b) synthetic data.

Based on these results, it can be seen that most of the correlation patterns in the synthetic data successfully replicate the correlation patterns in the original data. Correlations are evident in other variables, such as between YS with Ti and Zn, which show a fairly similar correlation between the two datasets. On the other hand, there are some minor differences in the correlations of certain variables between the original and synthetic data. For instance, the correlation between Si and Mn slightly differs between the original and synthetic datasets, with a correlation coefficient of 0.28 in the original data and 0.25 in the synthetic data. Although there are minor variations, overall, the correlation patterns among the main chemical variables are maintained. These small differences arise from natural variations in the data generation process using GAN, which attempts to replicate general patterns without matching every correlation exactly [27].

The element Mg has a relatively strong positive correlation with UTS in both datasets, with a correlation coefficient of 0.86 in the original data and 0.78 in the synthetic data. For a visualization of the similarity in the distribution shapes between the original and synthetic data, see Fig. 5. This demonstrates that the GAN used in this study is capable of capturing and mimicking the close relationship between magnesium and the maximum tensile strength of aluminum. Overall, the correlation comparison results indicate that the synthetic data has a structure similar to the original data, both in individual value distributions and in inter-variable relationships. This shows that the GAN model used in this study successfully generates data realistic enough for further analysis or predictive model training without sacrificing the original relationships between variables.

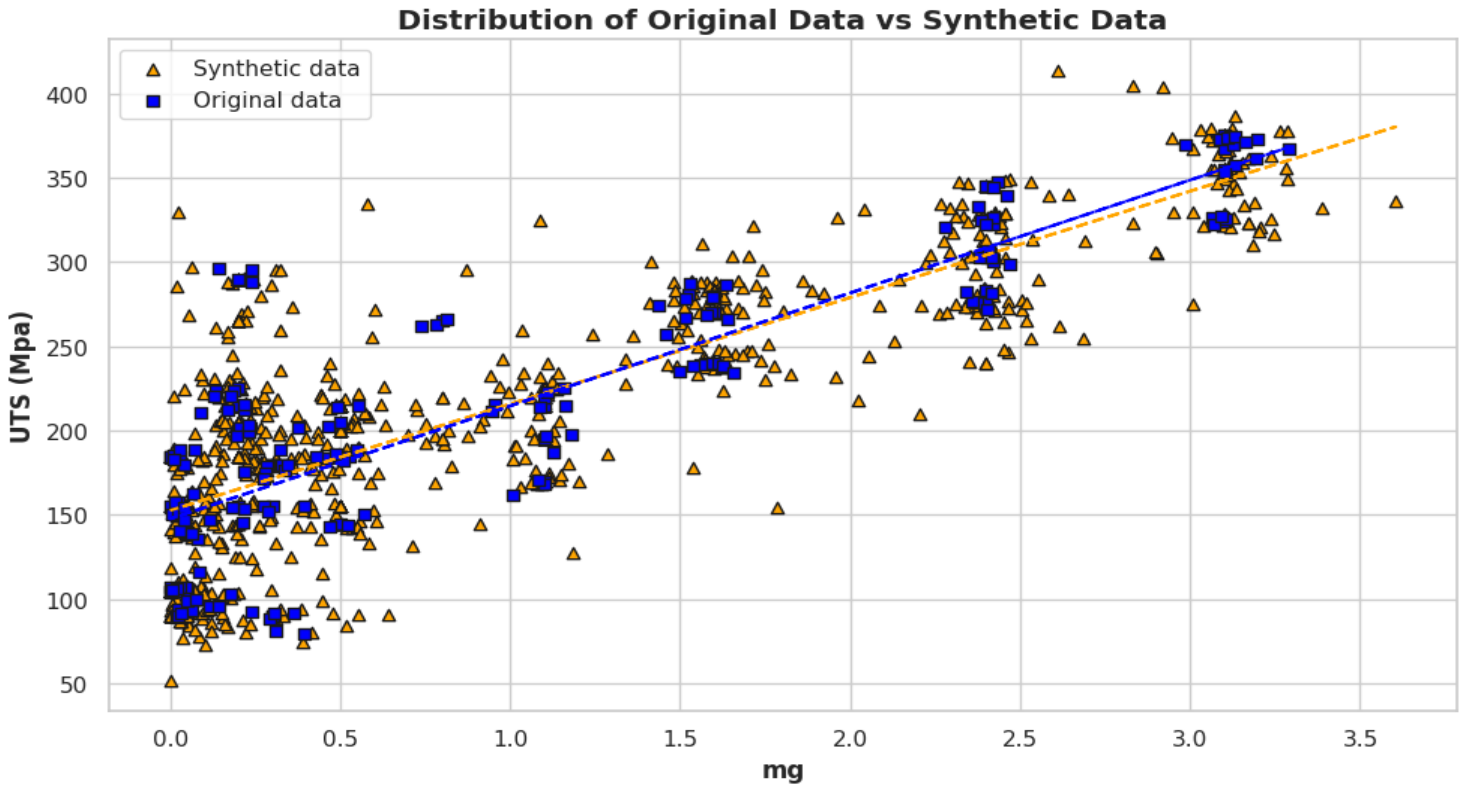


Fig. 5. Comparison of Mg distribution with UTS in both datasets.

3.3 Predictive modeling using machine learning

At this stage, predictions of aluminum's mechanical properties are made using various ensemble-based machine-learning algorithms. Ensemble models are chosen for their ability to improve prediction accuracy and stability by combining multiple base models into a stronger model [18]. Ensemble algorithms are also effective in reducing variance and addressing overfitting in datasets with high complexity, such as mechanical property data for aluminum involving multiple chemical variables. The algorithms applied include Random Forest Regressor (RF), Gradient Boosting Regressor (GBR), and Ada Boost Regressor (ABR). Each algorithm is trained and evaluated using the cross-validation (CV) method with K=10. Cross-validation with K=10 divides the dataset into 10 subsets, which are used alternately for training and testing. This method ensures that the model is thoroughly tested on all data, reducing the likelihood of overfitting and providing a more accurate estimate of model performance [28].

With cross-validation, model evaluation results are more stable because performance is measured across various data subsets. The settings and hyperparameter values for the three algorithms used in this study are shown in Table 4. Hyperparameters were selected through a tuning process to ensure that each model operates optimally and provides the best prediction results for aluminum's mechanical properties. At this stage, two testing schemes are conducted to evaluate the performance of the predictive model.

Table 4. Hyperparameters used for each model

Random Forest Regressor		Gradient Boosting Regressor		Ada Boost Regressor	
Hyperparameter	Value	Hyperparameter	Value	Hyperparameter	Value
bootstrap	TRUE	criterion	friedman_mse	estimator	None
criterion	squared error	learning rate	0.1	learning rate	1
max_features	1	loss	squared error	loss	linear
min_samples_leaf	1	max_depth	3	n_estimators	50
min_samples_split	2	min_samples_leaf	1	random state	42
n_estimators	100	min_samples_split	2		
random state	42	n_estimators	100		
		random state	42		
		subsample	1		
		tol	0.0001		
		validation fraction	0.1		

The first scheme is Synthetic Training-Real Testing, where synthetic data generated by GAN is used as model training data, while the original data is used as test data. The second scheme is Real Training-Real Testing, in which the model is fully trained and tested using the original data. These two schemes are designed to identify the impact of using synthetic data in a machine learning predictive model. By comparing the results from both schemes, we can assess how well synthetic data approximates the performance of original data in producing accurate predictions. To evaluate the predictive performance of each model in both schemes, several evaluation metrics are used: MAE, RMSE, and R-squared. These metrics measure the accuracy and reliability of the model in predicting the tensile strength of aluminum. MAE and RMSE measure the level of prediction error in absolute terms and squared terms, respectively, while R-squared assesses how much of the variability in the data is explained by the model. By comparing the evaluation metrics from the first and second schemes, we can assess the impact of using synthetic data on the accuracy of the machine learning predictive model.

Based on the test results for both schemes, it can be seen that Scheme 1 (Synthetic Training - Real Testing) shows better performance than Scheme 2 (Real Training - Real Testing) across all three predictive models. For the RF model in predicting YS, Scheme 1 yields an MAE of 27, RMSE of 35, and R-squared of 0.89, while Scheme 2 yields a higher MAE of 33.5, RMSE of 42.5, and R-squared of 0.8. This result indicates that using synthetic data generated by GAN as training data can improve prediction accuracy for the RF model compared to training with only original data. A similar trend is observed for UTS prediction, where Scheme 1 provides more accurate results with lower MAE and RMSE and higher R-squared than Scheme 2. For the GBR model, the performance difference between the two schemes is not very significant, although Scheme 1 still yields slightly better results for both YS and UTS predictions. The GBR model in Scheme 1 achieves an MAE of 20.54, RMSE of 24.73, and R-squared of 0.91 for YS, while Scheme 2 has an MAE of 26.5, RMSE of 34.5, and R-squared of 0.88. Similarly, for the ABR model, Scheme 1 outperforms Scheme 2. For example, in UTS prediction, Scheme 1 achieves an MAE of 20.5, RMSE of 22.5, and R-squared of 0.92, while Scheme 2 has an MAE of 22.1, RMSE of 27.28, and R-squared of 0.86. This suggests that although ABR is known to be more sensitive to noise or differences in the training data, this model still provides better predictions when trained with synthetic data. The test results for both schemes can be seen in Table 5.

These results indicate that synthetic data can produce predictions close to the performance of original data in the GBR model, while for RF and ABR, synthetic data can improve the predictive model. RF shows the most significant improvement, particularly in UTS prediction, with increases in MAE and RMSE of 38.3% and 46.3%, respectively. This suggests that using synthetic data is highly effective in enhancing the accuracy of this model. GBR also shows improvement, especially in YS prediction, with increases in MAE by 22.5% and RMSE by 28.3%. The improvement in UTS prediction is smaller but still positive. ABR shows smaller gains than the other two models, but it still demonstrates accuracy improvements in both prediction targets. The difference in performance can be explained by several key factors related to the characteristics of the algorithm and the impact of synthetic data on the model training process.

Table 5. Comparison of test results of the two schemes

Model	Target	Skema 1			Skema 2		
		MAE	RMSE	R-squared	MAE	RMSE	R-squared
Random Forest Regressor	YS	27	35	0.89	33.5	42.5	0.8
	UTS	18.5	21.5	0.94	30	40	0.86
Gradient Boosting Regressor	YS	20.54	24.73	0.91	26.5	34.5	0.88
	UTS	18.4	21.3	0.95	19.74	23.6	0.89
Ada Boost Regressor	YS	30	38	0.88	31.5	41	0.85
	UTS	20.5	22.5	0.92	22.1	27.28	0.86

Random Forest (RF) is an ensemble learning method based on bagging, which constructs multiple decision trees in parallel and combines their results to enhance prediction stability. This approach enables RF to handle complex data distributions and high variability, which are often challenges in small datasets.

In this study, the addition of synthetic data enriches the variation in the training data, allowing RF to build a more accurate model while minimizing the risk of overfitting. Previous research has also shown that RF excels in processing small datasets that have been expanded using synthetic data. For instance, a study [29] found that RF is more resilient to noise and data imbalance compared to boosting models, especially when the dataset is augmented using a generative approach.

These results align with previous studies [23], [30], [31] which found that synthetic data generated by GAN models can serve as an effective alternative training dataset in situations where original data is limited. Those studies also found that synthetic data from GAN can represent the distribution of the original data well, contributing to improved predictive model performance. Findings in this study, with better results in Scheme 1, indicate that the synthetic data generated not only resembles the original data in distribution but also retains a representative correlation structure between variables. This aligns with the results of the KS test and correlation analysis conducted earlier, which showed that the synthetic data is similar to the original data in terms of distribution and inter-variable relationships. The comparison of GBR predictions in the two testing schemes, Scheme 1 and Scheme 2, for YS and UTS predictions can be seen in Figs 6 and 7.

The GBR prediction results in Scheme 1 indicate that synthetic data successfully provides prediction results close to the actual data. The predicted data points are scattered around the diagonal line, indicating closeness between actual and predicted values for YS and UTS. Especially for UTS, most of the prediction points are close to the red line, indicating that the model can predict values with high accuracy. This suggests that the synthetic data generated by GAN can provide sufficiently representative information for the model to capture patterns and relationships between variables, enabling the model to produce predictions close to the original data. On the other hand, the prediction results in Scheme 2 show increased dispersion of prediction points around the diagonal line compared to Scheme 1. For YS prediction, some points begin to deviate further from the diagonal line, indicating an increase in prediction error. Similarly, in UTS prediction, some prediction points are outside the ideal line range, suggesting that the model in Scheme 2 is slightly less accurate than in Scheme 1.

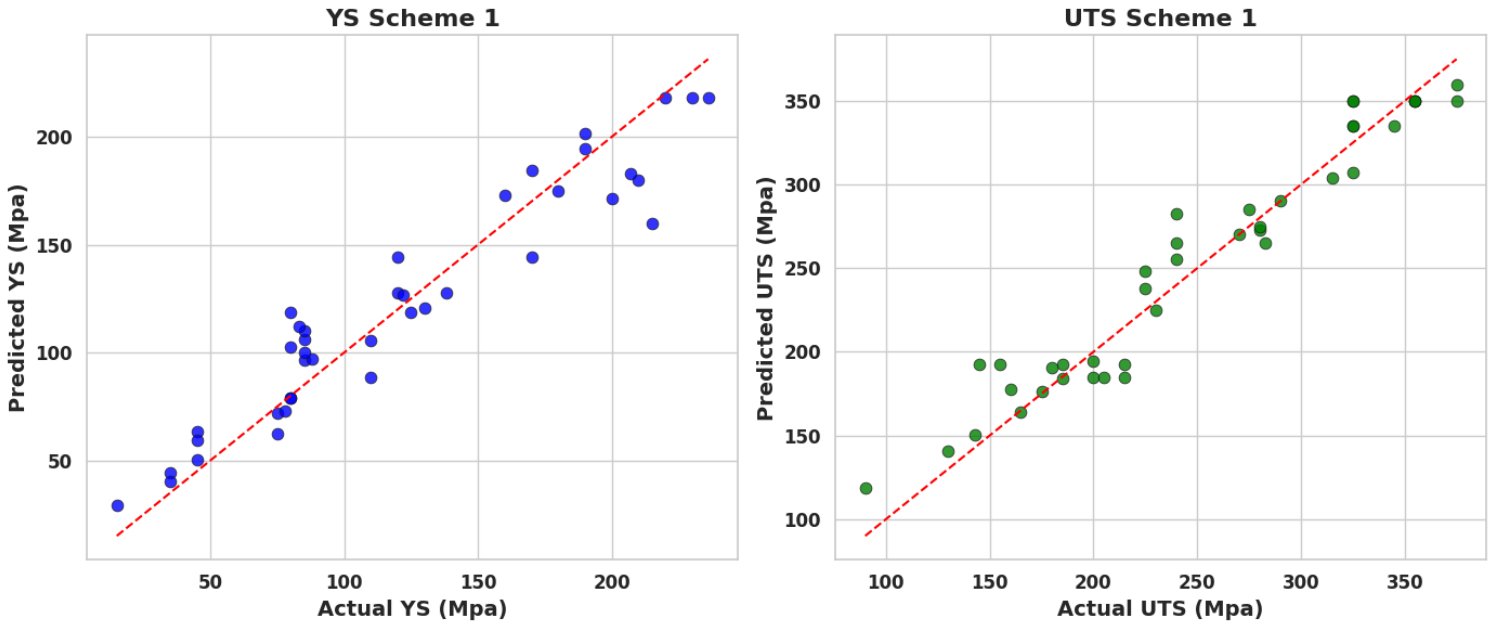


Fig. 6. Prediction results of GradientBoostingRegressor in Scheme 1

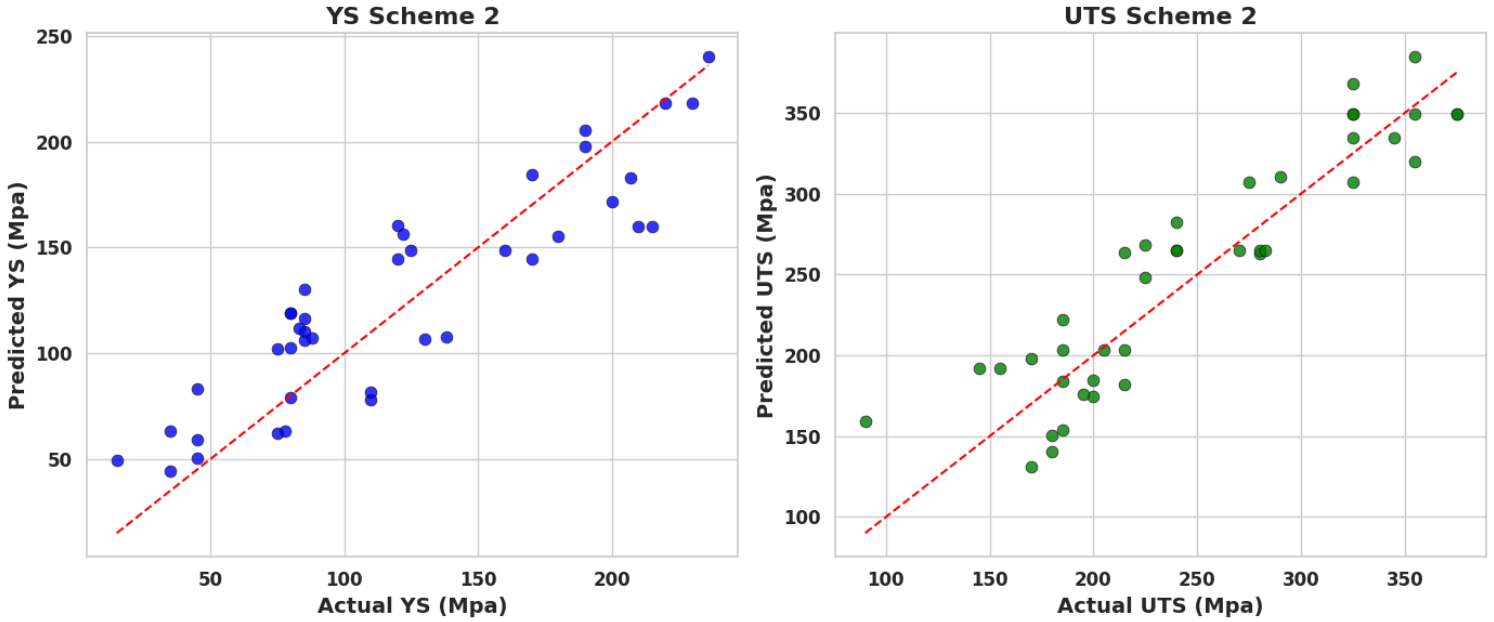


Fig. 7. Prediction results of Gradient Boosting Regressor in Scheme 2

Overall, these results show that Scheme 1 provides better and more consistent prediction results than Scheme 2, especially for UTS prediction. Using synthetic data for model training in Scheme 1 appears to maintain the pattern of inter-variable relationships. Additionally, the better MAE, RMSE, and R-squared values in Scheme 1, as presented in Table 5, support these findings, where the use of synthetic data results in better predictions than original data in some algorithms. These results are influenced by several factors, such as the larger amount of synthetic data compared to the original data and the diversity in synthetic data generated by GAN.

In Scheme 1, the model is trained using 1,000 synthetic data samples, while in Scheme 2, the model is trained with only 200 original data samples. The larger dataset in Scheme 1 allows the model to explore more variation in the data, ultimately improving the model's generalization ability and reducing the risk of overfitting. With more samples, the model can capture finer patterns and enhance the understanding of inter-variable relationships, increasing prediction accuracy when tested with original data. Additionally, synthetic data generated by GAN often reflects a broader diversity than the limited original data. GANs are designed to produce data with variation in range and distribution, allowing the model to learn from a broader set of patterns. This makes models trained with synthetic data more robust and capable of producing more accurate predictions when faced with diverse test data. These findings open up opportunities to use synthetic data in materials research and other applications that require high-quality data despite limitations in original data.

4 Conclusions

This study demonstrated that GAN can effectively generate synthetic tensile test data, significantly enhancing the accuracy of machine learning models for predicting the mechanical properties of aluminum alloys. The generated data closely mirrors the original dataset in statistical distribution and correlation patterns, as validated by the KS test and Pearson correlation analysis. Machine learning models trained with synthetic data exhibited substantial improvements, with the Random Forest Regressor (RF) model showing the highest enhancement in UTS prediction (38.3% and 46.3% reductions in MAE and RMSE, respectively). The GBR model also demonstrated notable performance gains in YS prediction (MAE reduced by 22.5% and RMSE by 28.3%). These findings underscore the potential of GAN-generated synthetic data to address data scarcity challenges in material science, enabling more reliable and cost-efficient machine learning-based predictions.

References

[1] G. Xu *et al.*, "Review on corrosion resistance of mild steels in liquid aluminum," *J. Mater. Sci. Technol.*, Vol. 71, Pp. 12–22, Apr. 2021, Doi: 10.1016/J.Jmst.2020.08.052.

[2] B. Zhou, B. Liu, And S. Zhang, "The advancement of 7xxx series aluminum alloys for aircraft structures: A Review," *Metals*, Vol. 11, No. 5, P. 718, Apr. 2021, Doi: 10.3390/Met11050718.

[3] N. C. G. Silveira, M. L. F. Martins, A. C. S. Bezerra, And F. G. S. Araújo, "Red mud from the aluminum industry: Production, characteristics, and alternative applications in construction materials—A review," *Sustainability*, Vol. 13, No. 22, P. 12741, Nov. 2021, Doi: 10.3390/Su132212741.

[4] M. Simoncini, A. Costa, S. Fichera, And A. Forcellese, "Experimental analysis and optimization to maximize ultimate tensile strength and ultimate elongation of friction stir welded aa6082 aluminum alloy," *Metals*, Vol. 11, No. 1, P. 69, Dec. 2020, Doi: 10.3390/Met11010069.

[5] A. Du Plessis *et al.*, "Properties and applications of additively manufactured metallic cellular materials: A review," *Prog. Mater. Sci.*, Vol. 125, P. 100918, Apr. 2022, Doi: 10.1016/J.Pmatsci.2021.100918.

[6] Y. Otani And S. Sasaki, "Effects of the addition of silicon to 7075 aluminum alloy on microstructure, mechanical properties, and selective laser melting processability," *Mater. Sci. Eng. A*, Vol. 777, P. 139079, Mar. 2020, Doi: 10.1016/J.Msea.2020.139079.

[7] M. D. Vijayakumar, V. Dhinakaran, T. Sathish, G. Muthu, And P. M. B. Ram, "Experimental Study of chemical composition of aluminum alloys," *Mater. Today Proc.*, Vol. 37, Pp. 1790–1793, 2021, Doi: 10.1016/J.Matpr.2020.07.391.

[8] A. Kazemi And S. Yang, "Effects of magnesium dopants on grain boundary migration in aluminum-magnesium alloys," *Comput. Mater. Sci.*, Vol. 188, P. 110130, Feb. 2021, Doi: 10.1016/J.Commatsci.2020.110130.

[9] D. Leni, "Prediction modeling of low alloy steel based on chemical composition and heat treatment using artificial neural network," *J. Polimesin*, Vol. 21, No. 5, Pp. 54–61, Oct. 2023, Doi: 10.30811/Jpl.V21i5.3896.

[10] H. Kulina, M. Koleva-Petrova, And S. Gocheva-Ilieva, "Ensemble learning for predicting the tensile strength of alloy steels from chemical composition and processing parameters," In *2022 International Conference On Electrical, Computer, Communications And Mechatronics Engineering (Iceccme)*, Maldives, Maldives: Ieee, Nov. 2022, Pp. 01–06. Doi: 10.1109/Iceccme55909.2022.9988668.

[11] C. Qian, R. K. Tan, And W. Ye, "Design of architectured composite materials with an efficient, adaptive artificial neural network-based generative design method," *Acta Mater.*, Vol. 225, P. 117548, Feb. 2022, Doi: 10.1016/J.Actamat.2021.117548.

[12] D. Leni, D. S. Kesuma, Maimuzar, Haris, And S. Afriyani, "Prediction of mechanical properties of austenitic stainless steels with the use of synthetic data via generative adversarial networks," In *The 7th Mechanical Engineering, Science And Technology International Conference*, Mdpi, Feb. 2024, P. 4. Doi: 10.3390/Engproc2024063004.

[13] s. mohammad, R. Akand, K. M. Cook, S. Nilufar, And F. Chowdhury, "Leveraging deep learning and generative ai for predicting rheological properties and material compositions of 3d printed polyacrylamide hydrogels," *Gels*, Vol. 10, No. 10, P. 660, Oct. 2024, Doi: 10.3390/Gels10100660.

[14] X. Sun, K. Zhou, S. Shi, K. Song, And X. Chen, "A new cyclical generative adversarial network based data augmentation method for multiaxial fatigue life prediction," *Int. J. Fatigue*, Vol. 162, P. 106996, Sep. 2022, Doi: 10.1016/J.Ijfatigue.2022.106996.

[15] A. R. Prabowo, R. Ridwan, T. Tuswan, J. M. Sohn, E. Surojo, And F. Imaduddin, "Effect of the selected parameters in idealizing material failures under tensile loads: benchmarks for damage analysis on thin-walled structures," *Curved Layer. Struct.*, Vol. 9, No. 1, Pp. 258–285, Jan. 2022, Doi: 10.1515/Cls-2022-0021.

[16] V. W. Berger And Y. Zhou, "Kolmogorov–Smirnov test: Overview," In *Wiley Statsref: Statistics Reference Online*, 1st Ed., R. S. Kenett, N. T. Longford, W. W. Piegorsch, And F. Ruggeri, Eds., Wiley, 2014. Doi: 10.1002/9781118445112.Stat06558.

[17] D. Leni, A. Karudin, M. R. Abbas, J. K. Sharma, And A. Adriansyah, "Optimizing stainless steel tensile strength analysis: through data exploration and machine learning design with streamlit," *Eureka Phys. Eng.*, No. 5, Pp. 73–88, Sep. 2024, Doi: 10.21303/2461-4262.2024.003296.

[18] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, And P. N. Suganthan, "Ensemble deep learning: A Review," *Eng. Appl. Artif. Intell.*, Vol. 115, P. 105151, Oct. 2022, Doi: 10.1016/J.Engappai.2022.105151.

[19] A. Karudin, D. Leni, R. Lapisa, Y. P. Kusuma, And M. R. Abbas, "Design of tools for visualizing thermodynamic concepts in steam power plant trainer processes with web-based exploratory data analysis (Eda)," *Joiv Int. J. Inform.*

Vis., Vol. 8, No. 3, P. 1134, Sep. 2024, Doi: 10.62527/Joiv.8.3.2139.
[20] V. Chak, H. Chattopadhyay, And T. L. Dora, "A review on fabrication methods, reinforcements and mechanical properties of aluminum matrix composites," *J. Manuf. Process.*, Vol. 56, Pp. 1059–1074, Aug. 2020, Doi: 10.1016/J.Jmapro.2020.05.042.
[21] H. R. Kotadia, G. Gibbons, A. Das, And P. D. Howes, "A review of laser powder bed fusion additive manufacturing of aluminum alloys: microstructure and properties," *Addit. Manuf.*, Vol. 46, P. 102155, Oct. 2021, Doi: 10.1016/J.Addma.2021.102155.
[22] A. T. Gnanha, W. Cao, X. Mao, S. Wu, H.-S. Wong, and Q. Li, "The residual generator: An improved divergence minimization framework for gan," *Pattern Recognit.*, Vol. 121, P. 108222, Jan. 2022, Doi: 10.1016/J.Patcog.2021.108222.
[23] A. Figueira And B. Vaz, "Survey on synthetic data generation, evaluation methods and gans," *Mathematics*, Vol. 10, No. 15, P. 2733, Aug. 2022, Doi: 10.3390/Math10152733.
[24] M. Allahyani *Et al.*, "Divgan: A diversity enforcing generative adversarial network for mode collapse reduction," *Artif. Intell.*, Vol. 317, P. 103863, Apr. 2023, Doi: 10.1016/J.Artint.2023.103863.
[25] D. Saxena And J. Cao, "Generative adversarial networks (Gans): Challenges, solutions, and future directions," *Acm Comput. Surv.*, Vol. 54, No. 3, Pp. 1–42, Apr. 2022, Doi: 10.1145/3446374.
[26] P. Samal, P. R. Vundavilli, A. Meher, And M. M. Mahapatra, "Recent progress in aluminum metal matrix composites: A review on processing, mechanical and wear properties," *J. Manuf. Process.*, Vol. 59, Pp. 131–152, Nov. 2020, Doi: 10.1016/J.Jmapro.2020.09.010.
[27] A. Borji, "Pros and cons of gan evaluation measures: new developments," *Comput. Vis. Image Underst.*, Vol. 215, P. 103329, Jan. 2022, Doi: 10.1016/J.Cviu.2021.103329.
[28] D. Leni, "The influence of heatmap correlation-based feature selection on predictive modeling of low alloy steel mechanical properties using artificial neural network (Ann) algorithm," *J. Energy Mater. Instrum. Technol.*, Vol. 4, No. 4, Pp. 152–162, Nov. 2023, Doi: 10.23960/Jemit.V4i4.203.
[29] J. Zhang, Z. Liu, W. Jiang, Y. Liu, X. Zhou, And X. Li, "Application of deep generative networks for sar/isar: A Review," *Artif. Intell. Rev.*, Vol. 56, No. 10, Pp. 11905–11983, Oct. 2023, Doi: 10.1007/S10462-023-10469-5.
[30] Y. Lu *Et Al.*, "Machine learning for synthetic data generation: A review," 2023, *Arxiv*. Doi: 10.48550/Arxiv.2302.04062.
[31] F. K. Dankar And M. Ibrahim, "Fake it till you make it: Guidelines for effective synthetic data generation," *Appl. Sci.*, vol. 11, no. 5, p. 2158, feb. 2021, doi: 10.3390/app11052158.