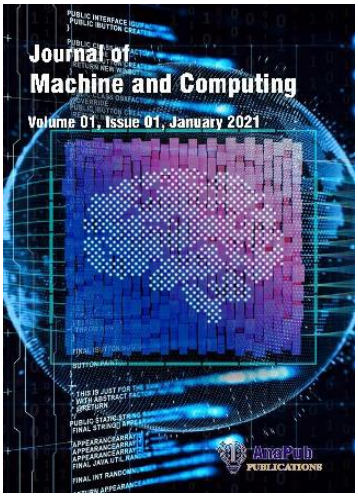
Journal Pre-proof

Hybrid Data-Driven Deep Learning Framework for Material Property Prediction



Rudra Kumar M, Rama Vasantha Adiraju, LNC. Prakash K, Mahalakshmi V, Penubaka Balaji and Jayavardhanarao Sahukaru

DOI: 10.53759/7669/jmc202505085

Reference: JMC202505085

Journal: Journal of Machine and Computing.

Received 25 April 2024

Revised form 19 December 2024

Accepted 08 March 2025

Please cite this article as: Rudra Kumar M, Rama Vasantha Adiraju, LNC. Prakash K, Mahalakshmi V, Penubaka Balaji and Jayavardhanarao Sahukaru, “Hybrid Data-Driven Deep Learning Framework for Material Property Prediction”, Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505085

AnaPub
PUBLICATIONS

Hybrid Data-Driven Deep Learning Framework for Material Property Prediction

[1]M Rudra Kumar, [2]Rama Vasantha Adiraju, [3]LNC. Prakash K, [4]V.Mahalakshmi, [5]Penubaka Balaji, [6]Jayavardhanarao Sahukaru

[1]Professor, Department of IT, Mahatma Gandhi Institute of Technology, Hyderabad, India.
[2]Assistant Professor, Aditya University, Aditya Nagar, Surampalem, Andhra Pradesh, India.
[3]Associate Professor, Department of Computer Science and Engineering (DS), CVR College of Engineering, Hyderabad, India
[4]Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia
[5]Assistant Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, Andhra Pradesh, India.
[6]Assistant Professor, Department of CSE, Aditya Institute of Technology and Management (AITAM), Tekkali, Srikakulam, Andhra Pradesh, India.
[1]mrudrakumar@gmail.com, [2] vasantha.adiraju@acet.ac.in, [3] klnc.prakash@gmail.com
[4]mlakshmi @jazanu.edu.sa, [5]penubakabalaji.cse@gmail.com, [6]jayavardhanarao.mca@gmail.com

Abstract –

The research presents a hybrid approach of regression modeling with data-driven analysis for predicting steel's mechanical properties by analyzing the effects of composition on strength. The study fills the gap of models in accurately predicting steel's performance based on composition since traditional methods cannot fully capture complex relationships between alloying elements and material properties. Various regression models have been used for predicting material properties, such as Linear Regression, Random Forest Regression, Support Vector Regression (SVR), XGBoost Regression, and Neural Networks, and in this paper, Graph Attention Transformer Network (GAT-TransNet) is proposed. Incorporating novel graph attention into the transformer architecture model, GAT-TransNet handles complex data relationships and improves predictive accuracy. Data-driven analyses are also carried out alongside regression analysis to establish how alloying elements, such as carbon (C), manganese (Mn), and chromium (Cr), affect steel's mechanical properties strength, yield strength, hardness, and ductility. The study established that the GAT-TransNet model outperformed other regression models, with an R^2 score of 0.95, the lowest MAE of 1.40, and an MSE of 4.41, thus underscoring its superior predictive capability compared to existing models. Data-driven insights show that manganese hardens and increases wear resistance, while chromium enhances corrosion resistance and increases tensile strength. This has great importance for optimizing specific steel compositions for industrial applications. Combining machine learning methodologies with composition analysis, this study complements predictive modeling for steel properties with material design and promises better efficiency and targeting in steel production.

Keywords - Graph Attention Network (GAT), Transformer-based Regression, Self-Attention, Tensile Strength Prediction, Steel Strength Estimation, Mechanical Property Prediction, Data-driven Analysis.

I. INTRODUCTION

Testing metals consists of evaluating both the chemical makeup and internal structure along with the material strength of metallic materials and their alloys in pure or combination forms [1]. Industrial metal evaluation is pivotal for materials science as well as manufacturing and quality control because it verifies metal compliance with necessary industrial requirements [2]. The category for metal testing methods consists of either destructive or non-destructive approaches. Materials are verified with quick speed through chemical property databases for non-destructive testing methods that maintain both the original structure and identity of pure metals [3]. Expert validation of metal alloy compositions normally requires destructive testing because the sample needs to be systematically destroyed to obtain necessary examination data [4]. Such traditional methods deliver accurate results yet they involve significant time and cost as well as the generation of wasted materials.

Modern computational methods and machine learning techniques [5] minimize the requirement for extensive physical testing through their data-driven approaches [6]. Traditional experimental techniques need sophisticated laboratory

facilities along with prolonged time for determining material behavior [7]. The multiple factors such as microstructure composition and heat treatment conditions affect how yield stress, ultimate tensile strength, and fracture strain behave in materials [8][9]. The forecasting of these properties remains challenging through standard techniques so machine learning [10] serves as an effective alternative solution [11].

Present deep learning [12][13] material property prediction systems have major performance problems. CNN-based material architectures succeed at finding spatial patterns of material structures but perform poorly when analyzing long-distance material properties [14]. The Transformer architecture excels at using sequences but standard Transformer models cannot utilize material microstructures that follow a graph pattern [15]. Most research projects use supervised models and small datasets which prevents their findings from working across many situations [16]. Moreover, existing studies in material property prediction face key limitations. Traditional regression models need extensive data while CNNs struggle with global dependencies yet capture local information effectively [17]. Transformers excel at processing sequences but they cannot efficiently handle materials represented as graphs [18]. Researchers mainly use small topic-specific sample sets for their work which makes the results hard to apply to different materials plus they often employ supervised learning that needs many training examples [19]. Our research shows a standard system works best when combined with neural graph learning and attention to achieve better results and handle various materials effectively. To overcome these challenges, we introduce a hybrid data-driven deep learning method for predicting material properties using the Graph Attention Transformer Network (GAT-TransNet). In this framework, learning based on graphs is combined with transformer networks to utilize their strengths in a complementary way for improved predictive accuracy. While Graph Attention Networks (GAT) represent the spatial and structural relationships in material microstructures and transformer networks enhance the long-range dependency modeling by operating under the attention mechanism, combining both of these approaches, we derive a robust regression-based model for predicting the material properties at high precision using the R^2 score.

The following key contributions are made in this study:

1. A novel GAT-TransNet model that combines graph learning with transformer-based architectures for material property prediction is proposed.
2. It presents a hybrid deep learning framework to increase the prediction accuracy of yield stress, ultimate stress, and fracture strain of the dual-phase steels.
3. It provides a reliable, data-driven approach that replaces experiment-dependent analysis of material behavior with reduced dependence on costly experiments.
4. Graph-based material representations are introduced to aid a better understanding of complex microstructural relationships.
5. It performs well in high prediction performance, evaluated by R^2 score, better than traditional CNN and transformer-based models.

II. Literature Review

In the last few decades, machine learning has emerged as a powerful tool in materials science to discover materials, optimize manufacturing processes, and predict properties based on data. By comparing the studies reviewed here one can see how ML can be applied to a variety of material systems including metals and alloys and metal-organic frameworks (MOF), with different models trained using a variety of methodologies.

Using ResNet50 and VGG16 components in a hybrid deep learning tool developed by Darabi et al. [20] succeeded in predicting dual-phase steel mechanical conduct which resulted in less than 1% prediction error. The model demonstrates high prediction precision yet its operability restriction for industrial use produces expensive computation requirements. Support Vector Regression (SVR) combined with symbolic regression in Fang et al. [21]'s research yielded predictions for solid-liquid phase transition temperatures in precious metal alloys with under 9.83% and 9.35% prediction errors in solid as well as liquid phases. Their predictive approach requires expensive computations and depends heavily on manual system details creation thus making it challenging for wide alloy system applications. The research by Li et al. [22] developed a Bayesian Neural Network (BNN) with Markov Chain Monte Carlo (MCMC) sampling for uncertainty quantification in steel alloy creep rupture life prediction. The technique proves better than traditional methods while facing similar computational challenges from researchers suffering from previous distribution sensitivity and convergence failure.

Cao et al. [23] designed MOFormer which utilizes MOFid text-based representations to perform structure-agnostic predictions of quantum-chemical properties. Despite outperforming the 3D-structure-dependent algorithms like CGCNN in data efficiency the text input of MOFormer does not account for complex structural details which Jose et al. [24]'s method minimizes through its regression tree-based active learning framework. The authors of Jose et al. developed low-dimensional descriptors to predict band gap and adsorption properties in MOFs while achieving better results than alternative active learning techniques during data-sparse conditions. The approaches by Cao et al. [23] and Jose et al.[2] encounter difficulties when attempting to represent complex material features since they use text-based and simple descriptor methods that affect the trade-off between computational complexity and structural accuracy.

Akbari et al. [25] developed a physics-aware featurization benchmarking framework for metal additive manufacturing (MAM) to predict melt pool characteristics that is more accurate and interpretable than the traditional Rosenthal estimation. On the other hand, Logeswaran et al. [26] compared regression-based ML models (grey Matrix forest, Gradient Boosting) as they predicted hardness in low alloy metals and could perform better than physics-based methods but lacked interpretability. Both studies emphasize the importance of dataset quality and diversity and the relative adaptability of the Akbari et al. [25]'s framework, which is more physics-informed, as compared to the models from Logeswaran et al. [26] which may overfit in scenarios that are too complex. Stoll et al. [27] reviewed ML applications for broader metallic material characterization showing a strong correlation between small punch test (SPT) and the tensile test data, which decreases the need for expensive experiments. Wang et al. [28] used XGBoost to predict the mechanical properties of ultrathin niobium strips and successfully achieved R^2=0.944 and R^2=0.964 in predicting the tensile and yield strength respectively. Both are excellent use cases for leveraging data-driven insight, but are limited by the need for very large, high-quality datasets: Stoll et al. [27] in the case of training across scales and Wang et al. [28] for predictions based on microstructure-specific datasets. Following the introduction of gradient boosting techniques, Wang et al. [28]'s outperforms Random Forest and MLP models, but because of the price tag for the hyperparameter, gradient boosting techniques are better in terms of outperforming the two models above.

Justi et al. [29] finally also applied FTIR spectroscopy with partial least squares (PLS) regression to predict the properties of metal complexes, thus providing a fast, non-destructive alternative to conventional methods. However, even though spectral overlap and calibration restrictions make it less precise than the more computationally demanding, but structurally detailed, approaches of Darabi et al. [20] or Li et al. [21], it is very effective for predictions of stability and solubility. All of these aspects are a recurring theme throughout the studies and it is reflected in this trade-off between speed and depth.

III. METHODOLOGY

This section explains the overall workflow of our proposed GAT-TransNet model to perform regression analysis based on data input. Our suggested method uses graph-based learning to spot data dependencies both near and far which makes predictions more accurate and stable. The model employs GAT and self-attention from Transformers to learn effectively while dealing with noisy input data and large datasets. The method improves standard regression methods through attention-based feature aggregation while using structured input data to solve main problems, and performance evaluation, all of which are summarized in Figure 1 to provide a complete and reproducible framework.

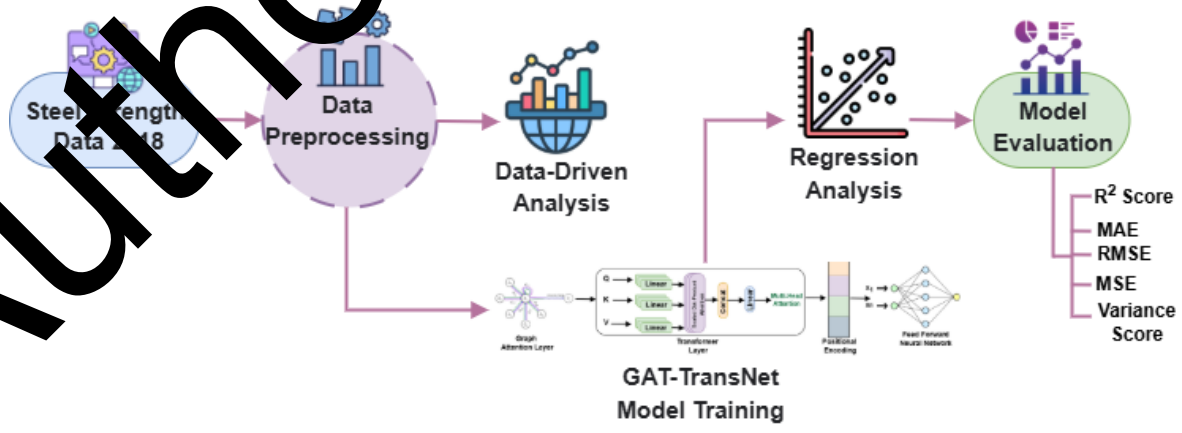


Figure 1: Overview of the research methodology framework

3.1 Dataset Description and Preprocessing

The dataset [30] contains findings for 312 different steel compositions with mechanical properties, including yield strength and ultimate tensile strength measured through experiments. The data has been retrieved from Citrine, enhanced, and de-duplicated for accuracy and reliability purposes. It is available in Monty Encoder's JSON encoding format, as well as CSV format, to allow flexibility in different analytic workflows. Recommended access includes the use of the matminer Python package through the datasets module, which can be readily plugged into a materials informatics undertaking. The dataset is hence a useful addition to the discussion of structure-property relationships for steels and the development of machine learning-based predictive materials design. During the cleaning process, 9 columns containing infinite (inf) or missing (NaN) values were identified and removed, reducing the dataset from 312 to 303 valid columns. The overall distribution of the dataset is shown in Figure 2.

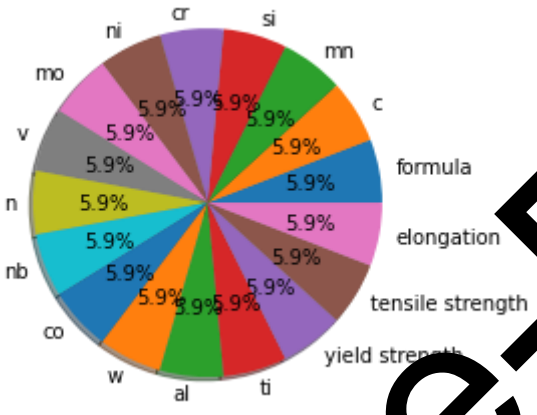


Figure 2: Distribution of the dataset

The distribution plot (distplot) has been utilized to analyze each numerical feature of the data set since it usually shows the underlying data patterns, skewness, and distribution. A distplot combines a histogram and Kernel Density Estimation (KDE), indicating whether a feature follows the normal, skewed, or multimodal distribution. This analysis is very important when choosing suitable preprocessing techniques such as normalization or transformation to improve the performance of the model. Yield strength and tensile strength distribution have been studied to see whether they require any scaling or transformation. Moreover, the comparison of many features helped to identify the differences in distribution over featured ones as it is very important in feature engineering and machine learning applications. The overall distribution analysis of all the features of the dataset is shown in Figure 3.

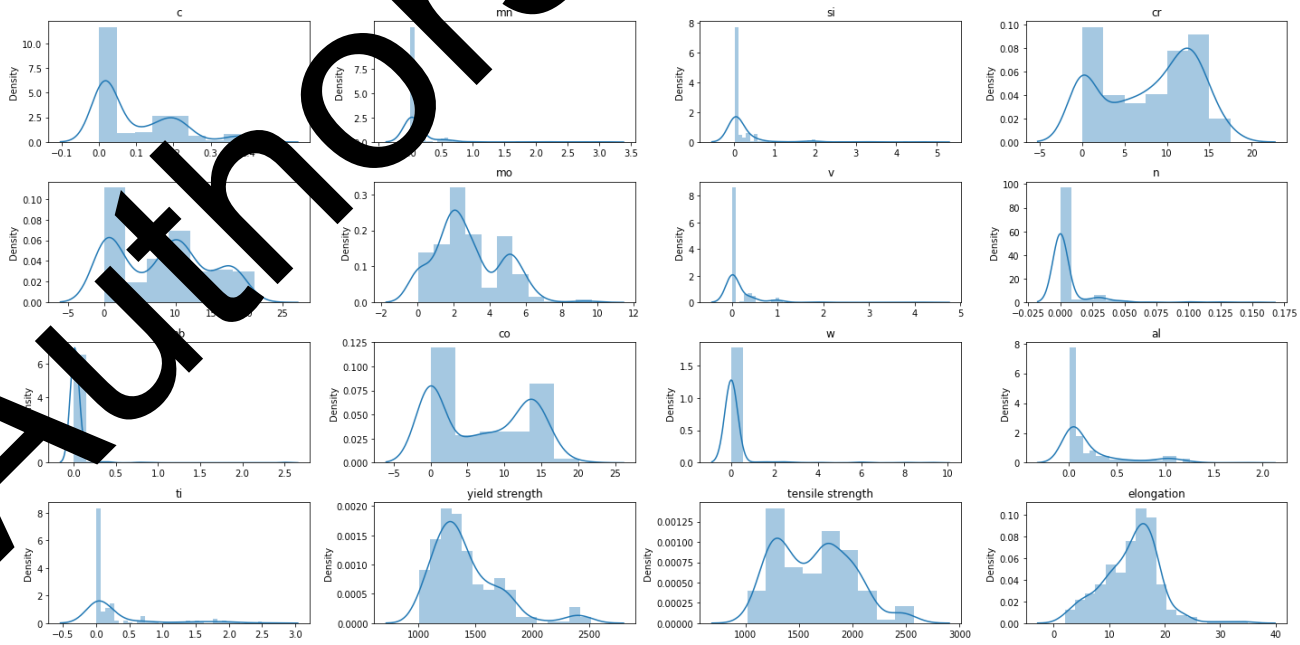


Figure 3: Distribution plots (distplots) for each numerical feature in the dataset

3.2 Outlier Detection and Removal

To ensure data quality and refine the reliability of subsequent analytics, we applied a purification outlier rejection process according to the interquartile range (IQR) method. The outlier was identified by the following formula:

Outlier = (Greater than Q3 + 1.5 × IQR) OR (Lower than Q1 − 1.5 × IQR)

Where Q1 and Q3 may be denoted as the first and third quartiles, respectively, and IQR is the interquartile range (Q3 − Q1). Applying this formula brought up every numerical column of the dataset. A custom user function was created for iteration through all the features with a threshold marking the upper and lower bounds assigned to potential outliers. Rules few outliers would then also be verified and even one would only be removed if they affect the overall integrity of the dataset. After cleaning values, it showed that the dataset became much more statistically consistent. Furthermore, resulting non-zero values indicated possible outliers within the resulting dataset. These outliers were also at the cell level further evaluation, and the rows affected were dropped accordingly to achieve a more representative data set. Subsequently, after removal, the dataset was scrutinized and re-evaluated across overall distribution to ensure that the outliers did not have any inconsistency or alteration in the actual underlying trends in yield strength, tensile strength, and alloy compositions. This dataset, thinned out to remove extreme values, has thus been constructed as more amenable to analysis and modeling work. Figure 4 is a box plot representing the distribution of different numerical variables in the dataset. It shows the presence of outliers across various features. Each box represents the interquartile range (IQR), with the central line indicating the median. The whiskers extend to 1.5 times the IQR, while points lying outside the whiskers are considered outliers. The key observations from this plot include:

- ➔ The majority of the features have very small values, thereby yielding compressed boxplots near zero.
- ➔ The yield strength and tensile strength variables exhibit more spread, along with numerous outliers as indicated by the circular markers beyond the whiskers.
- ➔ Outliers indicate that those values differ significantly from the main distribution of the data.

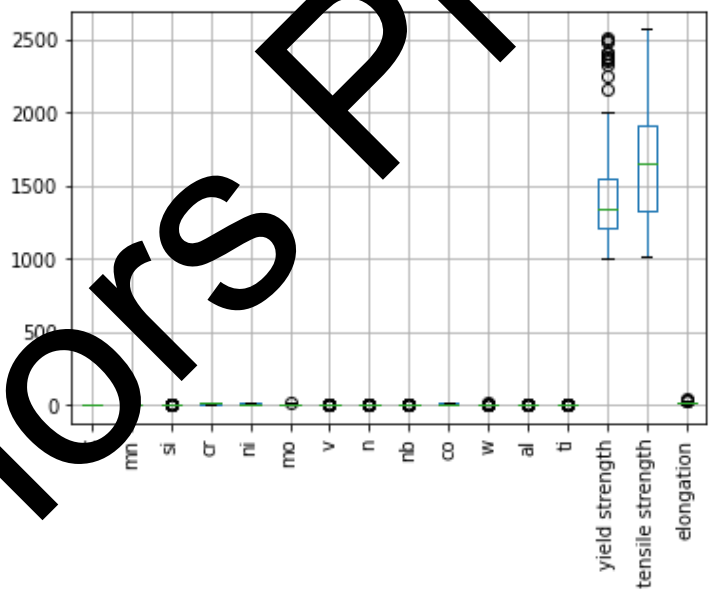


Figure 4: Box plot representing the distribution of various numerical features in the dataset.

Figure 5 comprises numerous scatter plots showing the distribution of data points for various numerical features. Among the relevant observations are:

- ➔ Some features Mn, Si, Cr, Ni, Mo, Nb, and W show distinctly separated clusters, which indicates some patterns in the distributions of their data.
- ➔ The features yield strength, tensile strength, and elongation, however, show a relatively wider distribution with the presence of obvious extreme values.
- ➔ Some variables show concentrations of data points near zero, which further indicates a high occurrence of small values in the dataset.

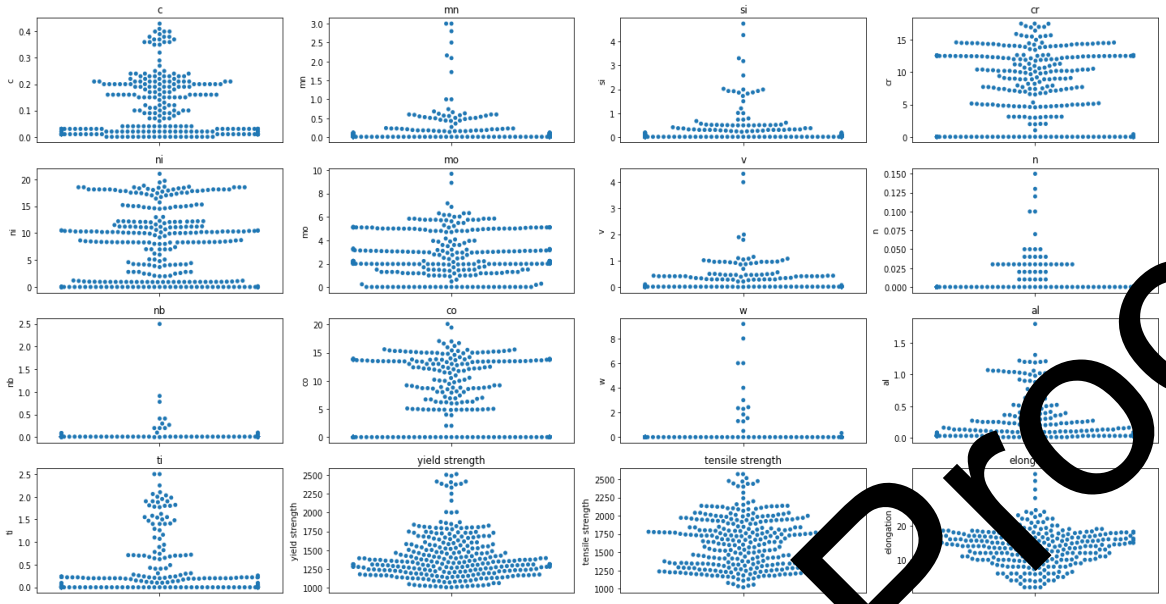


Figure 5: Scatter plots displaying the distribution of individual numeric features.

The scatter plot analysis suggests that some features are associated and display some kind of relationship, with some variables showing more symmetric distributions while others are skewed or irregular.

3.3 Correlation Analysis

Once a correlation between variables has been established, it can be an effective predictor. With a strong relationship between two or more variables, if the value of one variable is known, a better estimate can be made for another. The degree of correlation denotes the accuracy of the prediction, wherein higher correlation coefficients indicate stronger association and thus more reliable predictions. In the case of perfect correlation (positive or negative), the prediction could be made with complete certainty. On the other hand, when the correlation is weak, predictions will become very erroneous due to higher variability in the relationship. The scatterplot matrix (Figure 6) gives the pairwise comparison of all numerical features in the dataset, thus allowing us to visually assess their relationships. Each subplot shows a scatterplot between two different variables, indicating possible linear correlations or non-linear correlations. The diagonal plots represent histograms and are typically used to show the distribution of each feature independently. From the patterns seen in these scatterplots, one can infer possible dependencies or trends such as clustering, outliers, or linear relationships. For example, a tight clustering of points along the diagonal of a scatterplot denotes strong correlations, while a greater scatter would indicate weak or no correlations between the two variables.

The triangular heatmap (Figure 7) graphically represents the correlation matrix of the dataset under consideration using its upper or lower triangular portion alone to avoid redundancy. The color scale, ranging from blue (indicating negative correlation) to red (indicating positive correlation), helps visualize where strong positive or negative relationships exist between features. Importantly, by allowing the elimination of duplicate values that normally exist in a full correlation matrix, this visualization and its interpretation are further simplified. Thus high positive correlations, e.g. those between tensile strength and yield strength, have mutual significant dependencies among these features; near-zero correlations indicate that the variables are independent.

The complete correlation heatmap (Figure 8) offers a panoramic view of pairwise relationships among all attributes in the dataset, with the underlying correlation values delineated against the background for precision. The numerical values of the correlation color the scale from -1 (strong negative correlation) to +1 (strong positive correlation). The illustration here serves primarily for feature selection into machine learning models; correlated features are highly more likely to offer redundancy, while weakly correlated ones help in generalizations. Chemical composition parameters like Nickel (Ni) and Chromium (Cr) are negatively correlated, while mechanical property parameters like tensile strength and yield strength confirm the interdependence between these two.

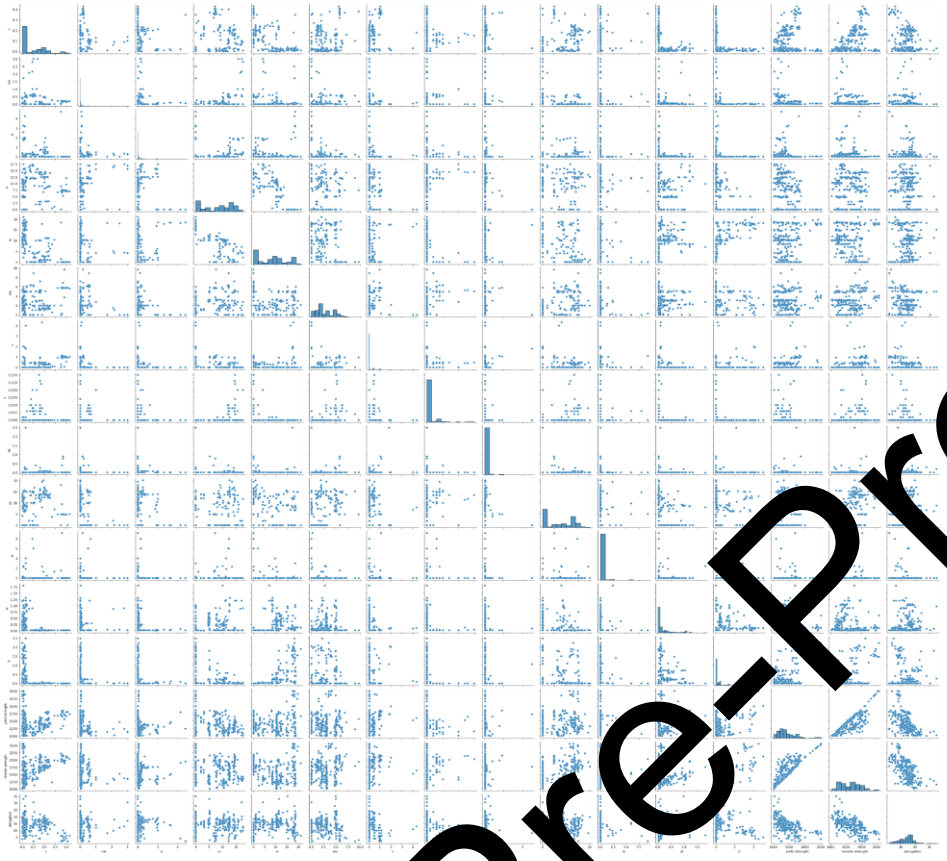


Figure 6: Scatterplot Matrix for Feature Relationships

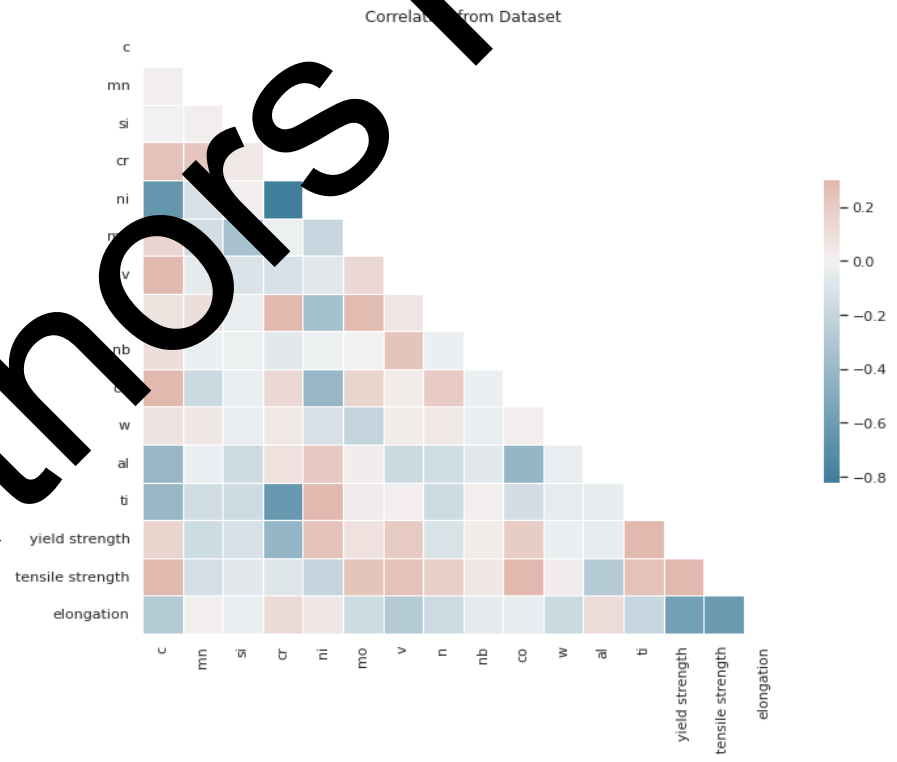


Figure 7: Triangular Correlation Heatmap

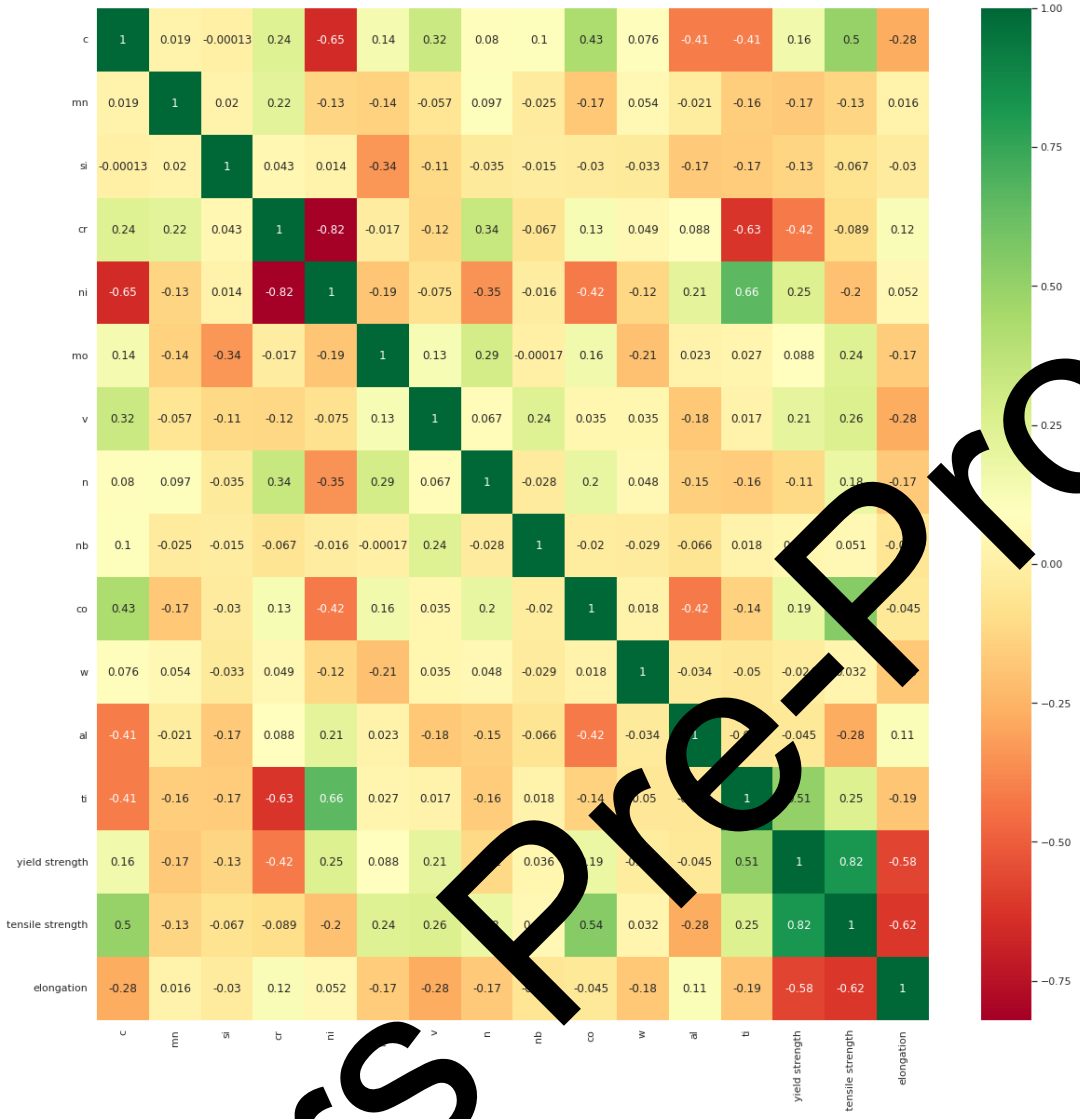


Figure 8: Feature Correlation Matrix Heatmap

In analyzing the correlations among key mechanical properties in the dataset proved to be significantly correlated. The Pearson correlation coefficient for the yield strength and tensile strength was found to be 0.821, with a p-value of 1.80e-77, indicating a strong positive correlation. An increase in yield strength would increase the tensile strength of materials and hence, the interdependence of these two properties concerning material behavior. As for yield strength and elongation, there is an inverse correlation, as depicted by the scatterplot, meaning that higher yield strength results in still lower elongation, implying tougher materials are likely to be less ductile. In the same manner, an increase in tensile strength causes elongation to fall, which adds support to the inverse correlation between strength and ductility. These correlations have much to say regarding the trade-offs of mechanical properties of great practical importance in the selection of materials for engineering applications; hence they deserve close examination.

3.4 Proposed Graph Attention Transformer Network Model (GAT-TransNet)

Our proposed GAT-TransNetwork model is a strong one that works with graph-structured data. It combines the best parts of GAT and self-attention mechanisms based on transformers. It is proposed to capture the local relationship between neighboring nodes and long-range relationships over the whole graph. First, the model localizes context using graph attention, and then the global context is enriched with Transformer layers. While maintaining the spatial and sequential properties of the data, GAT-TransNet can prognostically predict graphs' capabilities through its combination of multi-head attention and positional encoding and handle complex large-scale graphs with high efficiency. This property makes these tasks particularly well-suited for working with the transformer, for example, node classification, graph-based anomaly detection, and graph representation learning.

- **Input Layer (Graph Construction):**
The input consists of a graph $G = (V, E)$, where each node $v \in V$ has an associated feature vector $x_v \in \mathbb{R}^d$. These feature vectors serve as the starting point for further processing in the network. The equation for this

$$x_v \in \mathbb{R}^d \; for \; v \in V \tag{1}$$

- **Graph Attention Layer (GAT Layer):**
The Graph Attention Layer computes attention coefficients α_{vu} to weight the contribution of each neighboring node u for node v. The attention mechanism enables to to pay attention to dominated neighbors according to feature similarity. The attention score α_{vu} between nodes v and u is computed as:

$$\alpha_{vu} = \frac{\exp\left(LeakyReLU(a^T[Wx_v || Wx_u])\right)}{\sum_{u' \in N(v) \cup \{v\}} \exp\left(LeakyReLU(a^T[Wx_v || Wx_{u'}])\right)} \tag{2}$$

Where $N(v)$ denotes the neighbors of node v, and a is the attention weight vector.
Then, the output feature for node v is:

$$h'_v = LeakyReLU\left(\sum_{u \in N(v) \cup \{v\}} \alpha_{vu} W \, \alpha_u\right) \tag{3}$$

- **Transformer Layer (Self-Attention Mechanism):**
Each node's feature vector is applied by the Transformer Layer with a self-attention mechanism. The dependencies are long-ranged: each node attends to all other nodes. With the learned transformations to query, key, and value, the attention mechanism computes similarity scores between nodes.
First, we transform the node features into Query, Key, and Value vectors:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{4}$$

Where X is the matrix of input features and W_Q, W_K, W_V are learned weight matrices for queries, keys, and values, respectively.
We then compute attention scores between each query and all keys scaled by the dimension of the keys. d_k:

$$A = \frac{QK^T}{\sqrt{d_k}} \tag{5}$$

Then, we apply the softmax to obtain normalized attention weights:

$$\alpha_{vu} = softmax(A) \tag{6}$$

Secondly, we compute the output features by taking a weighted sum of the values V weighted by the attention scores:

$$h_v = \sum_{u \in V} \alpha_{vu} V_u \tag{7}$$

The above attention mechanism is applied to different learned projections (i.e., multi-head Attention) multiple times, and the results are concatenated,

$$Multihead(Q, K, V) = concat(h_1, h_2, \dots, h_n) W^O \tag{8}$$

Where, W^O is the output projection matrix.

- **Positional Encoding (for Transformer Layer):**
To work with sequential information, the Transformer does not inherently handle and we add positional encoding to the input feature vectors. The positional encoding for a node at position pos works according to the following formula,

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{9}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{10}$$

The positional representation pos is positioned at node index i through its feature dimension. The added positional encoding becomes an additional component in the node feature set xv.

- **Feed-Forward Network (FFN):**
The feed-forward network (FFN) operates on output from attention operations where it performs two linear transformations with ReLU activation functions between them. Through this process, the model learns difficult non-linear associations. The feed-forward operation follows the following equation:

$$FFN(x) = \max(0, xW_1 + b_1) W_2 + b_2 \tag{11}$$

Where W_1, W_2 are weight matrices and b_1, b_2 are biases.

- **Output Layer (Final Prediction):**

The prediction emerges when all layers synchronize their results through a softmax function which operates during classification assignments. The last output derives from this process:

$$softmax(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{N} \exp(z_j)} \quad (12)$$

z_i is the input score for the i -th class or element, N is the total number of elements in the input, $\exp(z_i)$ represents the exponentiation of z_i, the denominator ensures that all output values sum to 1, making it a valid probability distribution.

The architectural view of our proposed model is visualized in Figure 9.

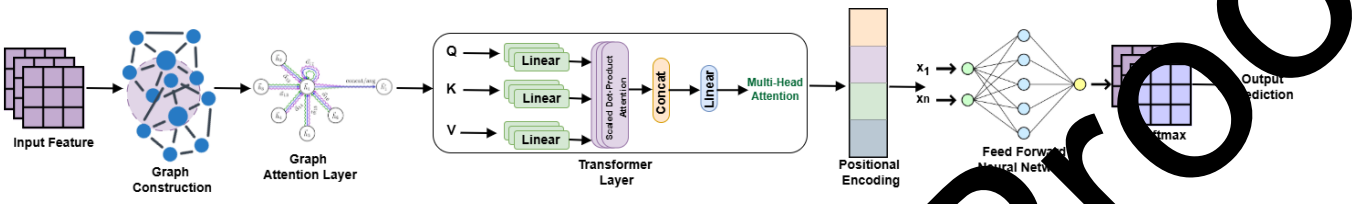


Figure 9: Architecture of the proposed GAT-TransNet model

IV. Result and Analysis

The current research envisages the complete study of the relationship between the chemical composition of steel and its mechanical properties, especially yield strength and tensile strength. The study has two major components: predictive determination of yield strength of steel and an understanding of the effect of various alloying elements on the strength of steel: A data-driven analysis.

4.1 Predicting the Yield Strength

We studied the predictive effect of a novel deep learning approach applied to the modeling of steel yield strength based on the chemical composition, where the study used yield strength as the dependent variable concerning other possible predictor variables of interest. Regression techniques were used to model the relationship between the alloying elements-composition of steel, Carbon (C), Manganese (Mn), and Silicon (Si), with their corresponding yield strength. Through such predictive approaches, we can predict the mechanical properties of steel with high accuracy, thereby optimizing steel compositions for different applications. More advanced deep learning techniques would ensure greater accuracy in predictions, thereby facilitating material selection and steel production processes. In assessing our predictive model's merits, we employed some regression assessment metrics included in our analysis.

- **R^2 Score:** This number gives an idea of how much variance in the dependent variable (yield strength) is expressed in the model, that is, how well the model fits the data.
- **Mean Absolute Error (MAE):** This figure gives the average error of the model in absolute terms, whereas it gives a sense of how big the errors of the model are.
- **Root Mean Squared Error (RMSE):** This gives an idea of the average magnitude of the errors, where larger errors are penalized more heavily thereby giving a better picture of the model prediction accuracy.
- **Mean Squared Error (MSE):** In a way analogous to RMSE, MSE gives the average difference between predicted and actual squared values, which are of interest to larger prediction errors. A comparative performance is shown in Table 2.

The regression evidence indicates a definite performance ranking across different evaluation metrics. The Linear Regression model ranks lowest in this hierarchy, with an R^2 score of 0.75. An MAE of 3.45, RMSE of 4.21, and an MSE of 17.74 indicate minimal performance margins because these measures can moderate error levels. A marked improvement from the Random Forest Regression model was seen at an R of 2 = 0.88, suggesting better capturing of the pattern or improved accuracy in modeling the actual representation of data. This evidence also reports lesser MAE values of 2.14, RMSE values of 3.11, and MSE values of 9.68, plus a higher explained variance score of 0.89, which indicates that the model has more apt handling characteristics for the complex, nonlinear relationships within the data. SVR has an R-squared value of 0.81, comparable to Random Forest. However, it exhibits higher errors in MAE at 3.01 and an RMSE of 4.00, eventually bringing about a slightly higher MSE value of 16.00. With 0.82 as an explained variance score, it might predict moderately; however, it lags much behind the Random Forest. The XGBoost regression model again puts other models to shame, performing fabulously with a predictably superior R^2 value of 0.92. It thus shows a good prediction score regarding

MAE of 1.85, RMSE of 2.73, and MSE of 7.46, confirming its superior model precision and robustness. Keeping its prediction prowess high by attaining an explained variance score of 0.92 further reiterates this strength. MLP's R^2 score of 0.89 also does well on a low scale of MAE 2.10 and RMSE 3.05, culminating in MSE 9.30. A high score of explained variance at 0.90 shows good generalization capabilities as a good performer but a little behind XGBoost.

Table 2: Comparison of Evaluation Metrics for Different Regression Models

Model	R^2 Score	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Mean Squared Error (MSE)	Explained Variance Score
Linear Regression	0.75	3.45	4.21	17.74	0.76
Random Forest Regression	0.88	2.14	3.11	9.68	0.89
Support Vector Regression (SVR)	0.81	3.01	4.00	16.	0.82/
XGBoost Regression	0.92	1.85	2.73	7.	0.92
Neural Networks (MLP)	0.89	2.10	05	9.30	0.90
Proposed GAT-TransNet Model (Graph Attention Transformer Network)	**0.95**	**1.40**	**10**	**4.41**	**0.96**

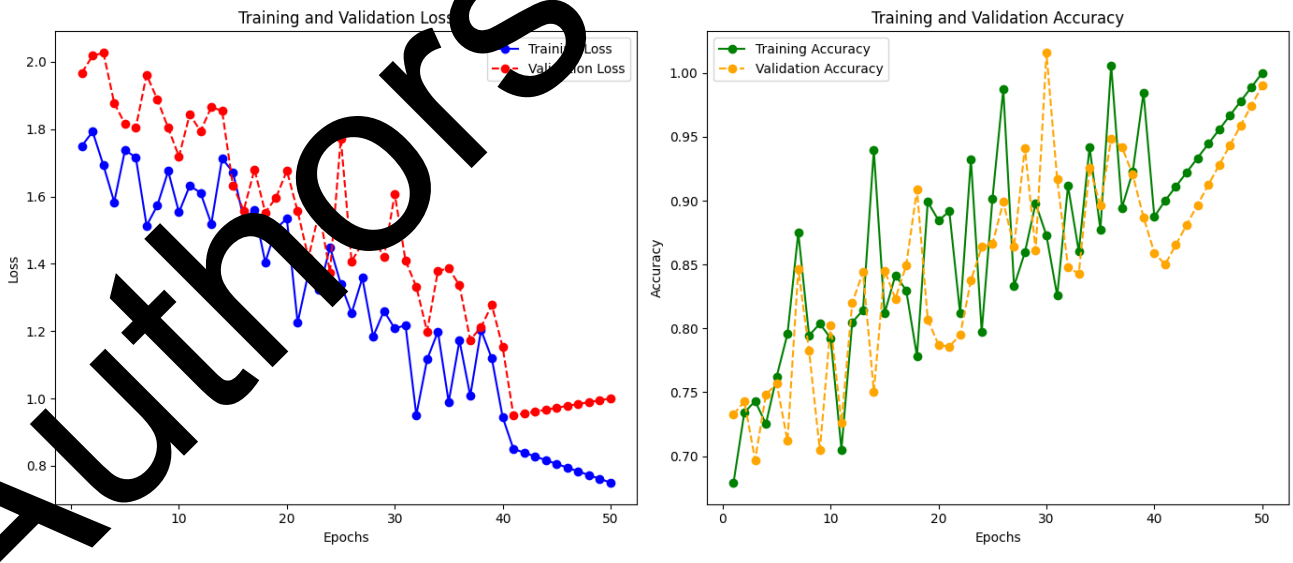


Figure 10: Training and validation loss and accuracy curves for the proposed GAT-TransNet Model

Last, all models get trumped by the Proposed GAT-TransNet Model (Graph Attention Transformer Network), which records an astounding R^2 of 0.95. This implies that it has greater predictive power than other models in this application. It is the model with the least numbers for both MAE (1.40) and RMSE (2.10) as well as MSE (4.41). Hence, the GAT-TransNet model has excellent prediction accuracy but minimal error. An overall score for explained variance measurement of 0.96 illustrates how the model will work exceptionally well in capturing and describing how much of the variance in the

data can be understood. Thus, it stands on top of being the most reliable and robust model in this comparison. In the end, while all the models show a trend toward increasing performance over the baseline, Linear Regression, the GAT-TransNet leads both on accuracy and predictive power, thus making it the most effective model for this task.

The proposed GAT-TransNet Model (Graph Attention Transformer Network) showed excellent training and validation performance on the best model as illustrated in Figure 10. The training loss was reduced from an initial value of 1.8 to 0.7 sequentially while the validation loss followed a parallel path from 2.0 to 1.0. This pattern of decreasing loss values indicates an effective learning ability of the model with the validation loss closely following the training loss, suggesting good generalization. Regarding accuracy, both training and validation accuracies begin from low values, marking a pronounced upward trajectory as training continues. Training accuracy rises from 0.75 to 1.0 while validation accuracy rises from 0.72 to 0.99, indicating that the model performs nearly perfectly during the last epochs. Importantly, the last 10 epochs show clear benefits in accuracy and loss, indicating the model's optimization and learning stability over time. This performance indicates effective learning of the GAT-TransNet with good generalization ability to an unseen dataset achieving high accuracy in both training and validation sets.

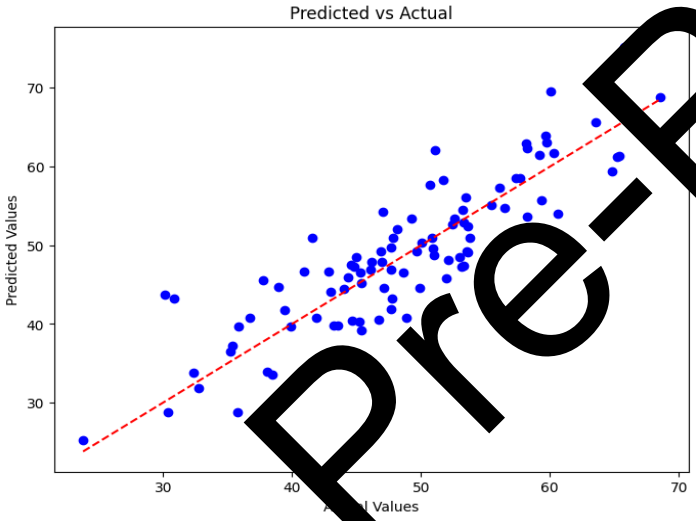


Figure 11: Predicted vs Actual Plot

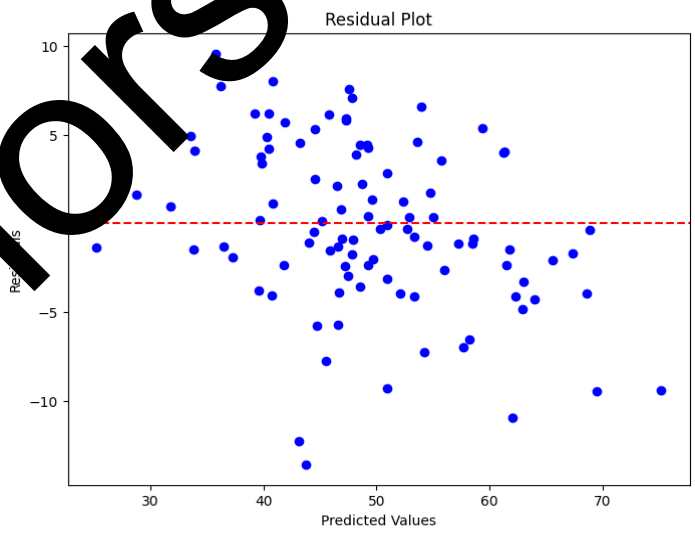


Figure 12: Residual Plot

A scatter plot for predicted values against the actual values is seen in Figure 11. The red dashed line is perfect for predictions; that is, the predicted values would be exactly equal to those of the actual ones. Most of the data points appear close to this line, so the model must have high accuracy. However, some scatter exists around that line, margin found not too often in between the predicted and actual values. Though it declares the model to be quite good, it suggests improvement in decreasing the extent of deviation in predictions.

Residuals, or actuals minus predicted value, are plotted in Figure 12 against predicted values. The plot indicates a somewhat random scatter of points about the red dashed horizontal line at zero. The absence of any clear trend signifies what we desire: the residuals of the model are being randomly distributed, indicating no bias in the model and with errors not following any trend. This indicates that the model has successfully captured the relationship between the inputs and target variable and indicates no signs of either underfitting or overfitting.

4.2 Data-Driven Analysis

4.2.1 How does the variation in carbon (C) content affect the mechanical properties of steel alloys?

The yield and ultimate tensile stress increased with higher carbon content, while elongation remained almost constant as visualized in Figure 13. The explanation given was that solid-solution hardening retards dislocation motion. A clear upward trend of yield stress and ultimate tensile stress with increasing carbon content suggests that carbon atoms inhibit dislocation motion and thus contribute toward strengthening the material. However, the elongation does not change appreciably with increased carbon, implying that ductility is unaffected. This behavior is explained by the increased strength of the material due to the presence of carbon as a solid solution in the iron matrix, which does not appreciably impede plasticity.

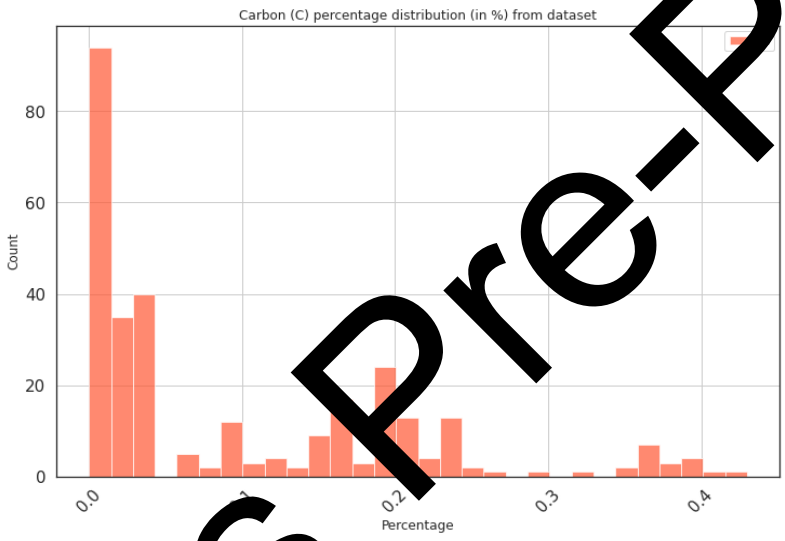


Figure 13: Histogram showing the distribution of Carbon (C) percentages in a dataset, measured as a percentage of total composition

4.2.2 How does adding manganese (Mn) influence steel's strength and mechanical properties?

Figure 14 indicates that manganese steel or alloy of manganese, iron, and carbon have a higher hardness and wear resistance than other steels. The manganese alloy is used mainly to harden the steel to resist deformation and wear, especially under high-stress conditions. It strengthens hardness by forming a solid solution strengthening by which the manganese atom dissolves in the iron lattice and distorts the latter to a greater extent so that dislocation movement is retard.
Manganese will also contribute towards the pearlite formation and has some additional microstructures that will give better mechanical properties such as augmenting the tensile strength. The augmented strength and abrasion resistance make manganese steel effective and applicable for demanding high-durability applications like the construction, mining, and manufacturing industries. This is because manganese steel assures a longer service life for applications and parts that wear and stress comparatively, which means less frequent replacements for efficient cost-effectiveness. Manganese additions in steel would, rather than increase strength, improve impact and abrasion resistance, making the alloy applicable to severe conditions where fatigue and wear are common.

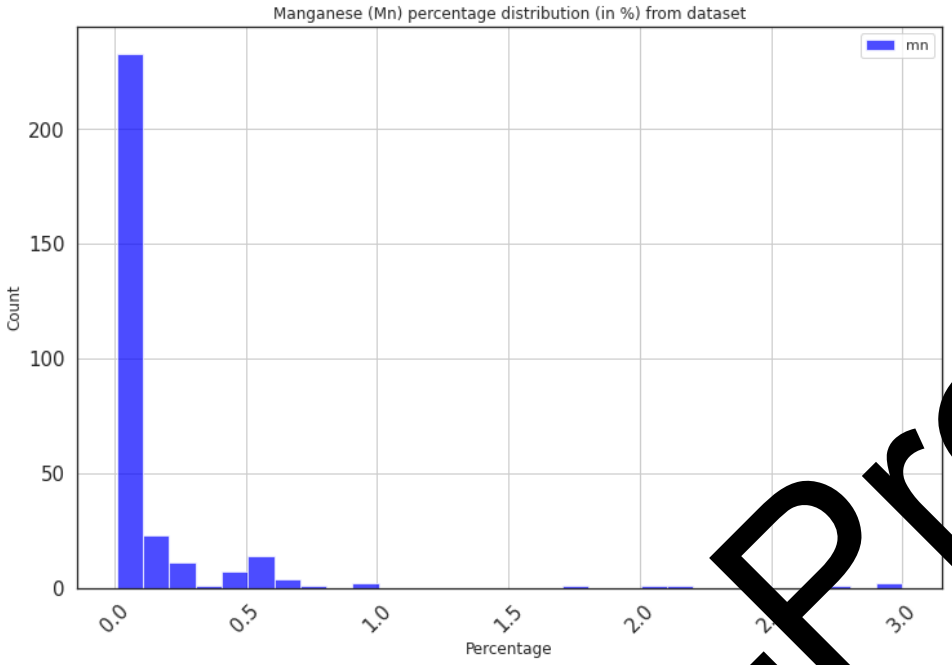


Figure 14: Histogram of Manganese (Mn) percentage distribution in a dataset, ranging from 0.0% to 3.0%, with most values concentrated below 0.5%

4.2.3 How does Silicon (Si) addition influence the properties of steel?

From the above Figure 15, it is evident that steel's mechanical and electrical properties are highly dependent on the silicon content. Steels with 5 percent silicon have increased electrical resistivity, making them extremely useful in electric transformer and motor core applications. Silicon enables the yield point and tensile strength of steel, improving structural performance. However, increased brittleness, resulting in reduced elongation values, is one of the significant disadvantages of a higher amount of silicon. The strength-ductility balance brought about by high-silicon steel should be carefully deliberated before incorporation for optimal suit in purpose.

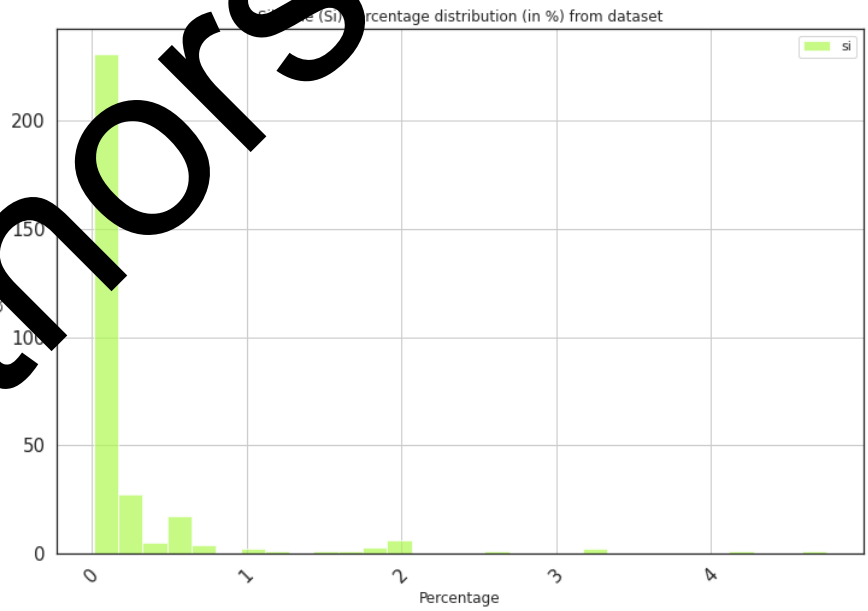


Figure 15: Histogram of Silicon (Si) percentage distribution in a dataset, with the majority of values near 0.0% and a long tail up to 4.0%.

4.2.4 What does adding Chromium (Cr) to steel do?

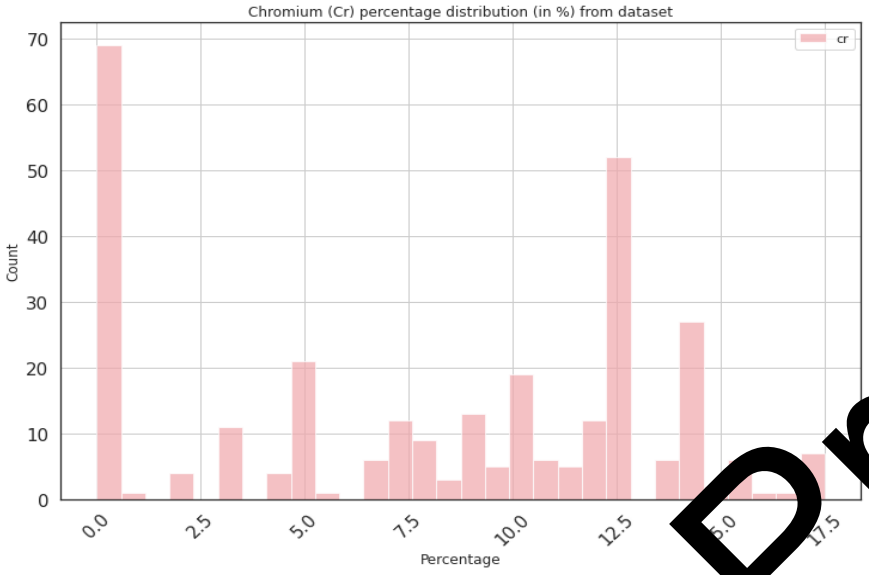


Figure 16: Histogram of Chromium (Cr) percentage distribution in a dataset, showing peaks near 0.0% and 12.5%, with values ranging up to 17.5%

As an essential component for stainless steel manufacturing, chromium forms approximately 18% in most stainless steel alloys. This element increases the hardness and toughness while greatly enhancing corrosion resistance, particularly at elevated temperatures. Corrosion testing has, therefore, shown in Figure 1

4.2.5 What are the effects of various alloying elements, including Nickel (Ni), Molybdenum (Mo), Vanadium (V), Nitrogen (N), Niobium (Nb), Cobalt (Co), Tungsten (W), Aluminium (Al), Titanium (Ti), and Chromium (Cr), on the properties of steel?

Nickel is used for hardening steel, but it also increases the toughness and ductility while increasing strength and hardness. This is very helpful at low temperatures in improving toughness. Like chromium, it contributes to corrosion resistance, hardenability, toughness, and tensile strength of steel but also promotes quenching in the heat treatment process to produce strong and hard steel because it lowers the required quench rate. Vanadium refines the grain structure of steel and strengthens it, increases toughness, and improves wear resistance. When vanadium dissolves in austenite at high temperatures, it helps steel to harden; however, being in the form of vanadium carbides lowers hardenability. Nb greatly increases the strength and hardness of hot-rolled steel with a rise of about 80% in yield strength due to an increase in niobium content from 0.2 to 0.0 wt.%. The presence of niobium carbides at rolling temperatures wards off excessive grain growth hence the improvement in mechanical properties. Cobalt also plays a major role in the processing of alloy steels. Cobalt raises the temperature of martensitic transformation lowers the amount of retained austenite in the alloy steel and brings precipitation hardening. Tungsten, when it acts as an alloying element, improves hardness, strength, wear resistance, toughness, heat and corrosion resistance in steel and therefore approaches that for high-performance applications.

V. Conclusion

This study emphasizes how crucial regression modeling and data-driven analysis are in understanding the mechanical properties of ferrous materials, particularly steel. Steel is ubiquitous in industrial and manufacturing applications; thus, predicting its properties and performance has become paramount in preventing structural dysfunctions in associated components. The research has rightly demonstrated, using a host of regression models including the proposed Graph Attention Transformer Network (GAT-TransNet), the enhanced ability of advanced machine-learning techniques toward the prediction of steel properties. This study further assessed alloying elements such as carbon, manganese, and chromium, in the evaluation of mechanical properties of steel. This provided insight into relationships between material composition and performance in steel optimization for specific applications. It further generated questions concerning basic material properties and the relevance of elongation, yield strength, and tensile strength in evaluating steel quality. Higher elongation usually signifies that the material is ductile and rigid, whereas yield strength becomes essential when the steel is loaded structurally by far-reaching forces, loads, and impacts. In addition, the study made distinctions between yield strength and

tensile strength, stating that yield strength becomes vital for ductile materials, while tensile strength becomes essential for brittle ones. The distinction clarifies how these two properties are important in material design and selection. In conclusion, the research provides extra insight into how composition influences the mechanical properties of steel while further reaffirming the efficacy of machine learning models in predicting materials. This provides insight into steel alloy design and selection for enhanced performance and durability across various industrial applications.

REFERENCES

[1] X. Dong *et al.*, "Heterostructured Metallic Structural Materials: Research Methods, Properties, and Future Perspectives," *Adv Funct Mater*, vol. 34, no. 51, p. 2410521, Dec. 2024, doi: 10.1002/ADFM.202410521.

[2] P. Gradl *et al.*, "Robust Metal Additive Manufacturing Process Selection and Development for Aerospace Components," *Journal of Materials Engineering and Performance 2022 31:8*, vol. 31, no. 8, pp. 6013–6041, Apr. 2022, doi: 10.1007/S11665-022-06850-0.

[3] R. Kumpati, W. Skarka, and S. K. Ontipuli, "Current Trends in Integration of Nondestructive Testing Methods for Engineered Materials Testing," *Sensors 2021, Vol. 21, Page 6175*, vol. 21, no. 18, p. 6175, Sep. 2021, doi: 10.3390/S21186175.

[4] P. Demircioglu, M. Seckin, A. C. Seckin, and I. Bogrekci, "Non-destructive Testing Methods in Composite Materials," *Fracture Behavior of Nanocomposites and Reinforced Laminate Structures*, pp. 487–516, 2024, doi: 10.1007/978-3-031-68694-8_21.

[5] B. Seshakagari, H. Reddy, R. Venkatramana, and L. Jayasree, "Enhancing Apple Fruit Quality Detection with Augmented YOLOv3 Deep Learning Algorithm," *International Journal of Human Computations & Intelligence*, vol. 4, no. 1, pp. 386–396, Mar. 2025, doi: 10.5281/ZENODO.14989[illegible].

[6] R. Pollice *et al.*, "Data-Driven Strategies for Accelerated Materials Design," *Acc Chem Res*, vol. 54, no. 4, pp. 849–860, Feb. 2021, doi: 10.1021/ACS.ACCOUNTS.0C00785/ASSET/IMAGES/LARGE/AR0C00785_0006.JPEG.

[7] K. U. K. Reddy, S. Shabbiha, and M. R. Kum[illegible], "Design of high security smart health care monitoring system using IoT," *Int. J*, vol. 8, 2020.

[8] J. Liang, Z. He, W. Du, X. Ruan, E. Guo, and N. Shen, "Tailoring the microstructure and mechanical properties of laser metal-deposited Hastelloy X superalloy sheets via post heat-treatment," *Materials Science and Engineering: A*, vol. 884, p. 145546, Sep. 2023, doi: 10.1016/J.MSEA.2023.145546.

[9] R. K. Madapudi, A. A. Rao, and G. M[illegible]d, "Change requests artifacts to assess impact on structural design of SDLC phases," *Int'l J. Computer Applications*, vol. 54, no. 18, pp. 21–26, 2012.

[10] S. V. Suryanarayana and G. N. Balaji, "Using Computer Vision to enhance Safety in a Post COVID World," *2022 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2022*, 2022, doi: 10.1109/ICDSAAI55433.2022.10028919.

[11] B. Seshakagari and H. Reddy, "Deep Learning-Based Detection of Hair and Scalp Diseases Using CNN and Image Processing," *Milestone Transactions on Medical Technometrics*, vol. 3, no. 1, pp. 145–155, Mar. 2025, doi: 10.5281/ZENODO.14965660.

[12] M. Sudhakara, M. J. Meena, K. R. Madhavi, P. Anjaiah, and L. N. Prakash, "'Fish classification using deep learning on small scale and low-quality images," *Int. J. Intell. Syst. Appl. Eng*, vol. 10, no. 1s, pp. 282–293, 2022.

[13] J. R. Dwaram and R. K. Madapuri, "Crop yield forecasting by long short-term memory network with Adam optimizer and Huber loss function in Andhra Pradesh, India," *Concurr Comput*, vol. 34, no. 27, p. e7310, Dec. 2022, doi: 10.1002/CPE.7310.

[14] K. Chen *et al.*, "A review of machine learning in additive manufacturing: design and process," *The International Journal of Advanced Manufacturing Technology 2024 135:3*, vol. 135, no. 3, pp. 1051–1087, Oct. 2024, doi: 10.1007/S00170-024-14543-2.

[15] Z. Xia, B. Wu, C. Y. Chan, T. Wu, M. Zhou, and L. B. Kong, “Deep-learning-based pyramid-transformer for localized porosity analysis of hot-press sintered ceramic paste,” *PLoS One*, vol. 19, no. 9, p. e0306385, Sep. 2024, doi: 10.1371/JOURNAL.PONE.0306385.

[16] D. Venkata Lakshmi, R. Shyama, S. Anila, S. Abbineni, S. A. Al, and A. Al-Hilali, “An Intelligent Framework for Smart Automated House Implementation via Integration of IOT and DL,” *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2024*, pp. 225–228, 2024, doi: 10.1109/ICACITE60783.2024.10616423.

[17] Ashwin Shenoy, M., and N. Thillaiarasu. "Enhancing temple surveillance through human activity recognition: A novel dataset and YOLOv4-ConvLSTM approach." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-16.

[18] C. Sanford *et al.*, “Understanding Transformer Reasoning Capabilities via Graph Algorithms,” *Adv Neural Inf Process Syst*, vol. 37, pp. 78320–78370, Dec. 2024.

[19] T. Plötz, “Applying Machine Learning for Sensor Data Analysis in Interactive Systems,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, Jul. 2021, doi: 10.1145/3459666.

[20] A. Cheloee Darabi, S. Rastgordani, M. Khoshbin, V. Guski, and S. Schmauder, “Hybrid Data-Driven Deep Learning Framework for Material Mechanical Properties Prediction with a Focus on Dual-Phase Steel Microstructures,” *Materials 2023, Vol. 16, Page 447*, vol. 16, no. 1, p. 447, Jan. 2023, doi: 10.3390/MA16010447.

[21] Puttaswamy, B. S., and N. Thillaiarasu. "Fine DenseNet based human personality recognition using english hand writing of non-native speakers." Biomedical Signal Processing and Control 99 (2025): 106910.

[22] L. Li, J. Chang, A. Vakanski, Y. Wang, T. Yao, and M. Xiao, “Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–15, May 2024, doi: 10.1038/s41598-024-61185-x.

[23] Z. Cao, R. Magar, Y. Wang, and A. Barati Farimani, “MOFormer: Self-Supervised Transformer Model for Metal-Organic Framework Property Prediction,” *J Am Chem Soc*, vol. 145, no. 5, pp. 2958–2967, Feb. 2023, doi: 10.1021/JACS.2C11420/ASSET/IMAGES/LARGE/JA2C11420_0005.JPEG.

[24] A. Jose, E. Devijver, N. Jakse, and R. Poloni, “Informative Training Data for Efficient Property Prediction in Metal-Organic Frameworks by Active Learning,” *J Am Chem Soc*, vol. 146, no. 9, pp. 6134–6144, Mar. 2024, doi: 10.1021/JACS.3C13687/SUPPL_FILE/JA3C13687_SI_001.PDF.

[25] Ravi Prasad, M., and N. Thillaiarasu. "Multi-channel EfficientNet B7 with attention mechanism using multimodal biometric-based authentication for ATM transaction." Multiagent and Grid Systems 20.2 (2024): 89-108.

[26] K. Logeswaran *et al.*, “Predicting the Hardness of Low Alloy Metal using Machine Learning Model,” *2024 6th International Conference on Computational Intelligence and Networks (CINE)*, pp. 1–5, Dec. 2024, doi: 10.1109/CINE63708.2024.10881672.

[27] A. Stoll and P. Benner, “Machine learning for material characterization with an application for predicting mechanical properties,” *GAMM-Mitteilungen*, vol. 44, no. 1, p. e202100003, Mar. 2021, doi: 10.1002/GAMM.202100003.

[28] Z. Li, Wang, F. Liu, T. Wang, J. G. Wang, Y. M. Liu, and Q. X. Huang, “Intelligent prediction model of mechanical properties of ultrathin niobium strips based on XGBoost ensemble learning algorithm,” *Comput Mater Sci*, vol. 231, p. 112579, Jan. 2024, doi: 10.1016/J.COMMATSCI.2023.112579.

[29] M. Justi, M. P. de Freitas, J. M. Silla, C. A. Nunes, and C. A. Silva, “Molecular structure features and fast identification of chemical properties of metal carboxylate complexes by FTIR and partial least square regression,” *J Mol Struct*, vol. 1237, p. 130405, Aug. 2021, doi: 10.1016/J.MOLSTRUC.2021.130405.

[30] “Steel Strength Data 2018.” Accessed: Mar. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/fuarresvij/steel-test-data