Predicting Mechanical Properties of Glass-Epoxy Composites Using Supervised Learning

Ronit Raj Baishya[1] and Scott W. Case[2]

[1]Virginia Tech, Blacksburg, USA

August 20, 2024

Abstract

The traditional testing methods for evaluating mechanical properties of composite laminates such as Interlaminar Shear Strength (ILSS) and In-Plane Shear Strength are known to be resource-intensive, time-consuming, and expensive. This often leads to setbacks and failures in the development process. In this study, the development of a predictive model was proposed to estimate these key mechanical properties based on specific test measurements and dimensions. The glass-epoxy family of composites was focused on as it is widely used in various industries. To explore the feasibility of this approach, supervised machine learning techniques were employed, which offer an efficient means to create predictive models based on test features. The selected tests for consideration include the Short Beam Strength Test (SBS Test) and Iosipescu or V–notch test for ILSS and in–plane shear strength respectively.

The performance of different machine learning algorithms such as decision tree, multiple linear regression, ridge regression, and artificial neural networks was evaluated to identify the most suitable model for the dataset. Given the limited availability of data, the study emphasizes the importance of achieving good performance even with small datasets. The findings from this research hold promise for streamlining the testing process and improving the efficiency of composite material development.

Keywords: Interlaminar Shear Strength, In-Plane Shear Strength, Composite Material, Glass Epoxy, Material Property Prediction, Supervised Learning, Decision Tree, Multiple Regression, Ridge Regression, Artificial Neural Network.

1 Introduction

Traditional composite laminate testing is resource-intensive, time-consuming, and expensive, often resulting in setbacks and failures. Developing a model to predict key measures of non-fiber controlled strength or toughness (e.g. Interlaminar Shear Strength (ILSS) and In-Plane Shear Strength) would have a significant impact. This would provide an estimate of Interlaminar Shear Strength and In-Plane Shear Strength based on the given measurements and dimensions for specific tests. To investigate and demonstrate feasibility, glass-epoxy family of composites was selected for the project, as one of the most commonly used forms of composite material. In the realm of predictive modeling, there are different techniques for predicting and the most emerging and efficient one is machine learning. Machine learning stands out as the most cutting-edge

and effective technique. Given the challenge of limited data availability from online sources, it was imperative to develop a model that could yield accurate predictions even with sparse datasets.

The main goal for this project was to compile experimental data from various published papers and open sources. It aimed to construct a prediction model tailored for small datasets, focusing on mechanical properties such as Interlaminar Shear Strength and In-Plane Shear. To achieve this, the effectiveness of different machine learning algorithms was explored, including decision trees, multiple linear regression, ridge regression, and artificial neural networks (ANN). The main objective was to capture the intricate relationships within the data and to evaluate and compare the performance of each model with limited data. By identifying the strengths and limitations of each algorithm, the aim was to select the most suitable model for accurate prediction, while also gaining insights into the challenges associated with the rejected algorithms. After selecting the model, the feature importance or influence will be analyzed to aid in the composite design process.

2 Overview of Tests and Machine Learning Algorithms

2.1 Short Beam Strength Test (SBS Test)

For the project, the Short Beam Strength Test (SBS) was taken for the ILSS measurement. Figure 1 illustrates the fixture for the Short-Beam-Strength (SBS) test, while Figure 2 provides an example of the dimensions utilized for the laminate within the project. The test standards for SBS test are followed according to ASTM D2344. The primary focus during data collection was to ensure adherence to ASTM standards within the selected sources. The Short Beam Strength test, previously known as the Short Beam Shear test, is a standard method for measuring matrix-dominated properties and is often assumed to represent interlaminar shear strength (ILSS). In the SBS test, a concentrated load is applied at the center of the specimen, inducing shear stress within the material. The SBS is then calculated using beam theory, assuming that shear failure occurs mid-way through the thickness.

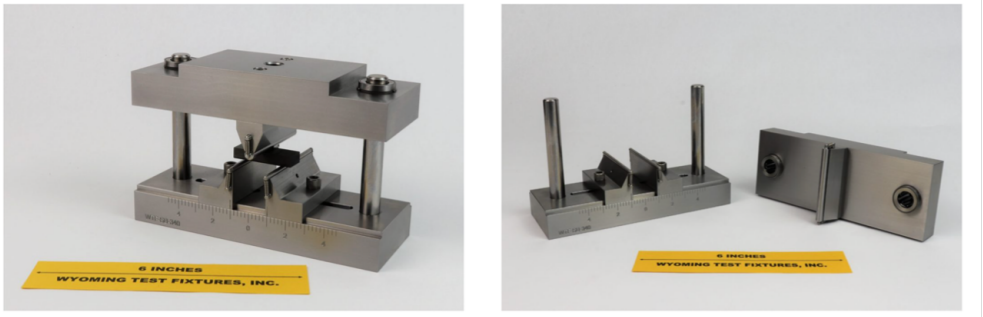

Fig. 1: Standard Short Beam Shear Test Fixture [3]

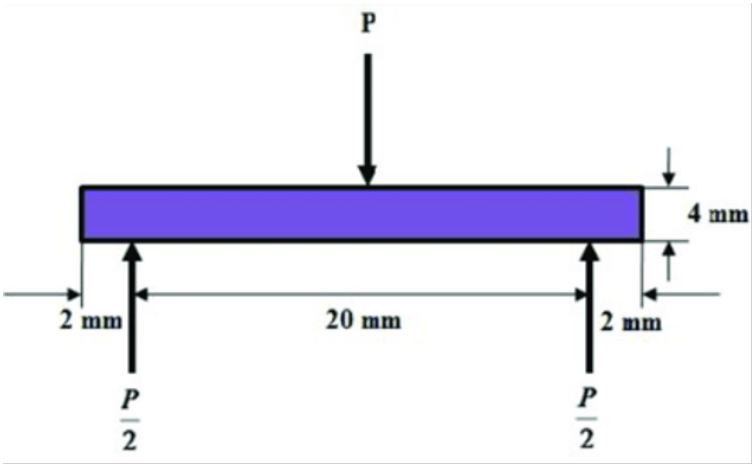

Fig. 2: Schematic Representation of SBS Test According to ASTM D2344 [24]

Interlaminar Shear Strength or ILSS is a material property used to measure the shear strength between adjacent layers, or laminas, in a composite material. ILSS is a crucial parameter in understanding how well the layers of composite material can resist sliding or shearing relative to each other, which is essential for predicting the material's overall strength and durability.

2.2 Iosipescu Shear Test (V–notch or ±45° tensile shear test)

In–plane shear strength refers to the ability of a material to resist deformation due to applied shear stresses within the plane of its structure. This property is particularly important in materials like composites, where multiple layers or plies are bonded together to form a laminate. For this project we selected the V–notch or Iosipescu shear test or the ±45° tensile shear test. The Iosipescu test is a mechanical test method used to determine the shear properties of composite materials. Figure 3 illustrates the setup of the V–notch or Iosipescu shear test, while Figure 4 provides an example of the dimensions utilized for the specimen within the project.

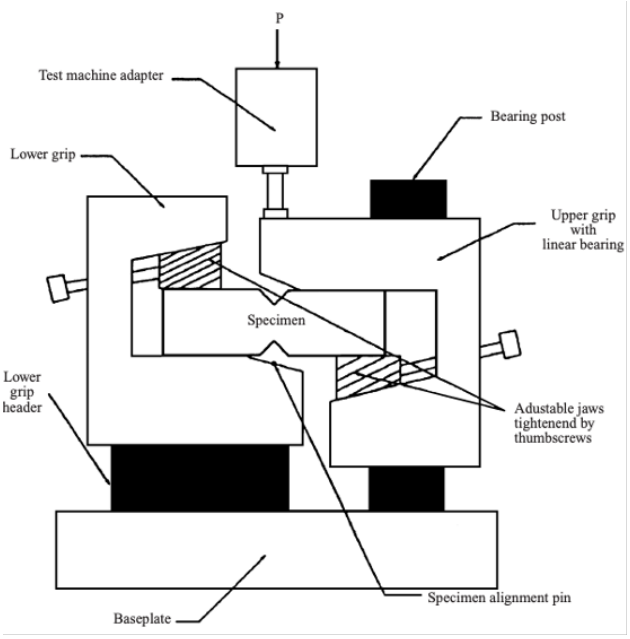

Fig. 3: Iosipescu Shear Test [26]

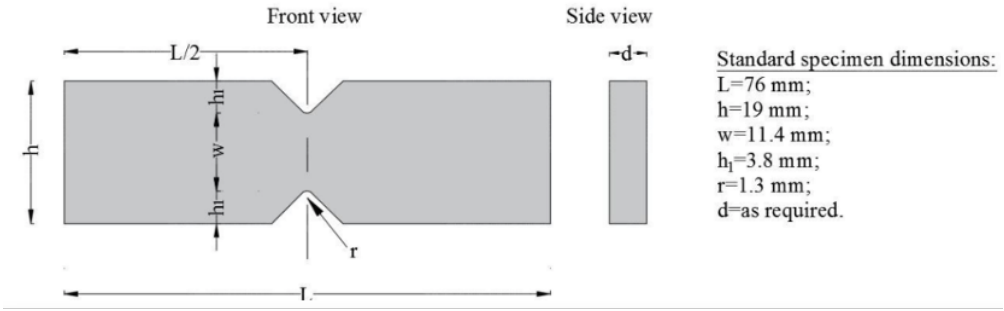

Fig. 4: Standard Geometry of Iosipescu Specimen [23]

The test standards for Iosipescu test are followed according to ASTM D5379. The specimen is placed between two grips, and a shear force is applied along the plane of the material, causing it to deform. The applied force causes the material to shear along the plane defined by the V-notch. The key principle behind the Iosipescu test is that the shear stress distribution near the notch tip can be analyzed to determine the shear strength of the material. Figure 5 shows the shear stress strain distribution of the specimen due to the load. By measuring the force applied and the deformation of the specimen, engineers can calculate the shear stress at the notch tip. This shear stress, combined with the dimensions of the specimen and the notch, allows for the calculation of the in-plane shear strength of the material.

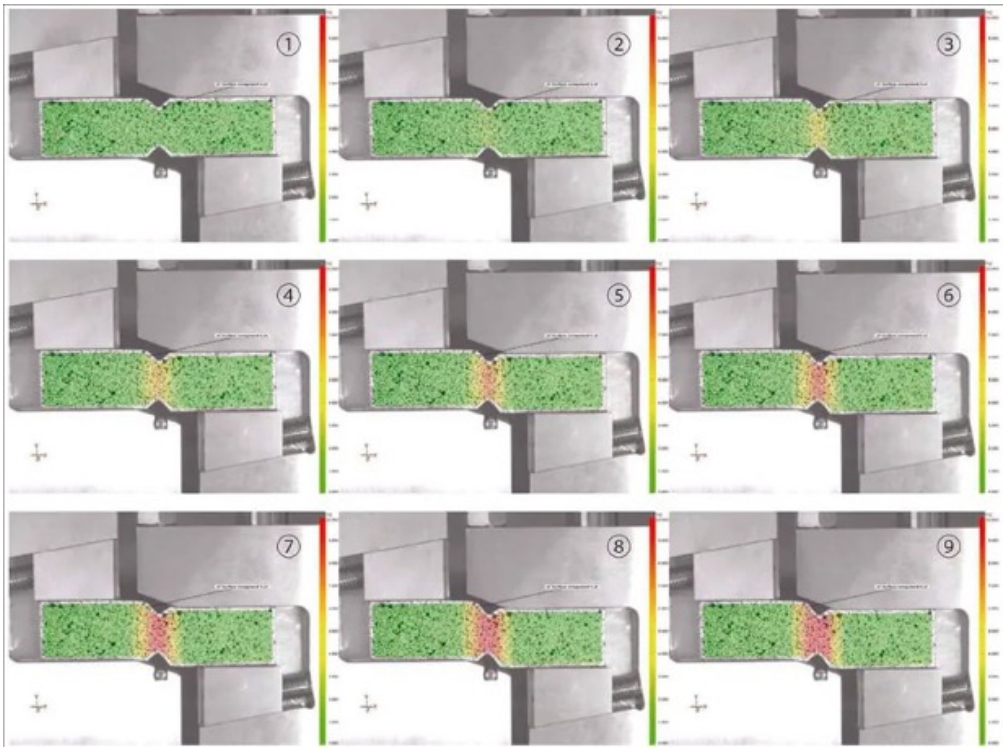

Fig. 5: Shear Strain Distribution in The Iosipescu Sample [21]

2.3 Decision Tree

In decision tree learning, the algorithm iteratively partitions the dataset into smaller subsets based on specific conditions or tests, represented as "leaves" (nodes) in the tree structure. Each leaf corresponds to a condition, and the possible outcomes are depicted as "branches" (edges) stemming from the leaves. This process of splitting continues until either no further improvement in prediction can be achieved or a predefined stopping criterion is met, such as reaching the maximum depth of the tree. This description outlines the fundamental principle of decision tree learning, which is widely recognized in the field of machine learning. Figure 6 shows a visual representation of a decision tree.

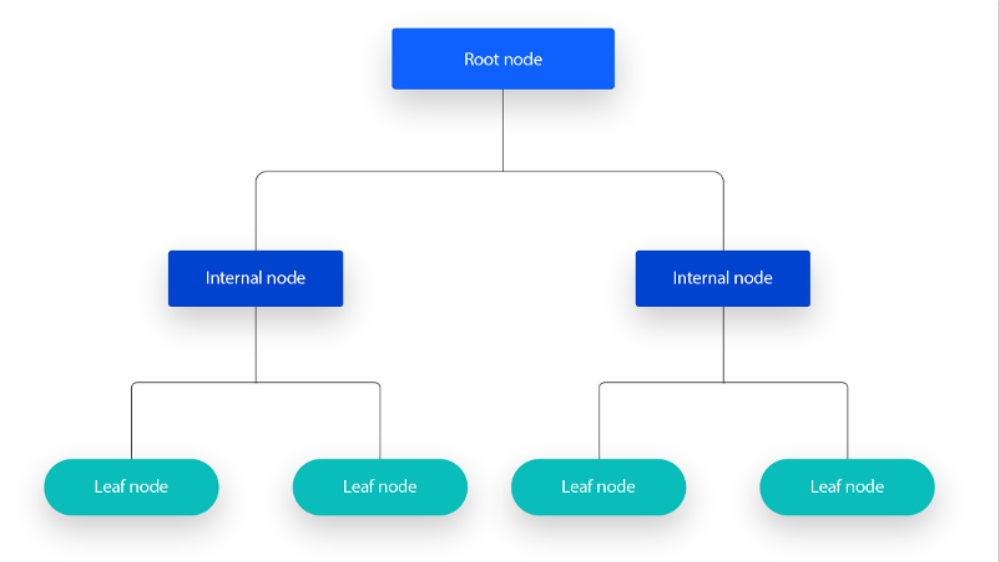

Fig. 6: Decision Tree Diagram [27]

In regression trees tailored for target variables with real numbers, conventional classification methods such as ID3 and C4.5 are unsuitable. Instead, these regression trees adopt a strategy centered around minimizing the standard deviation of the target data. Initially, the standard deviation of the target variable is computed across the dataset, serving as a measure of its spread. Subsequently, each feature column undergoes evaluation to ascertain its effectiveness in partitioning the data into groups that exhibit similarity in their target values. For every group formed through splitting based on a specific feature, the standard deviation of the target values within that group is determined. The weighted standard deviation is then derived by amalgamating the standard deviations of all groups, considering their respective proportions relative to the total sample size. The reduction in standard deviation is computed for each feature, indicating its contribution to diminishing the overall variability in the target values upon splitting.

$$\sigma_{global} = \sqrt{\sum_{n=1}^{n} \frac{(x_n - average)^2}{n}}. \tag{1}$$

Where, σ_{global} = standard deviation of the whole dataset, x_1 = value of target column and n = total number of values

$$\sigma_{weighted} = \sqrt{\frac{\text{Fraction of Instance}}{\text{Total Instance}} \times \sigma_{\text{group}}} \tag{2}$$

Where, σ_{weighted} = weighted standard deviation, σ_{group} = standard deviation of each tree

$$\sigma_{reduction} = \sigma_{global} - \sigma_{weighted} \tag{3}$$

This iterative process repeats for all features, ultimately identifying the feature that yields the highest reduction in standard deviation as the root node of the tree. Following this selection, the process recurs on

each subtree, with features chosen based on their potential to further reduce the standard deviation within each subset of the data. Through this methodical approach, regression trees effectively segment the data into increasingly homogeneous groups concerning the target variable, facilitating the creation of a predictive model capable of accurately estimating continuous target values.

2.4 Linear Regression

Regression is a statistical method to determine the relationship between one dependent variable Y, which is a continuous variable, and a series of other independent variables $x_1, x_2, \ldots, x_n$. Figure 7 shows a visual representation of a typical linear regression.

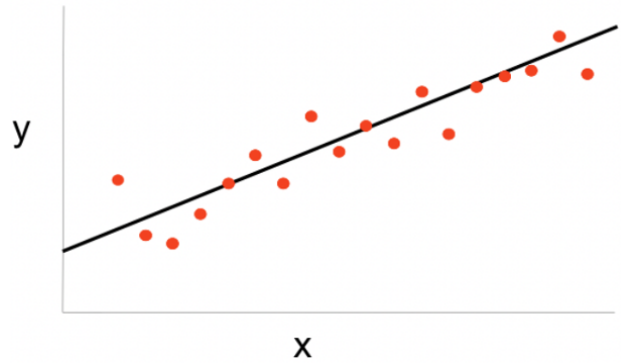

Fig. 7: Typical Linear Regression Plot [15]

Linear Regression, a cornerstone of Supervised Learning, leverages labeled historical data to predict the value of an output or dependent variable by analyzing the relationships with predictor or independent variables. As its name implies, Linear Regression assumes a linear relationship between the dependent and independent variables. By utilizing this method, the algorithm aims to uncover the underlying linear patterns within the data, enabling it to make accurate predictions based on observed historical patterns. This technique is widely used in various fields, from finance to healthcare, due to its simplicity and effectiveness in modeling and predicting continuous variables. There are two types of regression. Simple linear regression is denoted by the Equation of the line:

$$y = c + mX \tag{4}$$

Where, y : dependent variable, X: predictor variable, m: slope of the line defining relationship between X and y, also called coefficient of X, c: intercept.

Second is multiple linear regression where more than one predictor variables are used to predict the values of dependent variable. Equation of the line

$$y = c + m_1x_1 + m_2x_2 + m_3x_3.... + m_ix_i \tag{5}$$

where $x_1, x_2, \ldots, x_i$ are the n predictor variables and $m_1, m_2, \ldots, m_i$ are the respective coefficients.

In linear regression, the primary objective is to determine the optimal values for the coefficients c and m in the equation of a line that best fits the given set of predictor variables X and the corresponding target

variable y. To accomplish this, it is crucial to thoroughly grasp the graph depicting the relationship between the predictor and target variables. By close analysis of the graph, the aim is to discern the underlying patterns and trends within the data. Ultimately, through a comprehensive understanding of the graph's characteristics, the coefficients c and m can be iteratively adjusted (in terms of machine learning) to ensure that the resulting line minimizes the overall distance between the observed data points and the predicted values. This iterative process of fine-tuning the coefficients is essential for achieving the most accurate and optimal linear model. There are many methods of finding the fits but the most commonly used is squared error. The advantage of squared error over others is that it penalizes large errors more heavily, ensures that all errors are positive and facilitates mathematical analysis by yielding a smooth, differentiable function. The general equation for squared error is:

$$\text{mean squared error} = \frac{1}{n}\sum_{n=1}^{n}(\text{observed } y - \text{predicted } y)^2 \tag{6}$$

where n is number of terms. The visual representation of the error can be seen in Figure 8.

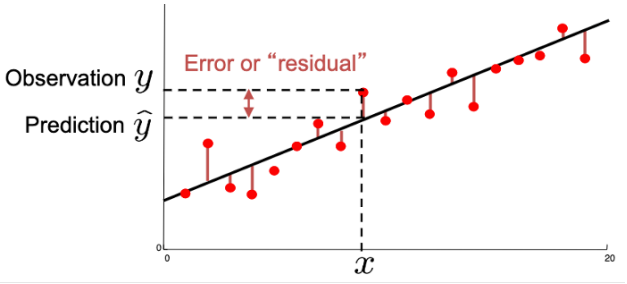

Fig. 8: Visualization of Error in A Linear Regression Plot [16]

After finding all the fits, the best fit is where the error is the minimum. The error equation can be also written as:

$$\text{Minimum Error} = \arg\min\left\{\frac{1}{n}\sum_{i=1}^{n}(\text{observed } y_i - \text{predicted } y_i)^2\right\} \tag{7}$$

2.5 Ridge Regression

Ridge regression stands out as an excellent starting point for regression analysis due to its stability across various datasets. Similar to lasso regression, ridge regression represents a regularized version of linear regression. In the linear regression loss function, it aims to minimize the discrepancy between observed and predicted values. However, ridge regression incorporates an additional penalty term, typically the square of the L2 norm of the coefficient vector, to address potential multicollinearity and overfitting issues. This penalty term controls the complexity of the model and helps prevent excessive reliance on any single feature, thereby enhancing generalization performance. The loss function for ridge regression just adds a regularization term to the end:

$$\text{Loss Function} = \frac{1}{n}\sum_{n=1}^{n}(\text{observed } y - \text{predicted } y)^2 + \lambda\phi^2 \tag{8}$$

where n is number of terms and

$$\phi = (X^T X + \lambda I)^{-1} X^T y \tag{9}$$

where, X: predictor variable and y : dependent variable

Ridge regression does have an analytical solution for its optimized parameters. The discovery of the optimal parameters in linear regression, as well as in ridge regression, follows a methodical process involving calculus. By taking the derivative of the loss function with respect to the regression coefficients (θ in linear regression), and equating it to zero, we can pinpoint the values that minimize the loss. Similarly, in ridge regression, this process is extended by including a regularization term controlled by the regularization parameter (α), and the identity matrix (I) to the equation. The optimized parameters for ridge regression, reflecting both the data fit and regularization, are derived through this meticulous calculus-based approach, ensuring a balanced trade-off between model complexity and performance.

2.6 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) have emerged as a cutting-edge technique for tasks like prediction, classification, translation, and more. Mimicking the structure of the human brain, an ANN comprises interconnected neurons that process information. Its architecture typically includes three layers: the input layer, where initial data is fed into the network; the hidden layer, which serves as an intermediary between the input and output layers, processing and transforming the information; and the output layer, responsible for delivering the final output or prediction. This layered structure allows ANNs to effectively learn complex patterns and relationships within data, making them a versatile and powerful tool in various fields of artificial intelligence and machine learning.

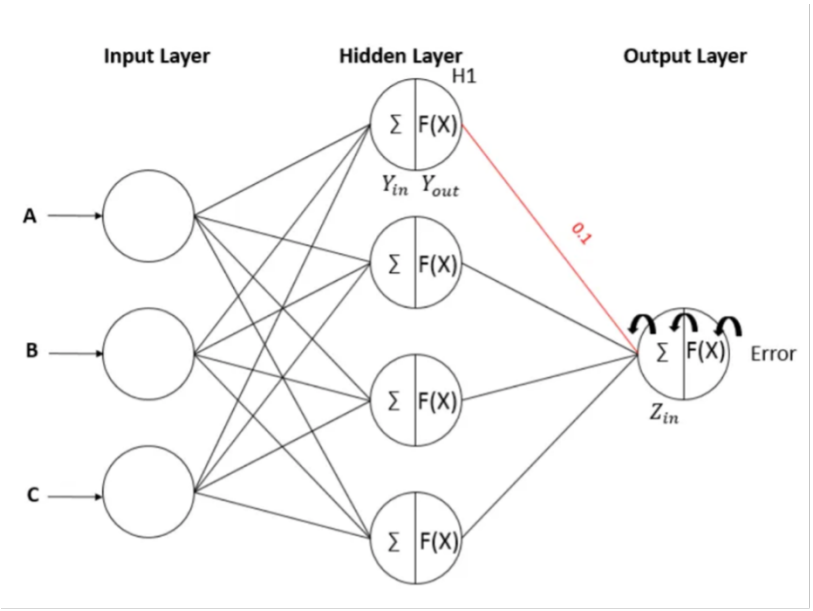

Fig. 9: Basic Structure of An Artificial Neural Network [19]

A fundamental building block in neural networks, the perceptron, is a simple artificial neuron featuring an input layer and an output layer. Within perceptrons, there exists a critical component known as the activation function, responsible for processing the weighted sum of input signals and producing the neuron's output signal. This output signals then serves as input for subsequent layers in the network. Activation functions play a pivotal role in introducing non-linearity to the neural network, enabling it to learn and represent complex relationships in the data. Among the array of activation functions available, two commonly used types include the step function, which produces a binary output based on a threshold, and the sigmoid function, which generates a smooth, continuous output between 0 and 1. These activation functions imbue neural networks with the capability to model intricate patterns and make accurate predictions across various domains.

Activation Function Name	Function	Remark
Sigmoid	$\frac{1}{1+e^{-x}}$	Commonly used in binary classification
Tanh	$\tanh(x)$	Scales output to [-1, 1]
ReLU	$\max(0, x)$	Popular in deep learning

Table 1: A table of activation functions

The process of training an Artificial Neural Network (ANN) comprises two key phases: forward propagation and backpropagation. In the first phase, forward propagation, arbitrary weights are initially assigned to the connections between neurons in the input layer. Each perceptron within the network computes the weighted sum of its inputs and applies an activation function to produce an output. This output is then multiplied by the corresponding weight and serves as the input for neurons in the subsequent hidden layer. This process iterates through the hidden layers until reaching the output layer, where the final prediction is made. The error is then calculated by comparing the predicted output to the actual target value for each data point.

The second phase, backpropagation, focuses on minimizing this error by adjusting the weights throughout the network. Utilizing the chain rule from calculus, the algorithm computes the gradient of the error with respect to each weight. By iteratively updating the weights in the direction that minimizes the error, typically through techniques like gradient descent, the network learns to make more accurate predictions. Gradient descent involves systematically adjusting the weights based on the calculated gradients, aiming to converge towards the local minimum of the error function.

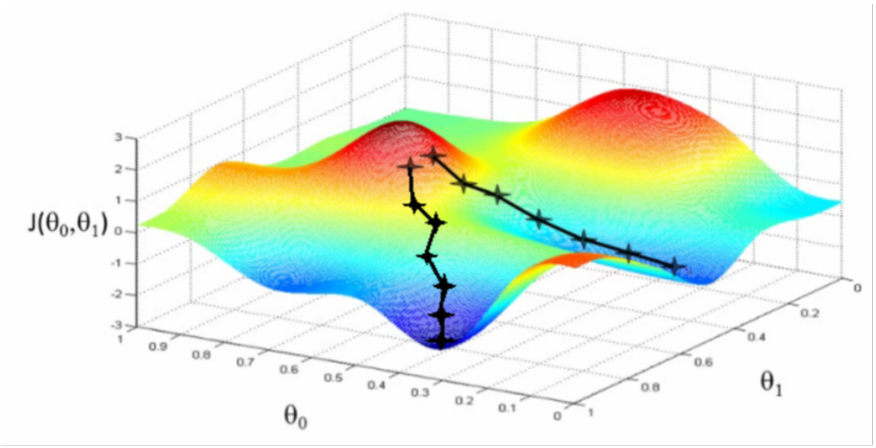

Fig. 10: Graphical Representation of Gradient Decent [18]

This iterative process of forward propagation followed by back propagation enables the neural network to learn from the data and continually improve its performance by adjusting its internal parameters.

2.7 R^2 Squared

The performance of models is often assessed using metrics such as the R2 score and adjusted R2 score. These scores gauge the effectiveness of regression models in explaining the variance observed in the dependent variable based on the independent variables. Specifically, the coefficient of determination, or R2, quantifies how well the regression line fits the given data. With values ranging from 0 to 1, an R2 score closer to 1 indicates a better fit, signifying that a substantial portion of the variance in the dependent variable is captured by the independent variables. The formula to calculate is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

The variables used in the R^2 formula are defined as follows:

y_i : Observed value of the dependent variable for the i-th data point
$\hat{y}_i$: Predicted value of the dependent variable for the i-th data point
$\bar{y}$: Mean of the observed values of the dependent variable

2.8 Adjusted R^2 Squared

As the features and data of the model expands by incorporating additional variables, the R-squared value tends to increase. However, this does not always guarantee an improvement in model performance, especially if the added variables lack significance. In such cases, the model may become overly complex and susceptible to overfitting, where it captures noise in the data rather than meaningful relationships. To address this concern, a metric known as Adjusted R-squared is utilized. Unlike R-squared, Adjusted R-squared penalizes the inclusion of insignificant variables, thereby providing a more accurate reflection of the model's effectiveness.

By accounting for the number of predictors and their significance, Adjusted R-squared offers a balanced assessment of model fit, helping to mitigate the risk of overfitting and ensuring that only relevant variables contribute meaningfully to the model's predictive power.The formula to calculate is:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{11}$$

The terms in the adjusted R^2 formula are defined as follows:

$\bar{R}^2$: Adjusted R^2, which accounts for the number of predictors

R^2 : Regular R^2, the coefficient of determination

n : Number of observations in the dataset

p : Number of predictors (independent variables) in the model

2.9 Cross Validation Score

Mean cross-validation score is a metric used to evaluate the performance of a machine learning model by estimating its generalization ability on unseen data. It is calculated by performing k-fold cross-validation on the dataset and averaging the performance scores (e.g., accuracy, F1-score, or mean squared error) obtained from each fold. The mean cross-validation score provides an estimate of the model's performance on unseen data, as each fold serves as a validation set for the model trained on the remaining folds. A higher mean cross-validation score generally indicates better generalization performance of the model.The formula to calculate is:

$$\text{Mean CV Score} = \frac{1}{K} \sum_{k=1}^{K} \text{CV}_k \tag{12}$$

The terms in the mean cross-validation score formula are defined as follows:

Mean CV Score : Average of the individual cross-validation scores

K : Number of folds in the cross-validation

CV_k : Cross-validation score for the k-th fold

2.10 Feature Importance

Random Forest Regressor algorithm is a popular machine learning algorithm used for both classification and regression tasks. In the context of regression, Random Forest Regressor is a powerful ensemble learning method that combines the predictions of multiple decision trees to improve the accuracy and robustness of the model. One of the key advantages of Random Forest is its ability to provide insights into feature importance, which can be crucial for understanding the underlying relationships in the data and making informed decisions.

Forest consists of a collection of decision trees, each trained on a random subset of the training data and a random subset of features. This randomness helps to reduce overfitting and improve generalization. After training the Random Forest, feature importance is calculated based on the impurity decrease caused by each feature. Features that result in the largest impurity decrease across all trees are considered the most important. The importance scores for each feature are typically normalized so that they sum up to 1 or 100,

allowing for easier interpretation and comparison. Finally, the feature importance scores can be visualized using bar plots or ranked in descending order to identify the most influential features in the model.

Random Forest is particularly effective at capturing complex nonlinear relationships in the data, making it well-suited for modeling nonlinear datasets. Unlike linear regression models, Random Forest can handle highly nonlinear relationships between features and the target variable without the need for feature engineering or transformation. This flexibility allows Random Forest to achieve higher accuracy on nonlinear datasets and outperform linear models in many cases. The formula used to calculate feature importance in Random Forest varies depending on the impurity measure used (e.g., Gini impurity or entropy) and the specific implementation of the algorithm. However, a common approach is to compute the weighted average of impurity decrease across all decision trees in the forest.

$$\text{Feature Importance} = \frac{\sum_{t=1}^{T} \text{Impurity Decrease}_t}{T} \tag{13}$$

T : the total number of trees in the Random Forest

$\text{Impurity Decrease}_t$: decrease in impurity caused by splitting on the feature in the t-th tree

In general machine learning concepts the two most commonly used impurities are Gini Impurity and Entropy. Impurity in decision trees refers to a measure of how mixed the classes are within a node. It indicates the degree of disorder or uncertainty. Lower impurity means the node is more homogenous. Common measures of impurity include Gini impurity and entropy. The Gini impurity for a node is calculated as 14 and The entropy for a node is calculated as 16 respectively.

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2 \tag{14}$$

where p_i is the probability of class i at a particular node.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{15}$$

where p_i is the probability of class i at a particular node.

3 Procedure

Initially, the focus lies on Data Collection, where efforts are concentrated on acquiring relevant data and delineating its features. This comprehensive database encompasses sixty-one data points for interlaminar shear strength and forty data points for in-plane shear. Subsequently, the Preprocessing stage ensues, marking the transformation of the raw dataset into a refined form conducive to model training. With the preprocessed dataset in hand, the Model Training phase commences. Employing various machine learning algorithms including decision trees, multiple linear regression, ridge regression, and Artificial Neural Networks. Lastly, Model Validation, here, the accuracy and effectiveness in predicting the target mechanical properties are assessed, and a reverse analysis is conducted to scrutinize the models' capabilities and refine them as necessary. Figure 11 visually illustrates the entire process, adopted in the project methodology.

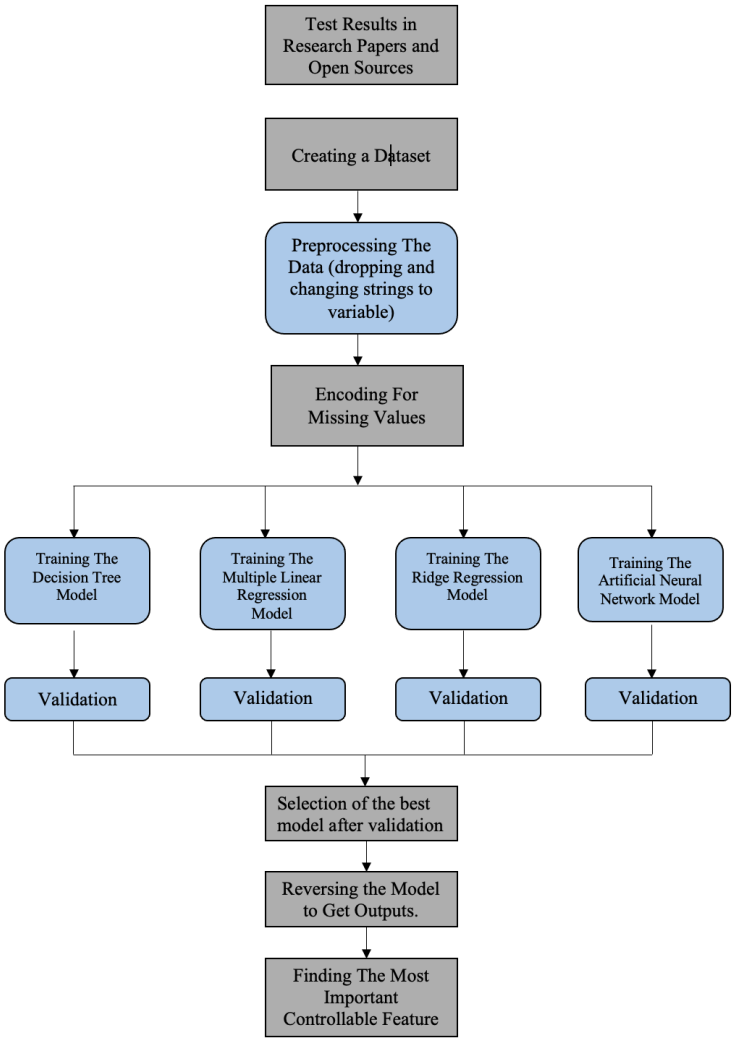

Fig. 11: General Framework of The Whole Project

3.1 Interlaminar Shear Strength

The dataset for interlaminar shear strength (ILSS) was derived from Short-Beam-Strength (SBS) tests, ensuring adherence to the ASTM D2344 standard. It encompasses detailed information on the composite's constituents, including E-glass fibers and epoxy matrix, as well as essential structural characteristics such as layer counts. The features considered for the models and their respective range of data are shown in table 2:

Column Name	Range
Thickness (mm)	1 mm–10.17 mm
Width (mm)	4.405 mm–50 mm
Length (mm)	15 mm–80.4 mm
Volume Fraction %	20 %–73 %
Load (N)	528 N–2280 N
Angle (Degree)	0°–90°
ILSS (MPa)	16.13 MPa N–90.9 MPa

Table 2: Limits of each feature of interlaminar shear strength in the model

3.2 In–Plane Shear Strength

The dataset for in-plane shear strength was meticulously collected using V–notch, also known as the Iosipescu shear test or the ±45° tensile shear test, following the rigorous guidelines outlined in the ASTM D5379 standard. This comprehensive dataset incorporates vital parameters such as fiber type (E–glass) and matrix composition (Epoxy), alongside key test details including the type of test conducted. The features considered for the models and their respective range of data are shown in table 3

Column Name	Range
Thickness (mm)	2.5 mm–6 mm
Width (mm)	20 mm–56 mm
Length (mm)	76 mm–82.7 mm
Volume Fraction %	34 %–50 %
Axial Strain Rate	0.00117–84.35
Shear Strain Rate	0.00243–130.95
Angle (Degree)	±45°
In Plane Shear (MPa)	14.3 MPa–59.24 MPa

Table 3: Limits of each feature of in–plane shear strength in the model

3.3 Preparation of The Dataset

In dataset preparation, it is crucial to streamline the features to ensure optimal model training. If the dataset contains consistent values across certain columns, such as fiber and matrix types, these columns are dropped as they do not contribute to the variability in the data. However, it's essential to retain these columns if the dataset comprises different fiber and matrix compositions. Once irrelevant columns are removed, the remaining features are primed for training the model. To handle missing data, empty cells were filled with zero values, and these zeros were encoded to ensure they didn't impact the model training process. This meticulous preprocessing step helps to refine the dataset, ensuring that only meaningful and diverse features are included in the model training process. To facilitate machine learning algorithms that are unable to interpret string data, it's necessary to convert categorical variables into numerical format through encoding.

3.4 Training of The Model

During the model training process, various features were considered depending on the specific mechanical property being analyzed. For Interlaminar Shear Strength, features such as number of layers, thickness, width, length, volume fraction, load, and angle were considered. In the case of In-Plane Shear Strength, features included thickness, width, length, volume fraction, axial strain rate, shear strain rate, and angle. The dataset was then split into training and validation sets, with 30% allocated to training and 70% to validation. Furthermore, the training set was randomized to prevent any bias in the model learning process. Various machine learning techniques were employed to train the models, including decision tree, multiple linear regression, ridge regression, and Artificial Neural Networks (ANN). For the ANN model, a single hidden layer was utilized. By leveraging these diverse approaches, predictions were generated to analyze and understand the relationships between the selected features and the respective mechanical properties.

3.5 Model Validation

The process of assessing whether the numerical outcomes that quantify hypothesized relationships between variables accurately represent the data is referred to as validation. Typically, after training a model, an evaluation of residuals is conducted to estimate errors. This involves numerically comparing the predicted responses with the original ones, known as the training error. To validate the models for this project, various metrics are employed to gauge its performance. Mean squared error, confidence interval, R-squared, and adjusted R-squared are utilized for this purpose. Among these metrics, confidence interval and adjusted R-squared holds the highest priority. Subsequently, all models undergo cross-validation to identify the best-performing model among them. This comprehensive validation process ensures that the selected model accurately captures the relationships within the data and generalizes well to unseen observations.

After numerical validation, a reverse model is created to obtain a target value, which is then compared with the training values. Following this comparison, feature importance is determined using a Random Forest Regressor (RFR) algorithm. The RFR algorithm is chosen for its effectiveness in interpreting nonlinear datasets.

4 Results and Discussion

The performance metrics are R-squared score, adjusted R-squared, Mean Squared Error (MSE), training and validation plot and cross-validation score for decision tree, multiple regression and ridge regression. The cross-validation score for ANN was ignored because of the assumption of poor performance in low number of data points.

After executing the code various parameters are generated to validate the performance of the models. These parameters are assessed across four different models: decision tree, multiple linear regression, ridge regression, and Artificial Neural Network (ANN). The performance metrics include R-squared score, adjusted R-squared, Mean Squared Error (MSE), cross-validation score, and confidence interval. The resulting metrics serve as indicators of the models' predictive accuracy and generalization ability.

4.1 Performance Metrics of The Model

The process of assessing whether the numerical outcomes that quantify hypothesized relationships between variables accurately represent the data is referred to as validation. Typically, after training a model, an

evaluation of residuals is conducted to estimate errors. This involves numerically comparing the predicted responses with the original ones, known as the training error. To validate the models for this project, various metrics are employed to gauge its performance. Mean squared error, confidence interval, R-squared, and adjusted R-squared are utilized for this purpose. Among these metrics, confidence interval and adjusted R-squared holds the highest priority. Subsequently, all models undergo cross-validation to identify the best-performing model among them. This comprehensive validation process ensures that the selected model accurately captures the relationships within the data and generalizes well to unseen observations.

Interlaminar Shear Strength			
Model	R^2	Adjusted R^2	Mean Squared Error
Decision Tree	0.99	0.99	6.11
Multiple Linear Regression	0.83	0.71	85.70
Ridge Regression	0.82	0.70	87.08
ANN	0.86	0.76	32.77

Table 4: Comparison of different models based on R^2, Adjusted R^2, and Mean Squared Error For Interlaminar Shear Strengh

In–Plane Shear Strength			
Model	R^2	Adjusted R^2	Mean Squared Error
Decision Tree	0.99	0.98	1.45
Multiple Linear Regression	0.92	0.77	9.59
Ridge Regression	0.91	0.76	9.71
ANN	0.13	0.02	574.26

Table 5: Comparison of different models based on R^2, Adjusted R^2, and Mean Squared Error for In–Plane Shear Strength

4.2 Model Selection

Upon thorough examination of the primary performance metrics R^2 Score, adjusted R^2 Score, and Mean Squared Error of each model for interlaminar shear strength and in-plane shear, it becomes evident that the decision tree model stands out as the most effective predictor in this context. The plots for the training set and validation set for Interlaminar Shear Strength and In–Plane Shear Strength are shown in Figure 12 and Figure 13 respectively.

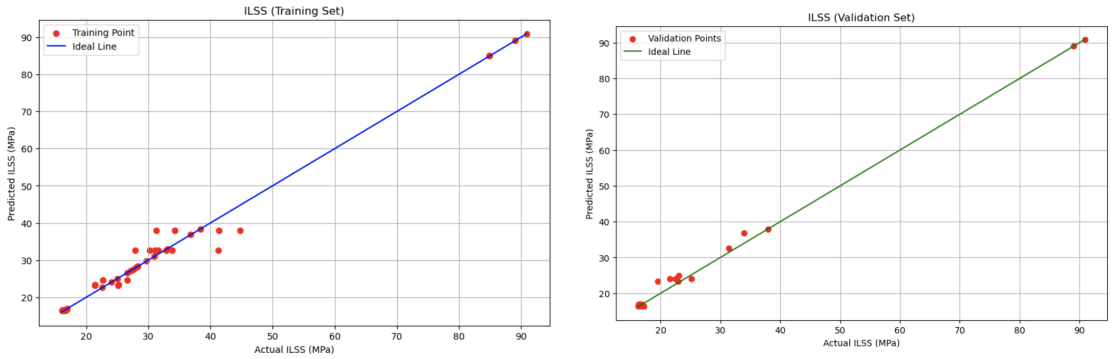

Fig. 12: Plot of Training Set and Validation Set For Interlaminar Shear Strength

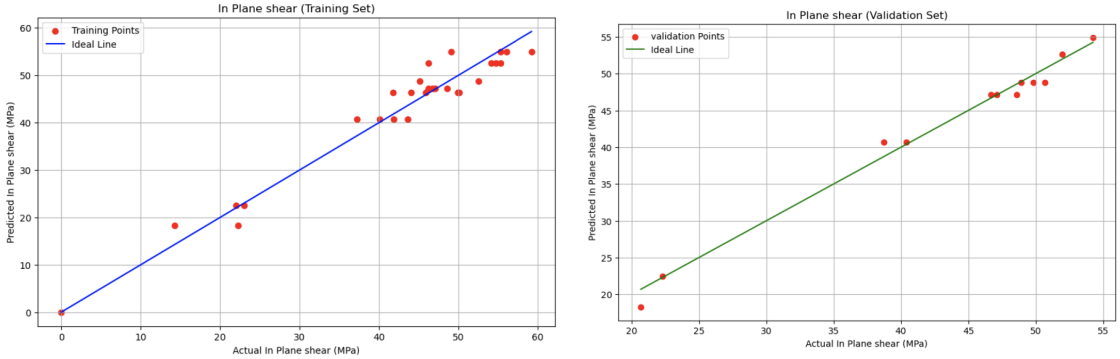

Fig. 13: Plot of Training Set and Validation Set For In–Plane Shear Strength

4.3 Model Validation

A more comprehensive validation was done of the decision tree model. Firstly, assessing the 95% confidence interval provides insight into the uncertainty surrounding the model's predictions. Secondly, comparing the mean cross-validation score offers an aggregate measure of the model's performance across multiple validation sets. Finally, reversing the model to predict random target values and comparing them with the corresponding values in the training dataset allows for a thorough examination of the model's predictive accuracy. This multifaceted approach ensures a robust evaluation of the decision tree model's reliability and effectiveness in making accurate predictions.

The 95% confidence interval of the decision tree model for interlaminar shear strength and in–plane shear is shown in Figure 14 and Figure 15 respectively.

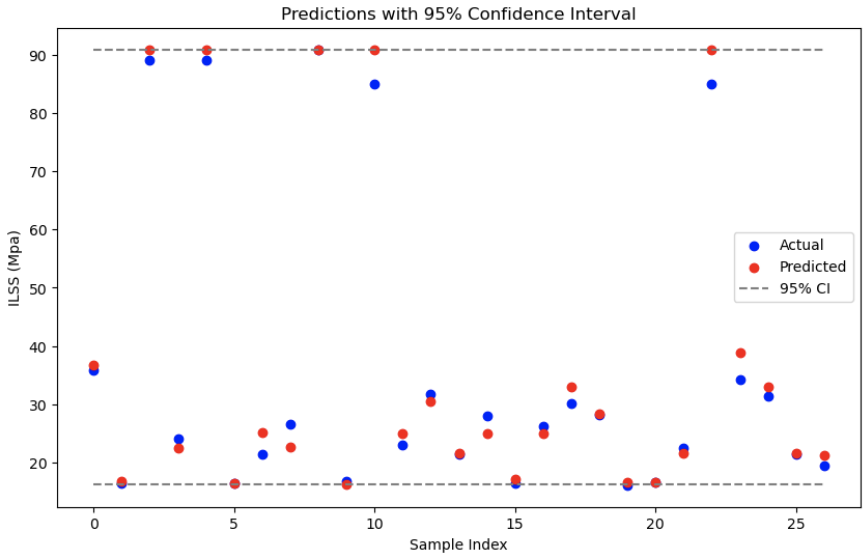

Fig. 14: Plot of Confidence Interval of Decision Tree Model For Interlaminar Shear Strength

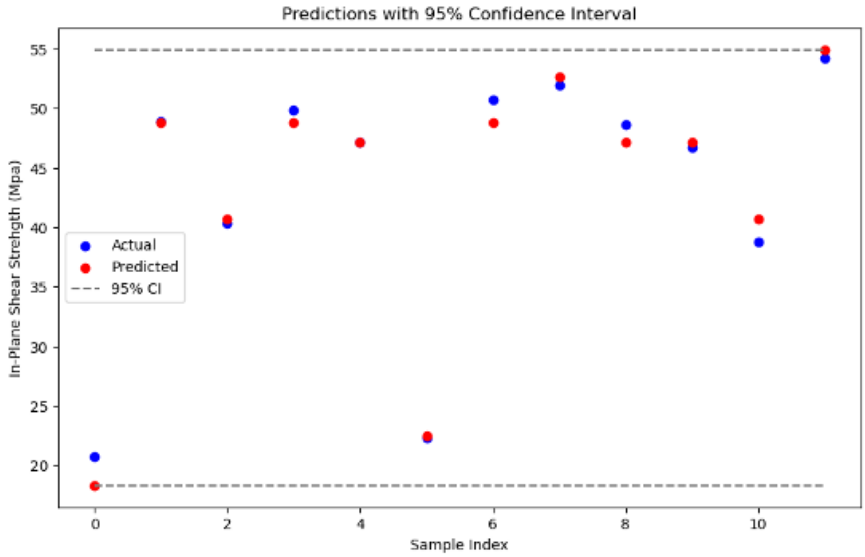

Fig. 15: Plot of Confidence Interval of Decision Tree Model For In–Plane Shear Strength

The comparison of mean cross validation score of the decision tree model, multiple linear regression model and ridge regression model for interlaminar shear strength and in–plane shear is shown in Figure 16 and Figure 17 respectively. ANN model was not taken into account because it had a poor performance compared

to other machine learning models, which would affect the comparison visually.

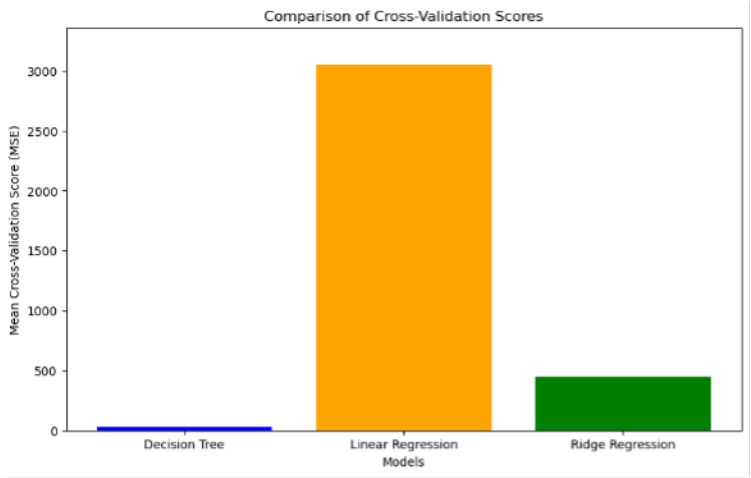

Fig. 16: Comparison Plot of Mean Cross-Validation Score for ILSS

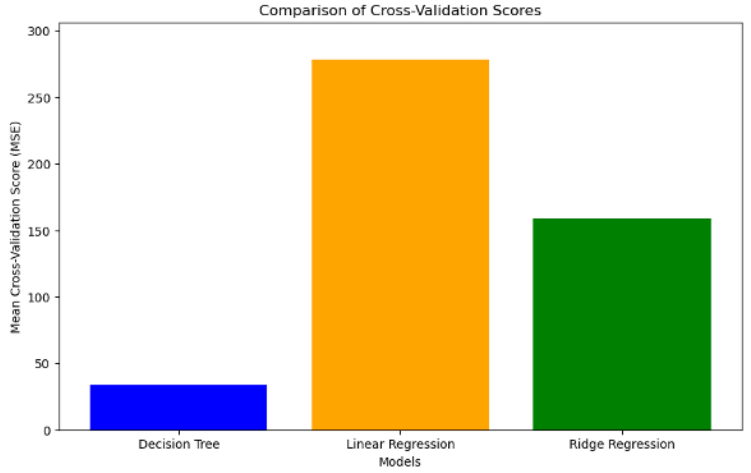

Fig. 17: Comparison Plot of Mean Cross-Validation Score for In–Plane Shear Strength

4.4 Reverse Model Prediction

Table 6 presents a comparative analysis of the predicted values generated by the decision tree model, the corresponding values from the dataset and the accuracy percentage for interlaminar shear strength and in–plane shear.

Upon analysis of the results, it is evident that the decision tree emerges as the most effective model for handling small datasets, outperforming both multiple linear regression and ridge regression. Interestingly,

Mechanical Property	Predicted Value	Corresponding Value	Accuracy
Interlaminar Shear Strength (MPa)	16.5	16.13	97.76%
In–Plane Shear Strenght (MPa)	54.92	54.24	98.74%

Table 6: Comparison of predicted value vs corresponding value

despite its complexity, the decision tree model demonstrates superior performance in capturing the nonlinear relationships present within the dataset. Conversely, the Artificial Neural Network (ANN) exhibits the poorest performance among the models considered, highlighting its limitations in handling smaller datasets effectively. This observation aligns with the initial hypothesis formulated during the model creation process, reinforcing the notion that decision tree models are well-suited for scenarios involving limited data points and nonlinear relationships.

4.5 Feature Importance

For feature importance, the entire dataset was analyzed. However, particular attention was given to controllable features such as length or span, width, thickness, and volume fraction. The focus was to determine the influence of these features and identify which changes would have the greatest effect.

The Random Forest Regressor (RFR) algorithm was used. The RFR algorithm is preferred over other algorithms such as Principal Component Analysis (PCA) or Linear Model Feature Importance due to its superior handling of nonlinear data. Given that decision trees demonstrated the best performance in our initial evaluations, leveraging the RFR algorithm is the most effective approach for feature importance in this context. This ensures a robust assessment of the predictive power of each feature within a nonlinear dataset. Figure 18 and Figure 19 show the feature importance for interlaminar shear strength and In–plane shear strength respectively.

	Feature	Importance
3	L to h ratio	0.521238
0	Thickness (mm)	0.311557
1	Width (mm)	0.091832
4	Vf (%)	0.048848
2	Length (mm) or Span	0.021371
5	Load (N)	0.005046
6	Angle	0.000109

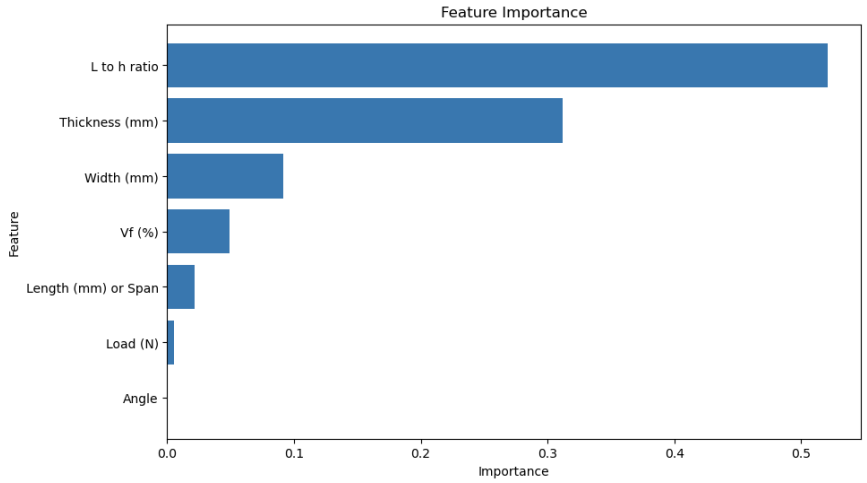

Fig. 18: Feature Importance of ILSS

	Feature	Importance
1	Thickness (mm)	0.830520
5	Axial Strain Rate	0.074286
4	Vf (%)	0.065832
6	Shear Strain Rate	0.025874
2	Width (mm)	0.003488
0	No. of Samples	0.000000
3	Length (mm) or Span	0.000000

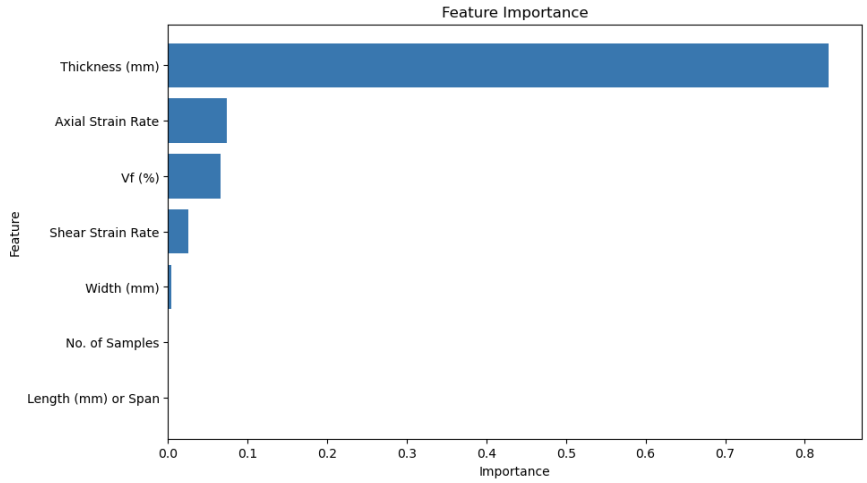

Fig. 19: Feature Importance of In–Plane Shear Strength

Conclusions

Upon thorough examination of the results, it becomes increasingly evident that the decision tree model emerges as the clear frontrunner for datasets characterized by a smaller sample size. Its ability to effectively capture complex, nonlinear relationships within the data sets it apart from both multiple linear regression and ridge regression models. Despite its inherent simplicity, the decision tree model showcases remarkable accuracy in navigating the intricacies of the small dataset. Conversely, the Artificial Neural Network (ANN) lags behind its counterparts, displaying inferior performance metrics, particularly in the context of smaller datasets. This observation further bolsters the initial hypothesis posited during the model's development, underscoring the decision tree's efficacy in scenarios where data points are limited, and relationships between variables are intricate and nonlinear which are also one of the disadvantages of Artificial Neural Network.

Based on the findings of this analysis, it is recommended to leverage decision tree models when dealing with datasets characterized by a limited number of observations and nonlinear relationships between variables. The decision tree's ability to effectively capture complex patterns within such datasets makes it a reliable choice for predictive modeling tasks.

On the other hand, multiple linear regression model and ridge regression model are not a perfect for the small dataset. The reason for the failure of the multiple linear regression model and ridge regression model is that both assumes that the relationship between the independent variables and the dependent variable is linear. If this assumption is violated, the model's predictions may be inaccurate. Also, ridge regression can help mitigate multicollinearity, but, it may not adequately capture complex non-linear relationships between the independent and dependent variables. In such cases, more flexible modeling techniques may be required. From the failure of the multiple linear regression and ridge regression it is also proven that the metrics used in the SBS Test, Iosipescu Test and DCB Test are nonlinear.

The results of feature importance analysis using the Random Forest Regressor (RFR) algorithm indicate that for interlaminar shear strength, the span-to-depth ratio of the laminate or specimen has the highest influence. It can be seen from [1], [6] and [5]. If the beam is too long compared to its depth, flexural failure (tensile or compressive) may occur at the outer plies of the beam. To ensure interlaminar shear failure prior to flexural failure, the span to depth ratio must satisfy the relationship.

$$\frac{L}{h} < \frac{F_1}{F_2} \tag{16}$$

where, L: Beam Length and F_i: flexure strength of beam in the fiber direction

Similarly, for in–plane shear strength, the thickness of the specimen has the greatest influence. Therefore, prioritizing changes in these factors is likely to have the most significant impact on achieving the desired mechanical property value. From [7] and [14] it can be seen that, in the Iosipescu shear test for in-plane shear strength, typically, the thickness of the specimen tends to have more importance. For carbon fibre composite it is seen that thicknesses have no significant difference in the shear strength in the elastic region, but the final shear stress is higher for the thicker specimen.

Finally, it can be concluded that decision tree is a perfect fit for small dataset with nonlinear mechanical property tastings, which may be positive as test data is limited and recreating more data just for the model is very time consuming and costly. Also, if new specimen of glass-epoxy is tried to design more focus should be put on the influential properties that would help in achieving the required mechanical property.

Future Works

It is crucial to continue exploring and experimenting with different modeling approaches, such as ensemble methods or feature engineering techniques, to further enhance predictive performance and gain deeper insights into the underlying data structures. Moreover, investing in data collection efforts to expand the dataset size could potentially improve the performance of more complex models like Artificial Neural Networks, thus warranting further exploration. Also, this approach can be extended for different mechanical property tests which might have an abundant data in open source. Overall, a balanced approach that combines the strengths of decision tree models with continuous exploration of alternative techniques will likely yield the most favorable outcomes in predictive modeling endeavors.

Looking ahead, there are several avenues for future exploration and enhancement in predictive modeling endeavors. Firstly, further research could focus on refining decision tree models by exploring advanced techniques such as ensemble methods (e.g., Random Forests, Gradient Boosting Machines) to improve predictive accuracy and robustness. Additionally, investigating feature engineering methods to extract more meaningful information from the existing dataset could lead to enhanced model performance. Expanding the dataset size through additional data collection efforts or incorporating external data sources could provide a richer and more comprehensive understanding of the underlying phenomena. This could potentially lead to improved model generalization and predictive capabilities.

Further research can be conducted to predict the behavior of long laminates in SBS tests. Additionally, investigations can explore the influence of thickness in In-Plane Shear on the interaction between glass fibers and carbon fibers. Future studies could also examine the effects of different resin matrices on interlaminar shear strength, and the impact of environmental factors such as temperature and humidity on the mechanical properties of composite materials. Advanced modeling techniques, such as finite element analysis, could be employed to simulate complex stress distributions and failure mechanisms in composite structures.

Overall, the future scope encompasses a wide range of possibilities, including refining existing models, exploring novel techniques, expanding datasets, and addressing the limitations observed in the current study. By embracing these opportunities, researchers can continue to advance predictive modeling capabilities and unlock new insights in various domains.

Acknowledgments

I would first like to express my gratitude to my committee chair, Dr. Scott W. Case, for his support and guidance throughout my time as a master's student at Virginia Tech. I would also like to thank him for inspiring me with different research ideas and guiding me throughout my project. In addition, I would like to thank Dr. Carin L. Roberts–Wollmann and Dr. Roberto T. Leon for serving on my committee. Finally, I would like to thank my parents, family, and friends, for their continued support and encouragement. I would not have been able to do this without them.

References

[1] ASTM D2344/D 2344M-00e1. *Standard Test Method for Short-Beam Strength of Polymer Matrix Composite Materials and Their Laminates.* (2017).

[2] ASTM D5379/D5379M-19e1. *Standard Test Method for Shear Properties of Composite Materials by the V-Notched Beam Method.* (2021).

[3] Wyoming Test Fixutures INC. *Short Beam Shear Test Fixture (ASTM D2344)*, (2024) Retrieved from: https://wyomingtestfixtures.com/products/shear/short-beam-shear-test-fixture-astm-d-2344

[4] Vinod Kushvaha, M. R. Sanjay, Priyanka Madhushri, Suchart Siengchin. *Machine Learning Applied to Composite Materials.* (2022) DOI: https://doi.org/10.1007/978-981-19-6278-3

[5] Krishan K. Chawla. *Composite Materials Science and Engineering Fouth Edition.* (2019)

[6] Isaac M. Daniel., and Ori Ishai. *Engineering Mechanics of Composite Materials, Second Edition.* (2006)

[7] David E. Walrath., and Donald F. Adams. *Analysis of the Stress State In An Iosipescu Shear Test Specimen* (June, 1983)

[8] Aslan, Z., and Almak, Y, *Characterization of interlaminar shear strength of laminated woven E-glass/epoxy composites by four point bend shear test*, The astrophysical journal **527** : 86 - 101 (1999) DOI: https://doi.org/10.1086/308062(2019)

[9] Fan, Z., Santare, M. H., and Advani, S. G., *Interlaminar shear strength of glass fiber reinforced epoxy composites enhanced with multi-walled carbon nanotubes*, (2008)

[10] Almeida, J. H. S., Angrizani, C. C., Botelho, E. C., and Amico S. C., *Effect of fiber orientation on the shear behavior of glass fiber/epoxy composites*, (2015)

[11] Turla, P., Kumar, S. S., Reddy, P. H., and Shekar, C., *Interlaminar Shear Strength of Carbon Fiber and Glass Fiber Reinforced Epoxy Matrix Hybrid Composite*, (2014)

[12] Selmy, A. I., Elsesi, A. R., Azab, N.A., and Abd El-baky, M. A., *In-plane shear properties of unidirectional glass fiber (U)/random glass fiber (R)/epoxy hybrid and non-hybrid composites*, (2012)

[13] Behrooz, F. T., Esmkhani, M., and Yaghoobi-Chatroodi, A., *Effect of testing procedure on the in-plane shear properties of CNF/glass/epoxy composites*, (2019)

[14] Ali H. Q., Aydin M. S., Khan R. M. A., and Yildiz M., *The role of "thickness effect" on the damage progression and crack growth inside the plain-woven carbon fiber composites.*, (2023)

[15] Medium. *The Mathematics behind Linear Regression.*, 8 March (2021) Retrieved from: https://pujappathak.medium.com/the-mathematics-behind-linear-regression

[16] Medium. *The Mathematics Behind the Regression Algorithms in Machine Learning.*, 24 August (2021) Retrieved from: https://organized-curiosity.medium.com/the-mathematics-behind-the-regression-algorithms-in-machine-learning

[17] Medium. *The Mathematics Behind Artificial Neural network.*, 17 November (2019) Retrieved from: https://towardsdatascience.com/the-heart-of-artificial-neural-networks

[18] Analytics Vidhya. *Introduction to Gradient Descent Algorithm (along with variants) in Machine Learning.*, 24 June (2019) Retrieved from: https://www.analyticsvidhya.com/blog/2017/03/introduction-to-gradient-descent-algorithm-along-its-variants/

[19] Medium. *Math Behind Artificial Neural network.*, 19 July (2020) Retrieved from: https://medium.com/analytics-vidhya/math-behind-artificial-neural-networks

[20] Belnoue, J., and Hallett, S. R., *Cohesive/Adhesive failure interaction in ductile adhesive joints Part I: A smeared-crack model for cohesive failure*, (2016)

[21] Shimadzu. *ASTM D5379 - Shear Properties of Composite Materials by the V–Notched Beam Method.*, 19 July (2020) Retrieved from: https://www.ssi.shimadzu.com/industries/automotive-materials-testing/composites/astm-d5379

[22] Stojcevskia, F., Hilditch, T., and Hendersona, L. C., *A modern account of Iosipescu testing*, (2018)

[23] Ungureanu, D., Taranu, N., Isopescu,D., Hudisteanu, I. and Lupasteanu, V., *IOSIPESCU SHEAR TEST FOR ADHESIVE MATERIALS AND BONDED FRP ELEMENTS*, (2019)

[24] Kumar, C., Rawat, P., Singh, K. K., BeheraR. P., and Deep, A., *Combined effect of loading rate and percentage by weight of MWCNTs on inter laminar shear strength (ILSS) and flexural strength of CFRP*, (2018)

[25] Kaya, Z., Balcioglu, H. E., and Gun, H., *The strain rate and temperature effects on the static and dynamic properties of S2 glass/epoxy composites*, (2018)

[26] Ramamuty, U., *Mechanical Testing Methods of Fibers and Composites. Mechanical Testing Methods of Fibers and Composites*, (2001)

[27] AskPython, *Decoding Entropy in Decision Trees: A Beginner's Guide*, 28 December (2023). Retrieved from: https://www.askpython.com/python/examples/entropy–decision–trees