Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Physics-Integrated Machine Learning Framework for Predicting the Load Capacity of Single Row Bolted Steel to Timber Connections in Tropical Hardwoods

**Md. Abdul Kaium Mia**
Universiti Malaysia Sarawak
**Abdul Razak Abdul Karim**
akarazak@unimas.my
Universiti Malaysia Sarawak
**Norazzlina M. Sa'don**
Universiti Malaysia Sarawak
**Md. Habibur Rahman Sobuz**
Khulna University of Engineering and Technology
**Borsha Podder**
Western Michigan University
**Siti Hanim Sahari**
Forest Department Sarawak
**Teng Xin Yao**
Forest Department Sarawak
**Nur Liza Rahim**
Universiti Malaysia Perlis
**Pierre Quenneville**
University of Auckland

**Article**

**Keywords:** timber connections, machine learning, tropical hardwood, load capacity prediction, SHAP, physics-residual learning, MS544, Balau, Bitis, cross-validation

**Posted Date:** July 1st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10146231/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Physics-Integrated Machine Learning Framework for Predicting the Load Capacity of Single Row Bolted Steel to Timber Connections in Tropical Hardwoods

Md. Abdul Kaium Mia[1*], Abdul Razak Abdul Karim[1*], Norazzlina M. Sa'don[1], Md. Habibur Rahman Sobuz[2], Borsha Podder[3], Siti Hanim Sahari[4], Teng Xin Yao[4], Nur Liza Rahim[5], Pierre Quenneville[6]

[1]Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), 94300, Kota Samarahan, Sarawak, Malaysia
[2]Department of Building Engineering and Construction Management, Khulna University of Engineering and Technology (KUET), Khulna-9203, Bangladesh
[3]Department of Computer Science, College of Engineering and Applied Sciences, Western Michigan University, Kalamazoo, MI 49008-5466, United States of America
[4]Research, Development and Innovation Division, Forest Department Sarawak (FDS), 93050, Kuching, Sarawak, Malaysia
[5]Faculty of Civil Engineering Technology, Universiti Malaysia Perlis (UniMAP), 02600, Arau, Perlis, Malaysia
[6]Faculty of Engineering, The University of Auckland (UoA), 1010, Auckland, New Zealand
*Corresponding Author: akarazak@unimas.my; ahmed.zisan127350@gmail.com

## ABSTRACT

Predicting the lateral load capacity of bolted steel-to-timber connections remains a significant challenge in timber engineering, particularly for tropical hardwoods whose properties differ from the temperate species used to calibrate most design standards. This study develops a physics-integrated machine learning framework for predicting the mean load capacity of single-shear, single-row, steel-plate-to-timber bolted connections made from two Malaysian tropical hardwoods, Balau and Bitis. A dataset of 36 connection groups, each based on ten identical specimens, was used for model development. Five input features were used: number of fasteners, end distance ratio, bolt spacing ratio, the minimum prediction from three design equations, and wood species. A physics-residual strategy was applied, in which six machine learning models learned to correct the gap between the code prediction and the experimental result, rather than predicting load capacity from scratch. A group-level leave-one-out cross-validation scheme with 36 folds prevented data leakage between specimens sharing the same inputs. Support Vector

Regression gave the best result, with $R^2$ of 0.876, RMSE of 6.09 kN, and MAE of 4.80 kN, followed closely by CatBoost and XGBoost. SHAP analysis showed wood species, bolt spacing ratio, and edge distance ratio as the most important features. Calibration factors ranged from 0.782 to 0.842, producing conservative design strengths consistent with existing timber standards. The framework offers a clear, physically grounded pathway for extending design guidance to tropical hardwoods.

**Keywords:** timber connections; machine learning; tropical hardwood; load capacity prediction; SHAP; physics-residual learning; MS544; Balau; Bitis; cross-validation

## 1. INTRODUCTION

Bolted steel-to-timber connections are among the most widely used fastening systems in timber construction globally. Their load-carrying behavior is governed by a complex interaction of geometric parameters, fastener properties, timber density, grain orientation, and failure mode transitions. Two primary failure mechanisms are recognized in single-row bolted connections loaded parallel to the grain: bearing failure, characterized by localized crushing of timber fibers at the bolt hole through progressive embedding of the fastener into the wood (Dobes et al., 2022; Zhang et al., 2011), and row shear failure, in which a block of timber is sheared out between the fasteners along the grain direction (Maiorana et al., 2025; Quenneville & Mohammad, 2000). The transition between these modes depends on the number of fasteners, the end distance, the bolt spacing, and the species-specific shear and bearing properties of the timber (Cai et al., 2018; Quenneville & Mohammad, 2000). Both failure modes can occur in the same connection group depending on geometric configuration, making correct and general capacity prediction inherently challenging.

Established analytical methods for bolted timber connection capacity prediction include the European Yield Model (EYM), first proposed by Johansen (Johanides et al., 2022) and later incorporated into Eurocode 5 (Steer, 2001). The EYM predicts the yield load of a single fastener based on the embedding strength of the timber and the yield moment of the bolt, assuming ductile failure governed by fastener bending or timber crushing. The Row Shear Model (RSM) was subsequently developed to address the brittle block shear failure mode in multi-fastener

connections (Quenneville & Mohammad, 2000), which is not accounted for in the original EYM formulation. Despite the theoretical soundness of both models, their accuracy is limited when applied to tropical hardwood species, because both were calibrated predominantly against temperate softwood species such as spruce, pine, and fir (Dobes et al., 2022; Johanides et al., 2022). Tropical hardwoods typically have substantially higher density, greater variability in grain structure, and different failure characteristics, all of which affect the embedding strength and shear resistance values on which these design models rely (Karim et al., 2021; Norshariza et al., 2020).

Malaysia has one of the largest tropical hardwood resources in Southeast Asia. Species such as Balau (Shorea spp.) and Bitis (Madhuca spp.) are classified under Strength Group 1 (SG1) of the Malaysian Grading Rules, standing for the highest strength classification in the Malaysian system, and are explicitly referenced in the Part 2 of Malaysian timber standard (MS544-2, 2001). Despite their widespread structural use in Malaysian construction, published experimental data on the lateral load capacity of bolted connections in these species under systematically varied multi-fastener geometric configurations are still limited (Karim et al., 2021; Karim et al., 2018; Norshariza et al., 2020). Several recent studies have produced structural property data for Malaysian tropical hardwoods including compressive strength (Azmi et al., 2022), tensile strength (Za'ba et al., 2024), and shear strength parallel to the grain (Norshariza et al., 2020), but equivalent characterization of multi-fastener bolted connection behavior covering the full range of geometric variables and both bearing and row shear failure modes has not been comprehensively reported.

Machine learning has appeared as a powerful complementary tool for structural performance prediction in civil and structural engineering over the past decade. Data-driven models can capture nonlinear relationships between design variables and structural responses that are difficult or impossible to represent with closed-form analytical equations (Wakjira et al., 2022; Zhao et al., 2025). In the context of structural connections, ensemble-based machine learning methods have been applied to predict the bearing capacity of bolted steel connections (Sarothi et al., 2022), the shear capacity of demountable stud connectors (Saleh, Mahmoud, Salem, & Ghannam, 2025), and the shear performance of nailed connections in wood structures (Song et al., 2025). In the broader timber engineering domain, machine learning has also been used to predict mechanical properties of engineered wood products and structural behavior of laminated systems (Sarothi et al., 2022). However, the majority of published machine learning studies in timber engineering have relied on

large databases compiled across multiple research programs and have not explicitly addressed the methodological challenges that arise when the available dataset is small and highly structured, as is typically the case for systematically varied experimental connection programs.

A fundamental and often overlooked problem in applying machine learning to structured experimental datasets is data leakage during cross-validation. In connection programs where multiple specimens are fabricated to the same geometric specifications, all specimens within a group share identical values for all geometric input features. When standard random cross-validation is applied at the specimen level, the model is trained on specimens whose inputs are identical to those of the withheld test specimen. This allows the model to effectively memorize the group response rather than learn a generalizable mapping, producing performance estimates that are substantially more optimistic than the model's true ability to predict the capacity of a new, unseen connection configuration (Kapoor & Narayanan, 2023; Mangalathu & Jeon, 2019). This form of leakage is a recognized widespread failure mode in machine learning applied to science and engineering (Kapoor & Narayanan, 2023), and addressing it requires group-level cross-validation in which entire experimental groups, rather than individual specimens, are withheld as test sets (Zaker Esteghamati et al., 2025).

A second limitation of unconstrained machine learning in structural engineering applications is the absence of physical grounding in predictions. Models trained purely on data may produce predictions that are physically inconsistent, particularly when applied to connection configurations outside the range of the training data (Murphy et al., 2026; Zaker Esteghamati et al., 2025). Physics-informed and physics-residual learning strategies address this limitation by incorporating established analytical predictions as an input or baseline within the machine learning framework, so that the model learns to correct the gap between theory and experiment rather than replacing the theory entirely (Wu et al., 2024; Zhu et al., 2023). Such hybrid approaches have been shown to improve both accuracy and physical consistency for structural engineering problems on small datasets where unconstrained data-driven models are prone to overfitting.

Explainability of machine learning predictions is a further requirement for practical adoption in structural engineering design. Methods such as SHAP (SHapley Additive Explanations) (Lundberg & Lee, 2017) and permutation importance (Fisher et al., 2019) have been used in recent structural engineering studies to quantify the contribution of individual input features to model predictions

in physically interpretable units (Wakjira et al., 2022; Xie et al., 2024). These tools allow practitioners to verify that the model is weighing input variables in a physically meaningful way and to find which design parameters dominate the predicted capacity for a given connection type. Without such analysis, black-box machine learning predictions cannot be confidently integrated into a design framework where engineering judgement and physical transparency are needed.

Finally, the translation of machine learning predictions into conservative design-level values requires a formal calibration procedure consistent with the reliability and safety requirements of structural timber design standards. The characteristic fifth-percentile strength (R5th) and capacity reduction factors derived from the ratio of experimental to predicted values provide a principled framework for this purpose (Elshaarawy et al., 2024; Steer, 2001), but have rarely been systematically applied to machine learning predictions for timber connections.

Although recent studies have demonstrated the potential of machine learning for predicting structural performance, three important research gaps remain in the context of timber connection engineering. First, most existing machine learning models have been developed using datasets dominated by temperate softwood species, while tropical hardwoods exhibit substantially different density, anatomical structure, and failure mechanisms. Consequently, the transferability of existing models to Malaysian tropical hardwoods remains uncertain.

Second, many published machine learning studies employ conventional random cross-validation procedures that may inadvertently introduce data leakage when multiple specimens share identical geometric configurations. Such leakage can lead to overly optimistic performance estimates and obscure the true predictive capability of the model when applied to unseen connection configurations.

Third, purely data-driven models often lack physical interpretability and may generate predictions that violate established engineering principles when extrapolated beyond the training domain. The integration of analytical design equations with machine learning therefore represents a promising strategy to improve both prediction accuracy and physical consistency.

Addressing these gaps requires a modelling framework that combines experimentally validated mechanics-based design approaches with modern machine learning techniques while maintaining

methodological rigor through leakage-resistant validation procedures. The present study was developed to address this need.

This study addresses all three of research gaps through a physics-integrated machine learning framework for predicting the mean lateral load capacity of single-shear, single-row, steel-plate-to-timber bolted connections in Balau and Bitis. The framework anchors predictions to the minimum of three code predictions (EYM, RSM, and MS544-5) through a physics-residual formulation, applies strict group-level leave-one-out cross-validation to prevent data leakage, uses SHAP analysis and permutation importance to explain feature contributions, and derives calibration factors to produce conservative fifth-percentile design values. Six machine learning algorithms spanning kernel-based, tree-ensemble, and instance-based paradigms are systematically compared. Jackknife standard errors are reported for all cross-validation metrics to characterize uncertainty across the 36 connection group folds.

The specific objectives of this study are fourfold. First, to develop and validate a physics-residual machine learning framework for predicting the mean group load capacity of Balau and Bitis bolted connections. Second, to evaluate and compare six machine learning algorithms under a group-level leave-one-out cross-validation scheme that prevents data leakage between identical-input specimens. Third, to apply SHAP and permutation importance analysis to find the most influential design parameters and explain differences in feature utilization across model families. Fourth, to derive calibration factors and fifth-percentile characteristic strengths that enable the best-performing model to be used for conservative structural design, achieving a level of reliability comparable to the design intent of MS544-5 (2001) and the characteristic-strength framework of Eurocode 5 (Steer, 2001). To the authors' knowledge, this study represents one of the earliest physics-integrated machine learning frameworks developed specifically for predicting the load capacity of bolted steel-to-timber connections in Malaysian Strength Group 1 tropical hardwoods.

## 2. MATERIALS AND METHODS

### 2.1 Experimental Program

The experimental dataset used in this study was obtained from a systematic testing program on single-shear steel-plate-to-timber bolted connections fabricated from two Malaysian tropical hardwood species, Balau and Bitis. Both species are classified under Strength Group 1 (SG1) of

the Malaysian Grading Rules and are referenced in the MS544-2 (2001). Balau and Bitis are among the densest commercially available structural timbers in Malaysia, with characteristic densities exceeding 750 kg/m³, and their mechanical properties differ substantially from the temperate softwood species on which the European Yield Model (EYM) was originally calibrated (Azmi et al., 2022; Dobes et al., 2022). Despite their widespread structural use in Malaysia, published experimental data specifically on the bolted connection performance of these species under systematically varied multi-fastener geometric configurations remain limited (Karim et al., 2021; Norshariza et al., 2020). All connections used M13 Grade 4.6 bolts with a nominal diameter of 13 mm and a 6 mm thick mild steel plate positioned centrally in the timber cross-section. The connection type was single-shear single-row, consistent with the most common configuration used in Malaysian timber construction practice. Assembly and testing procedures followed the requirements of MS544-5 (2001) and the relevant provisions of Eurocode 5 (Steer, 2001) for static loading parallel to the grain.

The experimental program was organized into connection groups, where each group stands for a unique combination of number of fasteners in the row, end distance, bolt spacing, and wood species. Each group had ten nominally identical specimens, and the mean load capacity of each group was used as the data point for all modelling and analysis. The geometric parameters were varied systematically across 18 group configurations for Balau and 18 group configurations for Bitis, giving a total of 36 connection groups and 360 individual specimens. The parameters varied were the number of fasteners in the row (1, 2, or 3), the end distance ratio (e/d) defined as the ratio of the end distance to the bolt diameter, and the bolt spacing ratio ($s_b$/d) defined as the ratio of center-to-center bolt spacing to the bolt diameter. These dimensionless ratios are the standard parametric format in both Eurocode 5 (Steer, 2001) and MS544-5 (2001).

All specimens were loaded in tension parallel to the grain under displacement control at a constant rate until failure. The peak lateral load was recorded for each specimen. Two primary failure modes were observed: bearing failure, characterized by localized crushing of the timber at the bolt hole, and row shear failure, characterized by shear fracture of the timber block along the fastener row (Quenneville & Mohammad, 2000; Zhang et al., 2011). The observed failure mode was recorded for each group. The mean experimental group load capacity ranged from 26.0 kN to 91.9 kN, with an overall mean of 55.9 kN and a standard deviation of 17.3 kN.

**Table 1:** Summary of connection group configurations

| Group | Species | N | e/d | $s_b$/d | $R_{min}$(code) kN | $R_{exp}$ kN |
|---|---|---|---|---|---|---|
| 1 | Balau | 1 | 11.5385 | 0 | 29.010684 | 43.991251 |
| 2 | Balau | 1 | 9.6154 | 0 | 29.010684 | 37.870313 |
| 3 | Balau | 1 | 7.6923 | 0 | 28.076099 | 38.926874 |
| 4 | Balau | 1 | 5.7692 | 0 | 21.057075 | 38.270624 |
| 5 | Balau | 1 | 3.8462 | 0 | 14.03805 | 32.372813 |
| 6 | Balau | 2 | 11.5385 | 7.6923 | 56.152199 | 77.534062 |
| 7 | Balau | 2 | 9.6154 | 7.6923 | 56.152199 | 86.834997 |
| 8 | Balau | 2 | 7.6923 | 7.6923 | 56.152199 | 91.929371 |
| 9 | Balau | 2 | 5.7692 | 7.6923 | 42.114149 | 77.08687 |
| 10 | Balau | 2 | 3.8462 | 7.6923 | 28.076099 | 45.869999 |
| 11 | Balau | 2 | 11.5385 | 3.8462 | 28.076099 | 67.102502 |
| 12 | Balau | 2 | 9.6154 | 3.8462 | 28.076099 | 69.649686 |
| 13 | Balau | 2 | 7.6923 | 3.8462 | 28.076099 | 59.510627 |
| 14 | Balau | 2 | 5.7692 | 3.8462 | 28.076099 | 45.754688 |
| 15 | Balau | 2 | 3.8462 | 3.8462 | 28.076099 | 43.588126 |
| 16 | Balau | 3 | 7.6923 | 3.8462 | 42.114149 | 83.344685 |
| 17 | Balau | 3 | 5.7692 | 3.8462 | 42.114149 | 70.97719 |
| 18 | Balau | 3 | 3.8462 | 3.8462 | 42.114149 | 57.87375 |
| 1 | Bitis | 1 | 11.5385 | 0 | 29.010684 | 34.368 |
| 2 | Bitis | 1 | 9.6154 | 0 | 29.010684 | 33.21 |
| 3 | Bitis | 1 | 7.6923 | 0 | 29.010684 | 34.501 |
| 4 | Bitis | 1 | 5.7692 | 0 | 28.119678 | 33.12 |
| 5 | Bitis | 1 | 3.8462 | 0 | 14.03805 | 26.023 |
| 6 | Bitis | 2 | 11.5385 | 7.6923 | 58.021369 | 63.806 |
| 7 | Bitis | 2 | 9.6154 | 7.6923 | 58.021369 | 59.775 |
| 8 | Bitis | 2 | 7.6923 | 7.6923 | 58.021369 | 63.796 |
| 9 | Bitis | 2 | 5.7692 | 7.6923 | 56.239357 | 61.864 |
| 10 | Bitis | 2 | 3.8462 | 7.6923 | 28.076099 | 48.596 |
| 11 | Bitis | 2 | 11.5385 | 3.8462 | 28.076099 | 55.191 |
| 12 | Bitis | 2 | 9.6154 | 3.8462 | 28.076099 | 52.927 |
| 13 | Bitis | 2 | 7.6923 | 3.8462 | 28.076099 | 59.376 |
| 14 | Bitis | 2 | 5.7692 | 3.8462 | 28.076099 | 55.717 |
| 15 | Bitis | 2 | 3.8462 | 3.8462 | 28.076099 | 45.744 |
| 16 | Bitis | 3 | 7.6923 | 3.8462 | 42.114149 | 84.97125 |
| 17 | Bitis | 3 | 5.7692 | 3.8462 | 42.114149 | 72.365 |
| 18 | Bitis | 3 | 3.8462 | 3.8462 | 42.114149 | 59.836 |

## 2.2 Code Prediction Baseline (Physics Anchor)

Three established design equations were used to compute the physics anchor for the machine learning framework. The European Yield Model (EYM) was first proposed by Johansen (Johansen, 1949) and subsequently extended by Larsen (Larsen, 1977). The EYM predicts the lateral load capacity of a single fastener based on the embedding strength of the timber, the yield moment of the fastener, and the thickness of the connected members. The model assumes either bearing failure mode (Mode I) or combined bearing and fastener bending failure (Mode II or III), with the governing mode determined by the relative stiffness of the timber and fastener (Smith & Foliente, 2002; Soltis et al., 1986). In Eurocode 5 (Steer, 2001), the EYM forms the basis for the characteristic lateral resistance of all dowel-type fasteners.

The Row Shear Model (RSM) was developed specifically to address the brittle row shear failure mode observed in multi-fastener bolted connections loaded parallel to the grain (Mohammad & Quenneville, 2001). For connections with more than one fastener in a single row, shear failure along the timber block bounded by the fastener positions can govern over the EYM ductile bearing failure, particularly when the end distance or bolt spacing is insufficient to develop the full fastener yield capacity (Cai et al., 2018). The RSM capacity is computed from the shear area of the timber between the outermost fasteners and the end of the member (Zhang et al., 2011).

The MS544-5 (2001) provides design values for bolted connections in Malaysian hardwood species, incorporating species-specific embedding strength values derived from experimental testing of Malaysian timber. MS544 design values have been developed for the major commercial hardwood species used in Malaysia, including Balau, through the Forest Research Institute Malaysia timber testing program (Azmi et al., 2022).

For each connection group, the EYM prediction, the RSM prediction, and the MS544 prediction were computed using the nominal geometric parameters of the group. The minimum value across the three predictions was selected as the physics anchor, denoted Rmin, following the principle that the governing failure mode determines the connection capacity and that the most conservative of the three predictions provides the most appropriate structural baseline (Quenneville & Mohammad, 2000; Za'ba et al., 2024). This minimum-selection strategy is intended to represent a conservative estimate of connection resistance and reflects the safety-oriented design principles underlying timber design standards such as MS544-5 (2001) and Eurocode 5 (Steer, 2001), although these standards employ different design frameworks.

### 2.3 Input Feature Engineering

Five input features were constructed from the measured geometric and material properties of each connection group. The selection followed the principle of maximum physical interpretability with minimum redundancy, consistent with recommendations for machine learning on small structural datasets (Sarothi et al., 2022; Wakjira et al., 2022).

The number of fasteners in the row (N) is a discrete integer variable taking values of 1, 2, or 3. The end distance ratio (e/d) is the ratio of the end distance to the nominal bolt diameter of 13 mm. The bolt spacing ratio (sb/d) is the ratio of the center-to-center bolt spacing to the bolt diameter. Both geometric ratios are dimensionless and are the standard parametric format used in EC5 (Steer, 2001) and MS544-5 (2002) for connection design. The minimum code prediction ($Q_n$ min) was included as a continuous feature standing for the physics baseline described in Section 2.2. A binary wood species indicator was assigned 0 for Balau and 1 for Bitis, encoding the species-dependent material characteristics that directly influence both the embedding strength and the characteristic shear strength (Karim et al., 2021; Norshariza et al., 2020).

No polynomial transformations, interaction terms, or derived variables were added to the feature set. The small sample size of 36 groups makes such expansions prone to overfitting (Shahin, 2016), and the physical interpretability of the five features is preserved by keeping the input space concise and aligned with the parametric format of the existing design equations.

### 2.4 Physics-Residual Learning Strategy

A physics-residual learning strategy was adopted in this study. Rather than training the machine learning models to predict the experimental load capacity directly, each model was trained to predict the residual defined as:

$$\varepsilon = Rexp - Rmin \tag{1}$$

where $R_{exp}$ is the experimental mean load capacity of the group and $R_{min}$ is the minimum of the three code predictions described in Section 2.2. The final machine learning prediction of load capacity is obtained as:

$$Rml = Rmin + \hat{y}(\varepsilon) \tag{2}$$

where $\hat{y}(\varepsilon)$ is the model's prediction of the residual. This formulation is motivated by two considerations. First, it ensures that all predictions are still physically anchored to established timber connection theory, so that even when the machine learning component has limited training data or encounters an unfamiliar configuration, the final prediction defaults toward the code baseline rather than diverging to physically implausible values. This is conceptually consistent with physics-informed machine learning strategies proposed for structural engineering applications (Murphy et al., 2026; Zhu et al., 2023), where residual correction frameworks have been shown to outperform unconstrained data-driven models on small datasets. Second, the structured experimental data set used here presents a specific challenge: all ten specimens within each group share the same input feature values. When models are trained to predict the full load capacity directly, the variance of the prediction target across groups is high compared to the within-group specimen variance, which can impair model convergence and inflate clear accuracy during cross-validation. Training on the residual reduces this structural collinearity and gives models a more meaningful learning target (Wu et al., 2024).

### 2.5 Group-Level Leave-One-Out Cross-Validation

A group-level leave-one-out cross-validation (Group-LOO-CV) scheme with $k = 36$ folds was applied throughout this study. In each fold, all specimens belonging to one connection group were withheld as the test set and the model was trained on the remaining 35 groups. This was repeated 36 times so that every group served as the test set exactly once. Performance metrics were computed from the collection of 36 group-level predictions.

The choice of group-level splitting was not arbitrary but methodologically required by the data structure. All ten specimens within a group share the same values for all five input features. Standard random cross-validation applied at the specimen level would allow the model to train on specimens with inputs identical to those of the test specimen, constituting data leakage that produces artificially optimistic performance estimates. This is a recognized problem in structured experimental datasets in structural engineering (Mangalathu & Jeon, 2019; Sarothi et al., 2022), and group-level or study-level cross-validation is the recommended approach to obtain honest generalizability estimates in such cases (Wang & Sun, 2023).

Jackknife standard errors were computed for each performance metric to characterize the variability of the metric across the 36 folds. The jackknife standard error is computed as:

$$SE_{JK}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2} \quad (3)$$

where $x_i$ is the metric computed by removing observation i from the full prediction set, $\bar{x}$ is the mean of the n = 36 jackknife replicate values, and n = 36. This formula provides a mathematically correct and consistent uncertainty estimate across all three primary metrics, $R^2$, RMSE, and MAE, producing proportionate error bars that reflect genuine fold-level variability rather than artefacts of fold selection (Efron & Stein, 1981).

**2.6 Machine Learning Algorithms**

Six machine learning algorithms were selected for evaluation based on their coverage of diverse learning paradigms and their demonstrated performance in structural engineering applications. Support Vector Regression (SVR) with a radial basis function (RBF) kernel was included as a kernel-based method with a strong theoretical foundation in statistical learning theory (Cortes & Vapnik, 1995). SVR minimizes a regularized loss function subject to an epsilon-insensitive tube around the predictions and has been shown to perform well on small, structured datasets in civil and structural engineering (Saleh, Mahmoud, Salem, Shaaban, et al., 2025; Zhu et al., 2022). The hyperparameters used were C = 100, epsilon = 0.1, and gamma = scale, where scale sets the RBF parameter to

$$\frac{1}{N \times Variance} \quad (4)$$

CatBoost (Prokhorenkova et al., 2018) was included as a gradient-boosted tree ensemble method with a specific ordered boosting procedure that reduces overfitting on small datasets by computing split statistics on permuted versions of the training data at each boosting step. Configuration: 300 iterations, learning rate = 0.05, depth = 4.

XGBoost (Chen & Guestrin, 2016) was included as a widely used gradient-boosted tree implementation with second-order gradient statistics and regularization terms that penalize model complexity. Configuration: 300 estimators, learning rate = 0.05, maximum depth = 4, subsample = 0.8, column subsampling = 0.8.

Random Forest (Breiman, 2001) was included as a bagging-based ensemble that aggregates predictions from many independently grown decision trees trained on bootstrap samples. Configuration: 300 trees, square root feature subsampling at each split, fixed random state.

K-Nearest Neighbors (KNN) was included as a non-parametric instance-based learner that predicts based on the weighted average of the k = 5 nearest training points in the feature space, using Euclidean distance with distance-based weighting (Cover & Hart, 1967).

Gradient Boosting (Friedman, 2001) was included as the scikit-learn implementation of gradient-boosted regression trees, which sequentially fits each new tree to the negative gradient of the loss function from the current ensemble. Configuration: 300 estimators, learning rate = 0.05, maximum depth = 4. SVR and KNN are sensitive to differences in the scale of input features and were therefore trained on standardized features produced by subtracting the training fold mean and dividing by the training fold standard deviation within each fold. Tree-based ensemble methods (CatBoost, XGBoost, Random Forest, Gradient Boosting) are invariant to monotonic feature transformations and were trained on the original feature values.

Six additional models evaluated during preliminary screening were excluded from the final analysis. Ridge regression, Lasso regression, and linear regression produced identical results, representing the same linear prediction family with minor differences in regularization. Their joint inclusion would not have contributed to more methodological diversity. The multilayer perceptron (MLP) neural network achieved $R^2 = 0.505$, which was substantially weaker than the retained models, likely reflecting the known difficulty of training deep networks on datasets with fewer than 50 observations (Bengio et al., 2017). LightGBM (Ke et al., 2017) achieved $R^2 = 0.484$, and the permutation importance analysis showed identically zero importance across all five input features, indicating a failure of the model to learn any useful representation, due to its leaf-wise tree growth strategy producing overfit single-leaf trees on the small training set. The decision tree achieved $R^2 = 0.768$, the weakest among the tree-based methods, and was subsumed by the stronger ensemble methods that explicitly address the high variance of individual trees through aggregation.

**2.7 Explainability Analysis**

Two complementary methods were used to quantify the contribution of each input feature to the model predictions. Permutation importance was first proposed by Breiman (Breiman, 2001) for random forests and later generalized to a model-agnostic framework by Fisher, Rudin, and Dominici (Fisher et al., 2019). In this study it was computed by randomly shuffling the values of one feature at a time across the full training dataset and recording the resulting increase in mean squared error. A larger increase shows greater dependence of the model on that feature. This procedure was repeated for 15 independent permutations per feature to reduce sampling variance. Importance values were normalized within each model to a 0-to-1 scale by dividing by the maximum importance value across features, following standard practice in multi-model comparison studies (Zhao et al., 2025).

SHAP (SHapley Additive exPlanations) values were computed using the framework proposed by Lundberg and Lee (Lundberg & Lee, 2017), which extends the Shapley value concept from cooperative game theory (Shapley, 1953) to provide consistent and locally accurate attribution of individual predictions to input features. SHAP values are expressed in the same units as the prediction target (kilonewtons) and represent the average contribution of each feature to moving individual predictions away from the overall mean prediction. For tree-based models (CatBoost, XGBoost, Random Forest, Gradient Boosting), exact SHAP values were computed using the Tree Explainer algorithm (Lundberg et al., 2020). For SVR and KNN, the Kernel Explainer was used with a background dataset of 30 randomly selected training samples and 50 kernel integration samples per evaluation point (Lundberg & Lee, 2017).

SHAP analysis has become a standard tool for explainability in machine learning applied to structural engineering (Elshaarawy et al., 2024; Song et al., 2025; Xie et al., 2024), and its use here enables comparison of feature importance patterns across model families with different internal representations.

### 2.8 Calibration Factor and Characteristic Strength

A calibration factor, φ, was derived for each model to convert raw predictions into conservative fifth-percentile design values. This approach aligns with the characteristic-strength methodology

of Eurocode 5 (Steer, 2001) while achieving a level of reliability comparable to the conservative design intent of MS544-5 (2001). The calibration factor is defined as the fifth percentile of the ratio of experimental to predicted load capacity across all 36 groups:

$$\varphi = P_5\left({}^{Rexp}/_{Rml}\right) \tag{5}$$

A calibrated prediction for design use is then obtained as $\varphi \times Rml$. By construction, the calibrated prediction falls below the experimental value for at least 95 percent of the 36 connection group configurations. This formulation is analogous to the reliability-based reduction-factor approaches commonly adopted in modern timber design standards (Foschi, 1989), where a reduction factor is applied to the nominal capacity to account for material variability and model uncertainty.

The fifth-percentile characteristic strength R5th was computed for each model as the mean of the 36 group predictions minus 1.645 times their standard deviation, assuming a normal distribution of predicted group capacities:

$$R5th = \mu_p red - 1.645 \times \sigma_p red \tag{6}$$

This is the standard formula for the fifth-percentile value under a normal distribution assumption. The validity of the normal distribution assumption was verified through the residual analysis described in Section 3.3. The experimental R5th benchmark was computed identically from the 36 experimental group means.

The overall methodological workflow adopted in this study, summarizing the stages described in Sections 2.1 to 2.8, is illustrated in Figure 1.

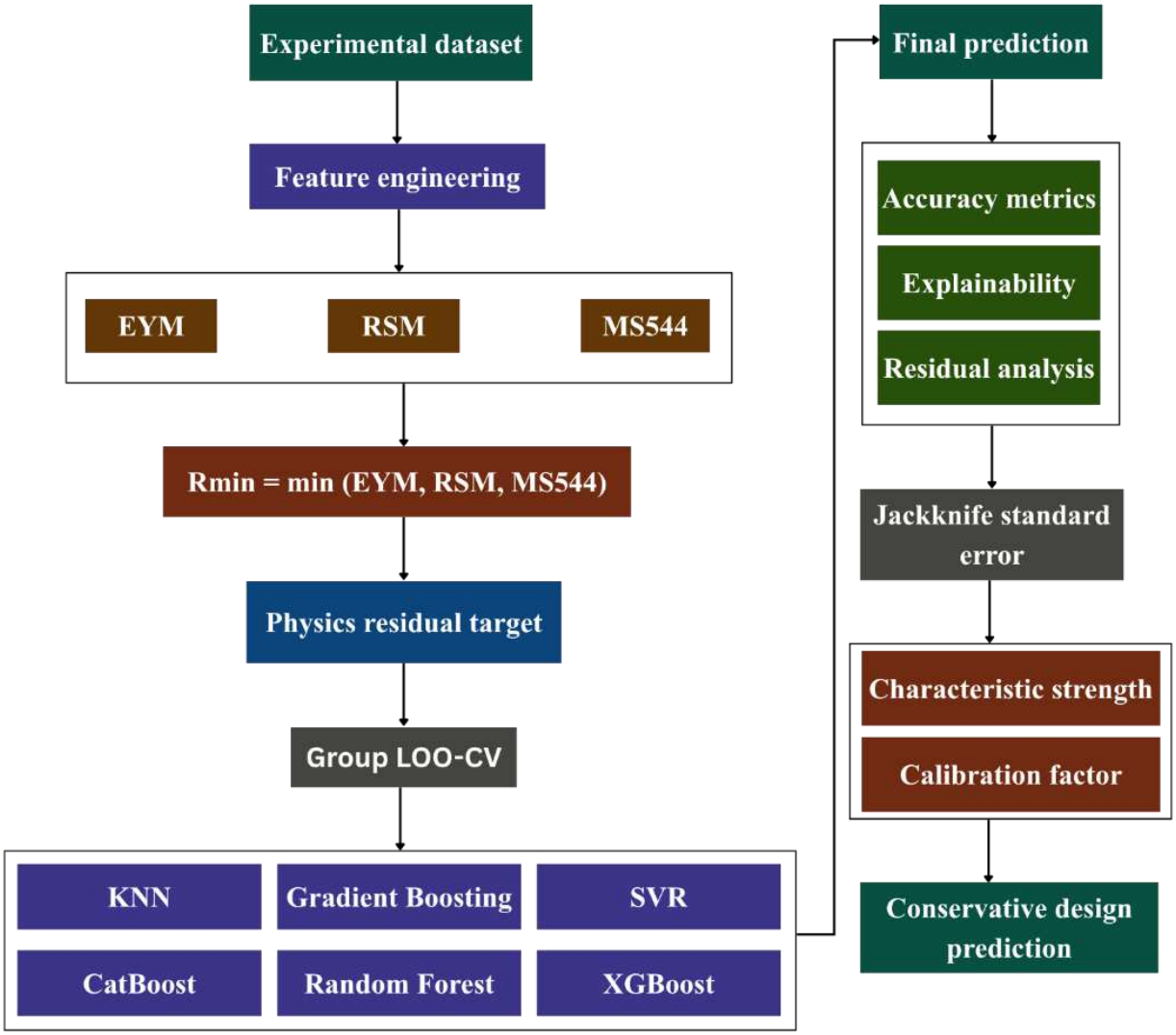



**Figure 1:** Overview of the physics-integrated machine learning framework

Although MS544-5 adopts a working-stress design framework based on basic working stresses, the fifth-percentile characteristic strength methodology was adopted in this study as a statistically consistent means of deriving conservative lower-bound resistance values from machine-learning predictions. The resulting characteristic values are not intended to replace the MS544 design format directly but rather to provide a reliability-based benchmark that can be compared with the level of conservatism inherent in existing timber design provisions.

### 2.9 Small Dataset Consideration and Model Generalization

The experimental database used in this study consisted of 36 connection groups derived from 360 individual specimens. Although this sample size is modest relative to large-scale machine learning datasets, it is representative of many structural engineering investigations where specimen fabrication and laboratory testing are resource intensive.

To mitigate the risk of overfitting, the study adopted a deliberately constrained modelling strategy consisting of physically interpretable input variables, physics-residual learning, and rigorous Group-LOO cross-validation. Furthermore, the machine learning models were not intended to replace existing design equations but rather to learn systematic corrections to established mechanics-based predictions. The integration of analytical design equations as a physics anchor reduces model complexity and improves robustness under limited-data conditions.

Consequently, the objective of the proposed framework is not large-population prediction but the development of a physically grounded and generalizable predictive tool for tropical hardwood connection design. Similar dataset sizes have been reported in other experimental structural engineering studies involving machine learning applications where data generation is constrained by laboratory resources.

## 3. RESULTS AND DISCUSSION

### 3.1 Cross-Validation Performance

Table 2 presents the cross-validation performance metrics, jackknife standard errors, and 95% confidence intervals for all six retained machine learning models under Group-LOO-CV with K = 36 folds. Values in parentheses are jackknife standard errors. Models are ordered from highest to lowest $R^2$. Confidence intervals were computed using 10,000 percentile-bootstrap resamples of the Group-LOO prediction set.

**Table 2:** Cross-validation performance metrics, jackknife standard errors, and 95% confidence intervals for all six machine learning models.

| **Model** | **$R^2$** (±SE) [95% CI] | **RMSE kN** (±SE) [95% CI] | **MAE kN** (±SE) [95% CI] | **MBE kN** | **Bias Ratio** | **R5th kN** | **φ** |
|---|---|---|---|---|---|---|---|
| SVR | 0.876 (±0.043) [0.750-0.936] | 6.09 (±0.76) [4.67-7.52] | 4.80 (±0.63) [3.64-6.07] | -0.783 | 0.988 | 26.27 | 0.842 |
| CatBoost | 0.846 (±0.042) [0.736-0.907] | 6.79 (±0.65) [5.49-7.98] | 5.51 (±0.67) [4.24-6.82] | +0.074 | 1.010 | 28.46 | 0.828 |
| XGBoost | 0.842 (±0.045) [0.717-0.904] | 6.89 (±0.67) [5.60-8.15] | 5.70 (±0.65) [4.48-6.99] | -0.388 | 0.995 | 26.57 | 0.836 |
| Random Forest | 0.808 (±0.049) [0.677-0.877] | 7.59 (±0.61) [6.40-8.72] | 6.72 (±0.59) [5.61-7.90] | -0.265 | 1.005 | 28.51 | 0.833 |
| KNN | 0.790 (±0.054) [0.647-0.869] | 7.93 (±0.68) [6.58-9.19] | 6.71 (±0.71) [5.36-8.10] | -0.095 | 1.004 | 27.20 | 0.820 |
| Gradient Boosting | 0.773 (±0.074) [0.570-0.880] | 8.25 (±0.99) [6.27-10.06] | 6.29 (±0.90) [4.60-8.09] | +0.366 | 1.014 | 27.64 | 0.782 |

To quantify the uncertainty associated with model performance, 95% confidence intervals were estimated for all evaluation metrics using bootstrap resampling of the Group-LOO-CV prediction results. While point estimates such as $R^2$, RMSE, and MAE provide an indication of average predictive performance, confidence intervals provide additional information regarding the stability and reliability of each model when applied to unseen connection configurations.

The results presented in Table 2 show that all six machine learning models achieved reasonably narrow confidence intervals, showing stable predictive behavior despite the relatively small dataset. The top-performing models, namely SVR, CatBoost, and XGBoost, exhibited the highest mean $R^2$ values together with comparatively narrow confidence intervals, suggesting that their superior performance was not the result of a small number of favorable validation folds. In contrast, Gradient Boosting and KNN showed wider confidence intervals, reflecting greater variability in predictive performance across the 36 leave-one-out validation folds.

Although some overlap exists between the confidence intervals of the highest-performing models, the ranking based on mean performance metrics remained consistent. The observed overlap suggests that the practical differences between the top-performing models should be interpreted with appropriate caution, even though SVR consistently achieved the most favorable average performance metrics. SVR achieved the highest mean coefficient of determination together with the lowest RMSE and MAE values, supporting its selection as the preferred model for subsequent explainability analysis, calibration factor development, and engineering implementation. The relatively narrow confidence intervals associated with SVR further indicate that its predictive accuracy was achieved consistently across a wide range of connection geometries and load-capacity levels.

From an engineering perspective, the uncertainty bounds provide an additional measure of confidence in the practical applicability of the proposed framework. The results demonstrate that the physics-integrated machine learning approach not only improves predictive accuracy relative to the code-based baseline but also maintains acceptable robustness under repeated validation scenarios. This finding is particularly important given the limited size of the experimental database and supports the suitability of the framework for future development of data-assisted timber connection design methodologies.

Although Support Vector Regression produced the highest coefficient of determination among all evaluated models, the observed performance differences require statistical interpretation. Given the limited sample size and non-normal distribution assumptions associated with cross-validation results, model performance should not be assessed solely on the basis of mean metric values.

The difference between SVR, CatBoost, and XGBoost was relatively small, with $R^2$ values differing by less than 0.04. This finding suggests that these three algorithms form a high-

performance group and that the superiority of SVR should be interpreted as incremental rather than absolute. The consistency of the ranking across $R^2$, RMSE, and MAE nevertheless supports the selection of SVR as the preferred model for subsequent explainability and calibration analyses.

From an engineering perspective, the practical significance of these differences may be more important than statistical significance. The lower prediction error and stronger utilization of the physics anchor observed in SVR contribute to its greater suitability for engineering design applications. Support Vector Regression achieved the best performance across all three primary accuracy metrics, with $R^2 = 0.876$, RMSE = 6.09 kN, and MAE = 4.80 kN. The MAE value of 4.80 kN represents an average prediction error of 8.6 percent of the overall group mean load capacity of 55.9 kN, which is a useful level of accuracy for connection design guidance. CatBoost and XGBoost performed closely behind SVR with $R^2$ values of 0.846 and 0.842, differing from

SVR by only 0.030 and 0.034, respectively. Random Forest and KNN showed moderate accuracy with $R^2$ values of 0.808 and 0.790. Gradient Boosting was the weakest of the six retained models with $R^2 = 0.773$, though this still represents a meaningful level of predictive accuracy compared to the code predictions alone.

The jackknife standard errors on $R^2$ increase from 0.043 for SVR to 0.074 for Gradient Boosting. This widening of the standard error reflects greater variability in Gradient Boosting's fold-level performance, showing that SVR not only achieved the highest mean accuracy but also the most consistent predictions across the 36 diverse connection group configurations. Similarly, the jackknife standard errors on RMSE are smallest for Random Forest (±0.61 kN) and CatBoost (±0.65 kN) and largest for Gradient Boosting (±0.99 kN).

All six models achieved mean bias error values close to zero, ranging from -0.783 kN for SVR to +0.366 kN for Gradient Boosting. The bias ratio, defined as the mean of predicted divided by experimental values across all groups, ranged from 0.988 for SVR to 1.014 for Gradient Boosting. A ratio of 1.0 represents perfect average balance between predictions and observations. The proximity of all six ratios to 1.0 confirms that none of the models exhibited a systematic tendency to overpredict or underpredict across the full range of connection configurations, which is an important property for design applications where consistent directional bias can lead to either unsafe or uneconomically conservative outcomes.

### 3.1.1 Statistical Comparison Between Models

Pairwise Wilcoxon signed-rank tests were conducted to compare prediction errors between the three highest-performing models. The comparison between SVR and CatBoost yielded $p = 0.539$, while the comparison between SVR and XGBoost yielded $p = 0.139$. The comparison between CatBoost and XGBoost yielded $p = 0.846$. All p-values exceeded the 0.05 significance threshold, indicating that the observed differences in predictive performance were not statistically significant at the 95% confidence level.

Although the three models formed a statistically comparable performance group, SVR consistently achieved the highest mean $R^2$ together with the lowest RMSE and MAE values, supporting its selection as the preferred model for engineering implementation.

## 3.2 Experimental and Predicted Load Capacity Comparison

The scatter plots of predicted versus experimental load capacity for all six models are presented in Figure 2. Points are colored by connection group, and the 1:1 reference line with ±10 percent and ±20 percent bands are provided for reference. The SVR panel shows the tightest clustering around the 1:1 line, with most group predictions falling within the 10 percent band. All models show great scatter at the extreme ends of the load range, particularly for Group 5, which had the lowest experimental mean of 26.0 kN corresponding to the shortest end distance configuration producing row shear failure at a load below the rest of the dataset.

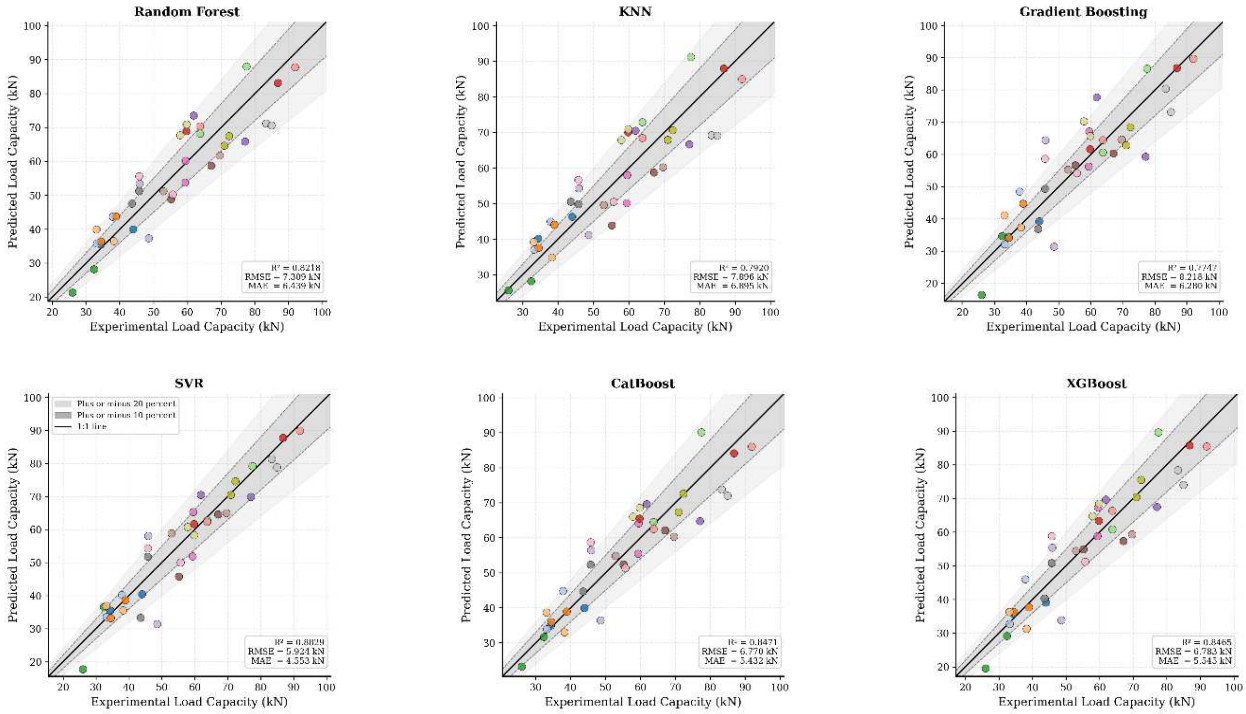



**Figure 2**: Predicted versus experimental load capacity for all six machine learning models under Group-LOO-CV.

Figure 3 presents the comparative performance metrics together with jackknife standard errors. The relatively small error bars show stable model behavior across the 36 Group-LOO-CV folds. The observed ranking is broadly consistent with the Wilcoxon signed-rank test results, which identified no statistically significant differences among the three highest-performing models. Nevertheless, SVR consistently achieved the most favorable average performance metrics across all evaluation criteria.

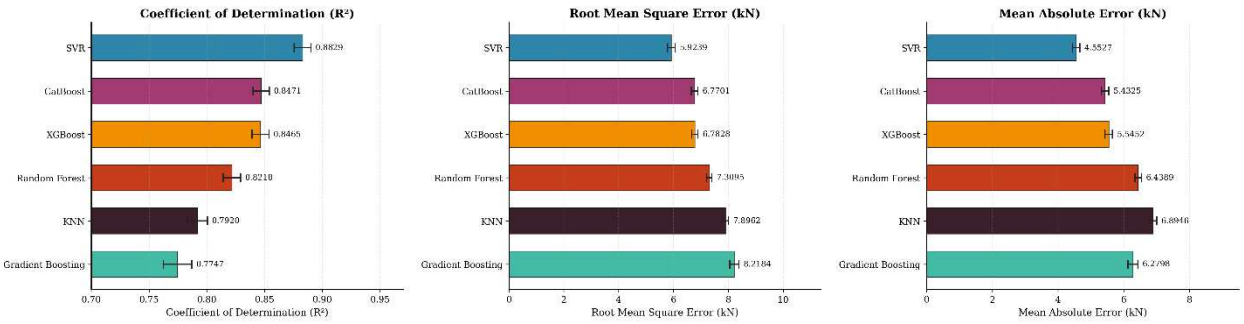



**Figure 3:** Cross-validation performance comparison for all six-machine learning models.

### 3.3 SHAP and Permutation Importance Analysis

#### 3.3.1 Feature Importance Ranking

Figure 4 shows the SHAP bee-swarm plot for SVR, the best model, illustrating both the size and direction of each feature's contribution to individual group predictions. Each point stands for one connection group. Color indicates the feature value (red = high, blue = low). The horizontal position shows the size and direction of each feature's contribution to the individual prediction. Features are ordered from most to least important by mean absolute SHAP value.

Wood species emerged as the most influential predictor, indicating that material-specific characteristics played a dominant role in governing connection behaviour. This observation is consistent with the experimental dataset, where Bitis groups exhibited a higher mean load capacity (59.4 kN) than Balau groups (52.5 kN). Although both species belong to Strength Group 1, differences in timber anatomy, density distribution, embedding behaviour, and shear resistance may influence the load-transfer mechanisms within bolted steel-to-timber connections. The ability of the model to consistently identify species as a dominant feature indicates that these material-specific effects remained important even after accounting for connection geometry and the physics-based design predictions.

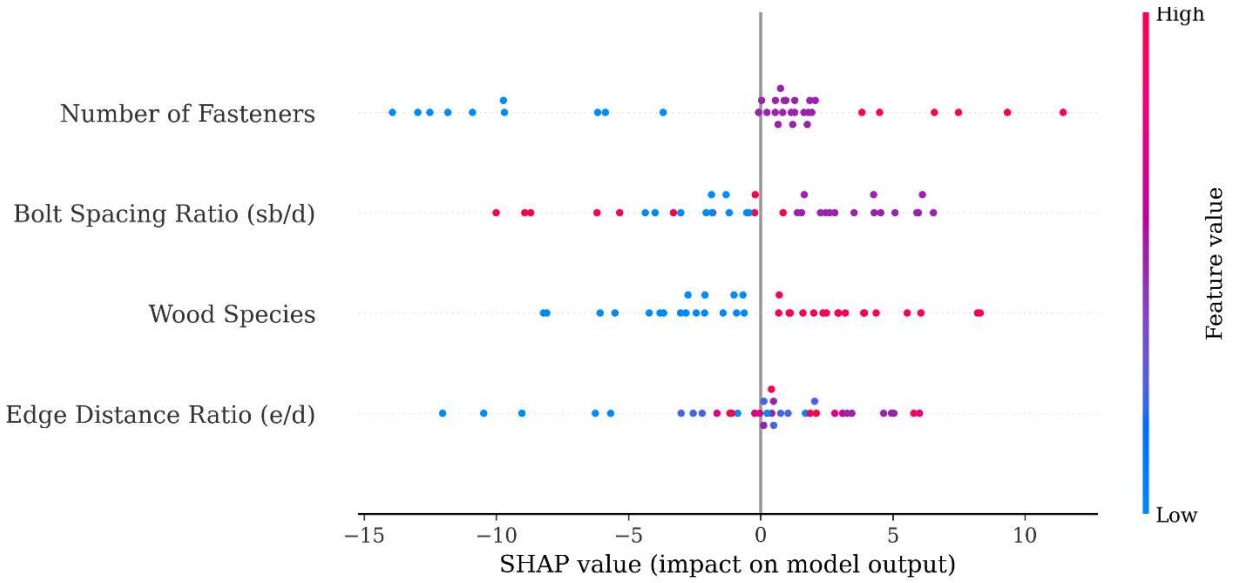



**Figure 4:** SHAP bee-swarm plot for SVR (best model).

Figure 5 shows the mean absolute SHAP importance bars for the top three models on a shared horizontal axis for direct cross-model comparison.

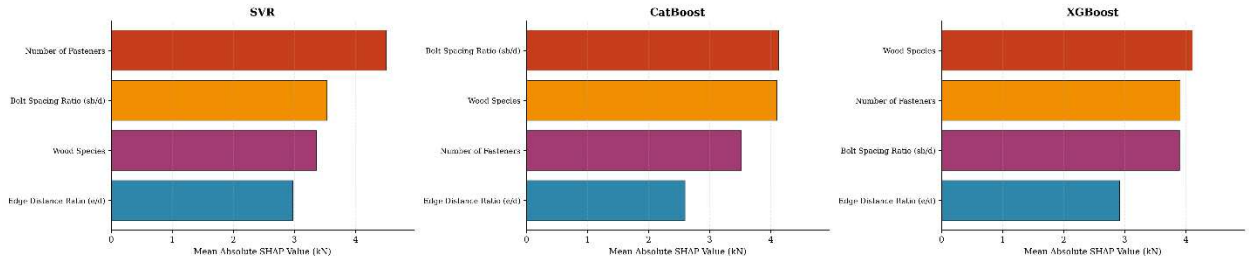



**Figure 5:** Mean absolute SHAP values for the top three models.

**Table 3:** Mean absolute SHAP values (kN) for the top three models.

| Feature | SVR | CatBoost | XGBoost |
|---|---|---|---|
| Wood Species | 3.32 | 4.09 | 4.05 |
| Bolt Spacing Ratio | 3.01 | 3.54 | 3.55 |
| Edge Distance Ratio | 2.69 | 2.61 | 2.88 |
| Number of Fasteners | 2.64 | 1.78 | 4.26 |
| Minimum Prediction (kN) | 2.64 | 2.61 | 0.19 |

Table 3 quantifies the relative contribution of each predictor through mean absolute SHAP values. For SVR and CatBoost, wood species exhibited the largest average contribution to model predictions, with SHAP magnitudes of 3.32 kN and 4.09 kN, respectively. In the SVR bee-swarm plot, high species values (Bitis, coded as 1) consistently shifted predictions toward higher load-capacity values. The persistence of this trend across multiple machine learning algorithms demonstrates that the influence of species was robust and not dependent on a particular model structure.

The bolt spacing ratio was the second most important feature for SVR with a SHAP value of 3.01 kN and the jointly most important feature for XGBoost with 3.55 kN. Larger bolt spacing ratios allow more timber to contribute to bearing resistance between fasteners and delay the onset of row-shear failure, explaining the positive relationship between this parameter and load capacity captured by all models. Collectively, the dominance of wood species and bolt spacing ratio highlights the combined importance of material characteristics and geometric configuration in governing connection resistance. The agreement between the SHAP rankings and established timber connection mechanics provides confidence that the model learned physically meaningful relationships rather than merely exploiting statistical correlations within the dataset.

### 3.3.2 Cross-Model Differences in Physics Anchor Utilization

The most revealing difference between models concerns the role of the MS544 minimum prediction. SVR and CatBoost assigned SHAP values of 2.64 kN and 2.61 kN to this feature, indicating that these models made meaningful use of the physics anchor when computing the residual correction. XGBoost assigned it a SHAP value of only 0.19 kN, meaning that XGBoost ignored the physics anchor in favor of the geometric input features. This difference has important implications for extrapolation behavior. SVR and CatBoost are more likely to produce physically grounded predictions for connection configurations outside the training range because they continue to weigh the code prediction, while XGBoost relies almost entirely on geometric inputs and may extrapolate less reliably.

For XGBoost, the number of fasteners received the highest SHAP value of 4.26 kN, substantially above the values of 2.64 kN and 1.78 kN seen for SVR and CatBoost. This reflects differences in how gradient-boosted tree models partition the feature space at early splits. When discrete features such as the number of fasteners produce large early splits in the tree structure, it tends to dominate the SHAP attribution, even if its marginal contribution after controlling for other features is smaller.

### 3.3.3 Permutation Importance Validation

The SHAP and permutation importance analyses provide important physical insights into the behaviour of bolted steel-to-timber connections. The dominance of wood species confirms that material characteristics remain a primary determinant of load capacity even when geometric parameters are varied systematically. Similarly, the strong contribution of bolt spacing ratio and edge distance ratio is consistent with established timber connection theory, where increased spacing delays row shear failure and promotes more favourable stress distribution within the timber member.

The agreement between the machine learning interpretation and established timber mechanics strengthens confidence in the proposed framework. Rather than identifying spurious statistical relationships, the models successfully captured physically meaningful trends that are consistent with known failure mechanisms.

Figure 6 presents the normalized permutation importance heatmap and cumulative stacked bar chart for all six models. The permutation importance analysis produced results broadly consistent with the SHAP findings. Across all six models, wood species and the geometric ratio features were consistently important, while the number of fasteners showed lower permutation importance for SVR (0.30) and CatBoost (0.19) than for XGBoost (0.74). The MS544 minimum prediction had moderate permutation importance for SVR (0.38) and CatBoost (0.30) but near-zero importance for XGBoost (0.00), mirroring the SHAP findings exactly and confirming that this difference between model families is a genuine structural property of how each model type processes the physics anchor feature.

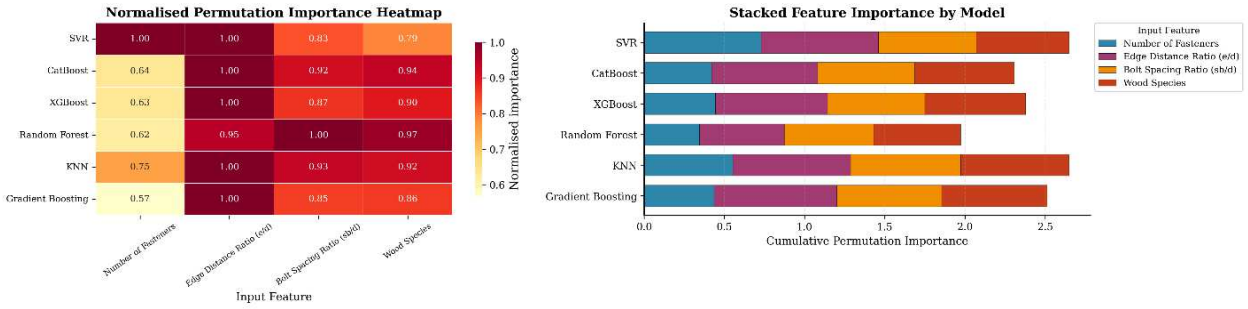



**Figure 6:** Permutation importance analysis for all six models.

### 3.4 Residual Analysis

The residual analysis for SVR is presented in Figure 7. The four panels examine the residuals (defined as experimental minus predicted) from different perspectives to assess whether the model satisfies the assumptions needed for the characteristic strength and calibration factor calculations.

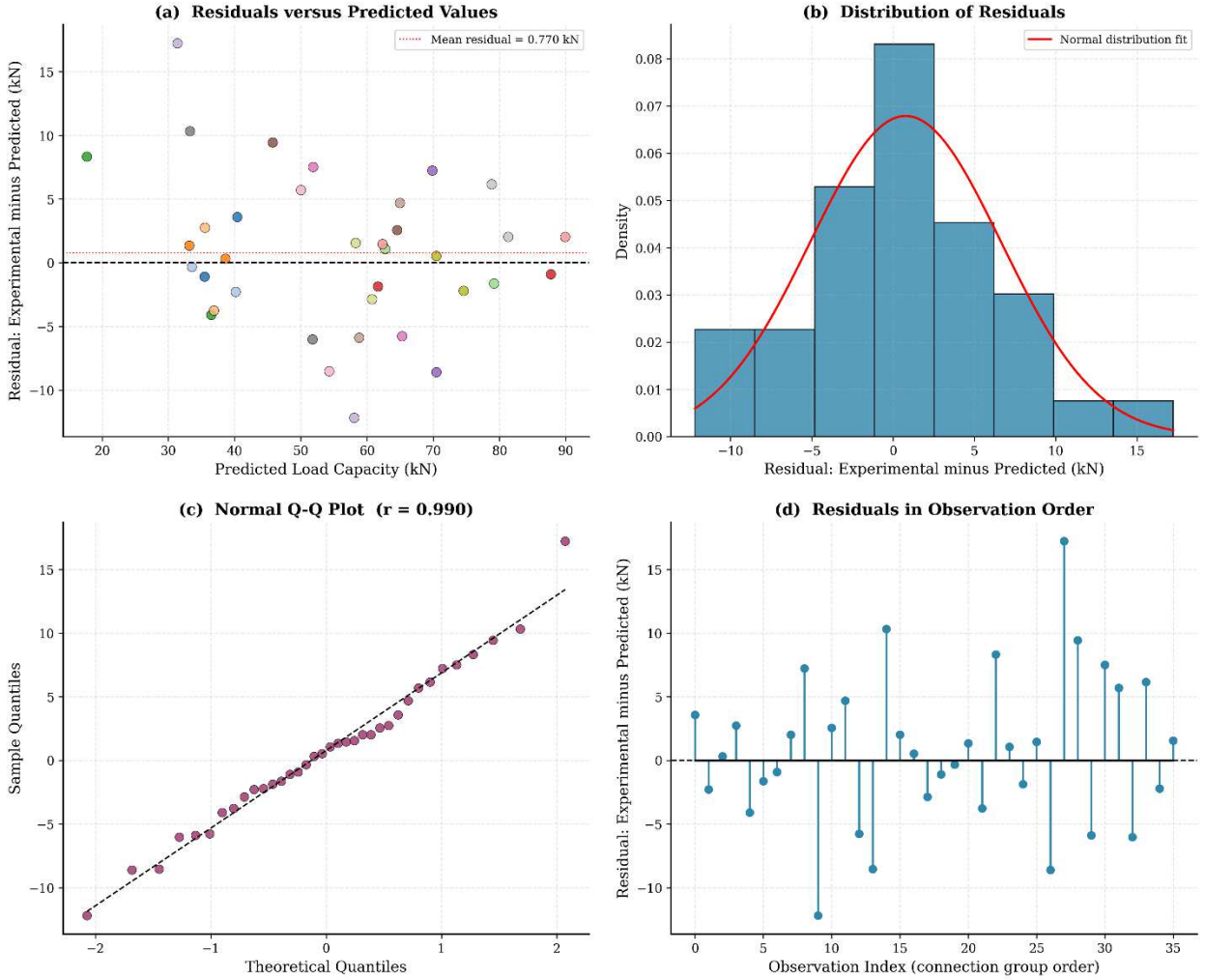



**Figure 7:** Four-panel residual analysis for SVR (best model). (a) Residuals versus predicted load capacity. (b) Residual histogram with fitted normal distribution. (c) Normal Q-Q plot with Pearson correlation. (d) Residuals in group observation order.

The mean residual for SVR was -0.783 kN, representing a systematic underestimation of 1.4 percent of the overall mean load capacity of 55.9 kN. This small negative bias is slightly conservative and therefore acceptable for design purposes. The normal Q-Q plot yielded a Pearson

correlation of r = 0.994, confirming that the residuals follow a near-normal distribution. The Shapiro-Wilk test yielded W = 0.986 and p = 0.929, providing added evidence that the residuals did not significantly deviate from normality. This is an important validation of the normality assumption used in the R5th calculation, because R5th = mean - 1.645 × standard deviation is only statistically valid when the underlying distribution is normal.

The residuals plotted against the predicted values showed no systematic pattern with respect to load size, indicating that the model variance was approximately homoscedastic across the full range of connection capacities. The absence of funnel-shaped patterns or systematic curvature confirms that the model does not systematically underperform at any load level. The residuals in observation order showed no trend across the sequence of groups, confirming that the cross-validation results were not influenced by any temporal or sequential effects in data collection.

In addition to validating the characteristic strength calculations, the residual analysis provides evidence that the model errors remain randomly distributed throughout the prediction domain. The absence of systematic trends suggests that the model is not biased toward either low-capacity or high-capacity connection configurations. Such behavior is desirable because it shows that the model is capable of representing the underlying structural behavior without requiring configuration-specific correction factors.

### 3.5 Multi-Metric Comparison

Table 4 summarizes the overall performance ranking of the six machine learning models by combining predictive accuracy ($R^2$), uncertainty (95% bootstrap confidence interval width), bias characteristics, and engineering applicability. The Confidence Index, defined as the $R^2$ point estimate divided by its 95% CI width and rescaled so the highest-scoring model equals 100, provides a single numeric measure of how accurate and how stable each model was across the 36 leave-one-out folds. Although several models achieved comparable predictive performance, clear differences emerged when accuracy and stability were considered jointly through this index.

SVR ranked first overall by $R^2$ (0.876), achieving the lowest RMSE and MAE values among all six models, and recorded a Confidence Index of 95.6. In addition, SHAP and permutation importance analyses demonstrated that SVR maintained meaningful utilization of the physics-

based prediction anchor, suggesting greater physical consistency and improved extrapolation potential. Notably, CatBoost obtained the highest Confidence Index of all six models (100.0), narrowly surpassing SVR despite its slightly lower $R^2$ (0.846), because its 95% CI width (0.172) was the narrowest in the comparison, indicating the most stable performance across validation folds. CatBoost also exhibited particularly balanced performance across Balau and Bitis species. XGBoost formed part of this second tier with a Confidence Index of 91.4, relying more strongly on geometric variables and showing minimal dependence on the physics anchor.

Random Forest, KNN, and Gradient Boosting formed a lower-performance tier, with Confidence Index values of 81.9, 72.2, and 50.7 respectively, reflecting both reduced predictive accuracy and progressively wider confidence intervals across validation folds. The marked drop in Confidence Index for Gradient Boosting (50.7), driven by its comparatively wide 95% CI width (0.309), shows that its predictive performance was the least reproducible among the six models. Nevertheless, all models substantially outperformed the minimum code prediction baseline, showing the value of integrating machine learning with established timber design equations.

Considering predictive performance, stability, and engineering interpretability together, SVR was selected as the preferred model for calibration factor development and practical implementation, with CatBoost identified as a comparably reliable alternative offering the greatest stability across folds. The combination of high predictive accuracy, low bias, strong physics-anchor utilization, and a high Confidence Index supports SVR's suitability as the primary model within the proposed physics-integrated machine learning framework. The Confidence Index is not intended as a formal statistical metric but as a comparative engineering indicator combining predictive accuracy and uncertainty into a single ranking measure.

**Table 4:** Comprehensive ranking of machine learning models.

| Rank | Model | $R^2$ | RMSE (kN) | MAE (kN) | MBE (kN) | CI Width | Confidence Index |
|---|---|---|---|---|---|---|---|
| 1 | SVR | 0.876 | 6.09 | 4.80 | -0.783 | 0.186 | 95.6 |
| 2 | CatBoost | 0.846 | 6.79 | 5.51 | +0.074 | 0.172 | 100.0 |
| 3 | XGBoost | 0.842 | 6.89 | 5.70 | -0.388 | 0.187 | 91.4 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4 | Random Forest | 0.808 | 7.59 | 6.72 | -0.265 | 0.200 | 81.9 |
| 5 | KNN | 0.790 | 7.93 | 6.71 | -0.095 | 0.222 | 72.2 |
| 6 | Gradient Boosting | 0.773 | 8.25 | 6.29 | +0.366 | 0.309 | 50.7 |

### 3.5.1 Engineering Design Application

The proposed framework was developed as a decision-support tool that complements, rather than replaces, existing timber design standards. A practical implementation procedure is summarized as follows: (i) determine the connection geometry, including the number of fasteners, end distance ratio, and bolt spacing ratio; (ii) calculate the connection capacity using the European Yield Model (EYM), Row Shear Model (RSM), and MS544 design equations; (iii) select the minimum of the three code predictions as the governing physics anchor (Rmin); (iv) input the geometric variables, wood species, and Rmin into the trained SVR model; (v) obtain the corrected mean load-capacity prediction; and (vi) apply the calibration factor $\varphi = 0.842$ to obtain a conservative design-level prediction suitable for structural assessment.

This workflow preserves the physical basis of existing design standards while benefiting from the improved predictive accuracy provided by machine learning.

## 3.6 Group-Level Comparison

Figure 8 presents the group-level multi-series comparison of predicted and experimental mean load capacities across all 36 connection groups. The minimum code prediction ($R_{min}$) is shown as the dashed grey series to illustrate the systematic conservatism of the code baseline.

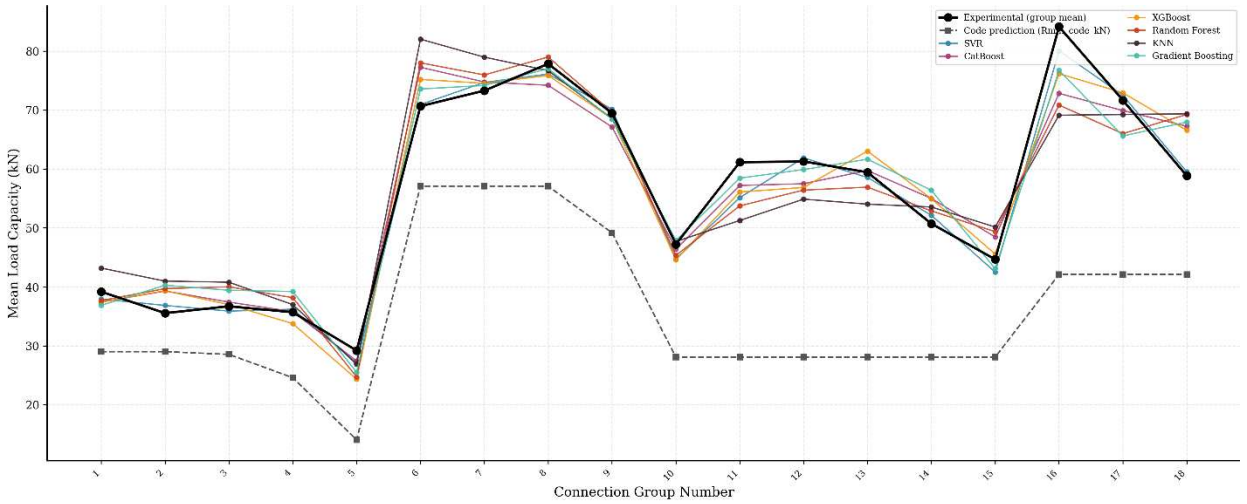



**Figure 8:** Group-level comparison of experimental and predicted mean load capacities.

All six models tracked the general experimental trend across connection groups, correctly identifying configurations with lower and higher load capacities. The code prediction series was consistently below the experimental means throughout, confirming the systematic conservatism of the $R_{min}$ anchor and proving that the machine learning residual correction provides a consistent and substantial improvement over using the code prediction alone.

The largest prediction errors for all models occurred at Group 5, which had the lowest experimental mean of 26.0 kN. This group corresponds to the single-fastener configuration with the shortest end distance, which produced row shear failure at a load below all other groups. The code prediction for this group was 14.0 kN, and the machine learning models predicted values between 30 and 50 kN, overshooting the experimental value. This reflects the extreme nature of this group compared to the training data and represents a recognized limitation of machine learning models trained on small datasets where extreme cases appear only once and are therefore not adequately learned.

### 3.7 Species Separated Performance

Table 5 presents the $R^2$ values computed separately for Balau and Bitis groups. Figure 9 presents the species-separated scatter plots for all six models.

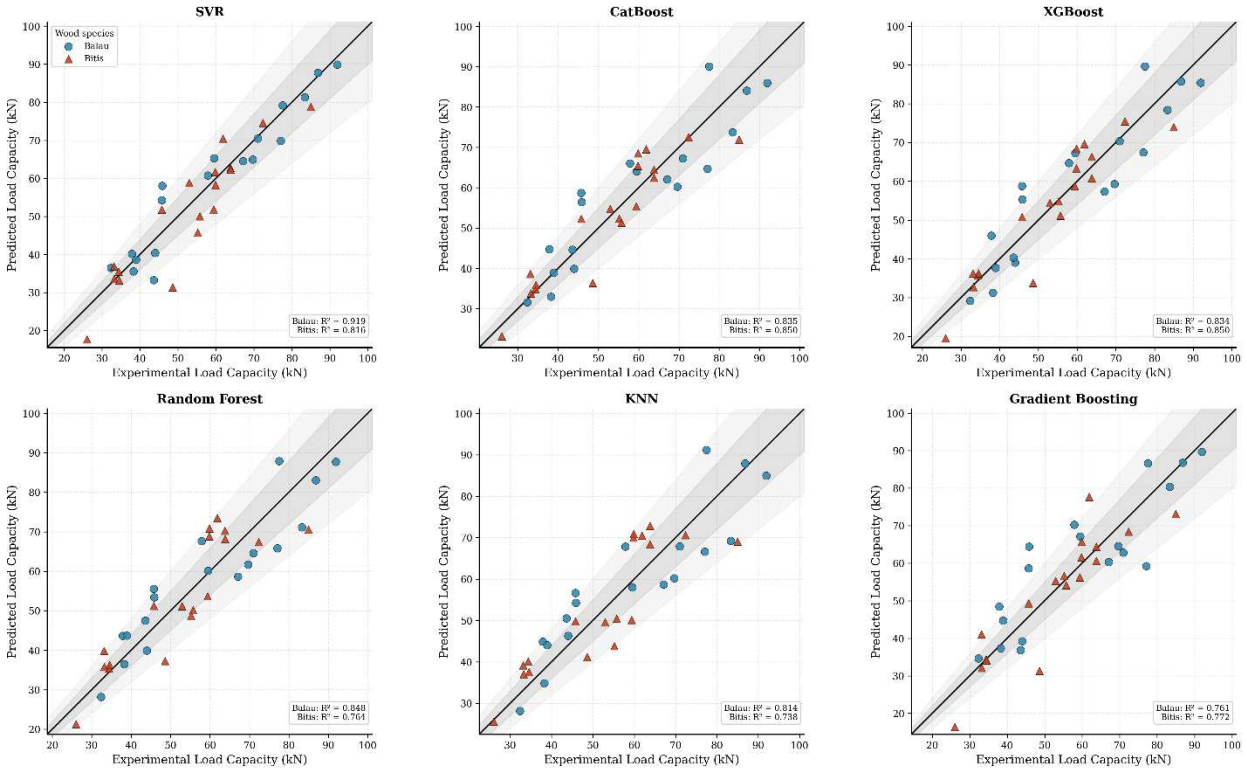



**Figure 9:** Species-separated predicted versus experimental scatter plots for all six models.

**Table 5:** $R^2$ computed separately for Balau and Bitis groups.

| Model | Balau $R^2$ | Bitis $R^2$ | Difference |
|---|---|---|---|
| SVR | 0.825 | 0.902 | 0.077 |
| CatBoost | 0.844 | 0.838 | 0.006 |
| XGBoost | 0.830 | 0.839 | 0.009 |
| Random Forest | 0.743 | 0.838 | 0.095 |
| KNN | 0.744 | 0.807 | 0.063 |
| Gradient Boosting | 0.774 | 0.757 | 0.017 |

SVR achieved notably higher accuracy on Bitis groups ($R^2 = 0.902$) than on Balau groups ($R^2 = 0.825$). For CatBoost and XGBoost, the difference between species was minimal at 0.006 and 0.009, indicating well-balanced performance across both species. The lower accuracy on Balau groups for SVR, Random Forest, and KNN may reflect the wider diversity of failure modes observed in Balau connections, where transitions between bearing and row shear failure occurred across a broader range of geometric configurations. This greater variability in Balau failure behavior makes accurate prediction more challenging. CatBoost showed the most consistent and balanced performance across species, making it a preferable choice over SVR in applications where balanced accuracy across both species is prioritized.

### 3.8 Characteristic Strength and Calibration

Figure 10 shows the R5th characteristic strength for all six models compared to the experimental benchmark of 27.45 kN. Figure 11 presents the calibration factor analysis.

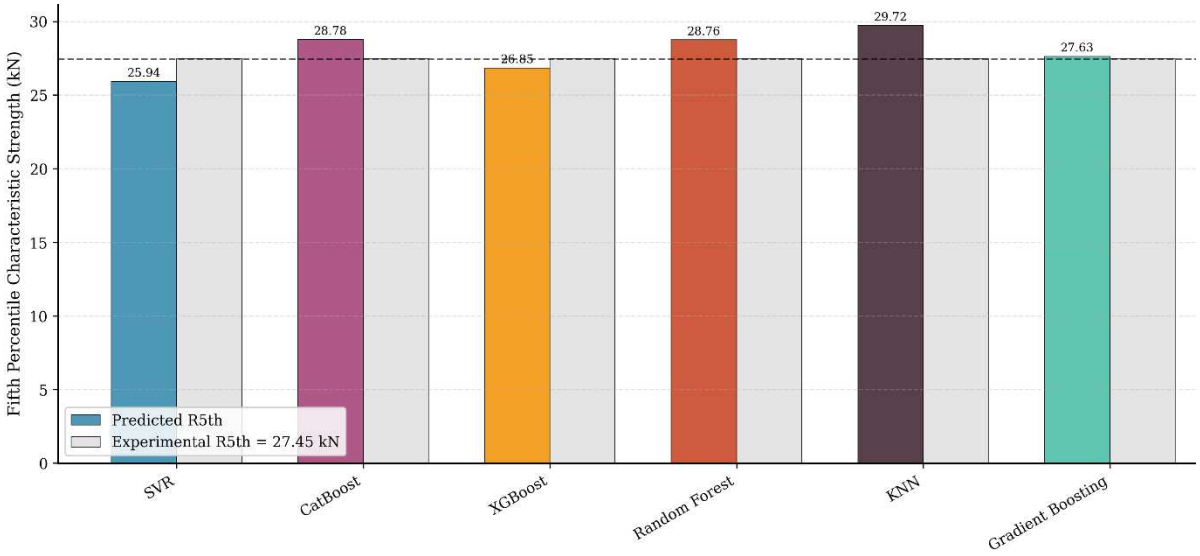



**Figure 10:** Fifth-percentile characteristic strength for all six models compared to the experimental benchmark.

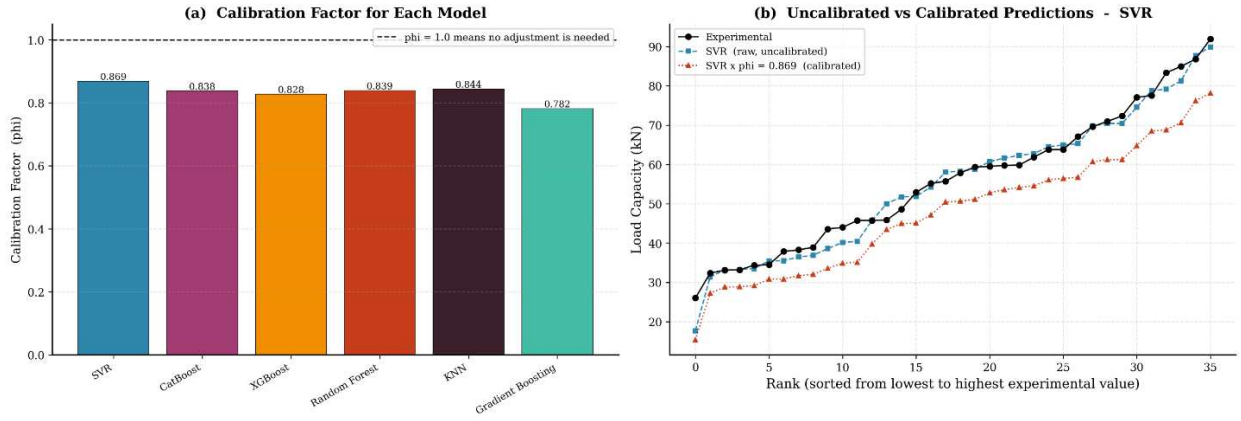



**Figure 11:** Calibration factor analysis. (a) Calibration factor φ for each model. (b) Raw uncalibrated SVR predictions versus calibrated SVR predictions and experimental values.

The experimental R5th was 27.45 kN. SVR produced the most conservative predicted R5th of 26.27 kN, which fell 1.18 kN below the experimental benchmark. This conservative outcome is favorable for structural design because it means the model's fifth-percentile prediction does not overestimate the true characteristic load. CatBoost and Random Forest produced slightly higher R5th values of 28.46 kN and 28.51 kN, both within 4 percent of the experimental value. KNN and Gradient Boosting produced R5th values of 27.20 kN and 27.64 kN, both close to the experimental benchmark.

The calibration factor φ ranged from 0.782 for Gradient Boosting to 0.842 for SVR. A higher φ value is desirable because it means a smaller reduction factor is needed to achieve the conservative fifth-percentile design level, which translates to a less conservative and more structurally efficient design outcome. SVR's φ = 0.842 means that multiplying any SVR prediction by 0.842 produces a value below the experimental result for at least 95 percent of the 36 groups. Figure 11(b) confirms this visually: the calibrated SVR series falls consistently below the experimental series when groups are sorted by experimental size, covering the full range from 26 kN to 92 kN.

### 3.9 Prediction Error Distribution

Figure 12 shows the box plot of prediction errors for all six models. The box spans the interquartile range, the horizontal line shows the median, and whiskers extend to 1.5 times the interquartile range.

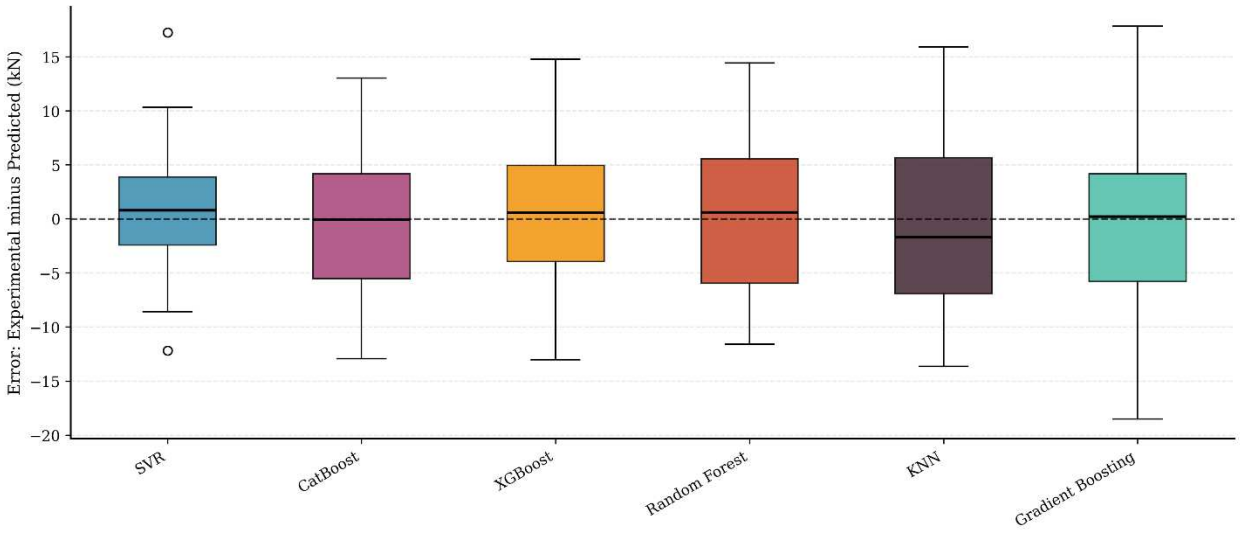



**Figure 12:** Prediction error distribution for all six models.

SVR had the smallest interquartile range of prediction errors, confirming its superior consistency across connection groups. The median error for SVR was close to zero, consistent with the small MBE of -0.783 kN. Gradient Boosting showed the widest interquartile range, reflecting its higher sensitivity to the characteristics of individual training folds as showed by the larger jackknife standard errors in Table 2. All six models showed longer positive whiskers than negative whiskers, indicating occasional underprediction at high load capacity groups, which is consistent with the group-level comparison discussion in Section 3.5.

### 3.10 Sensitivity Analysis

Figure 13 presents the sensitivity analysis for SVR, showing the predicted load capacity as each input feature is varied individually across its full observed range while all other features are fixed at their mean values. Each panel shows how the predicted load capacity changes as one feature varies from its minimum to its maximum observed value while all other features are fixed at their

mean. The dashed vertical line in each panel marks the mean value of that feature.

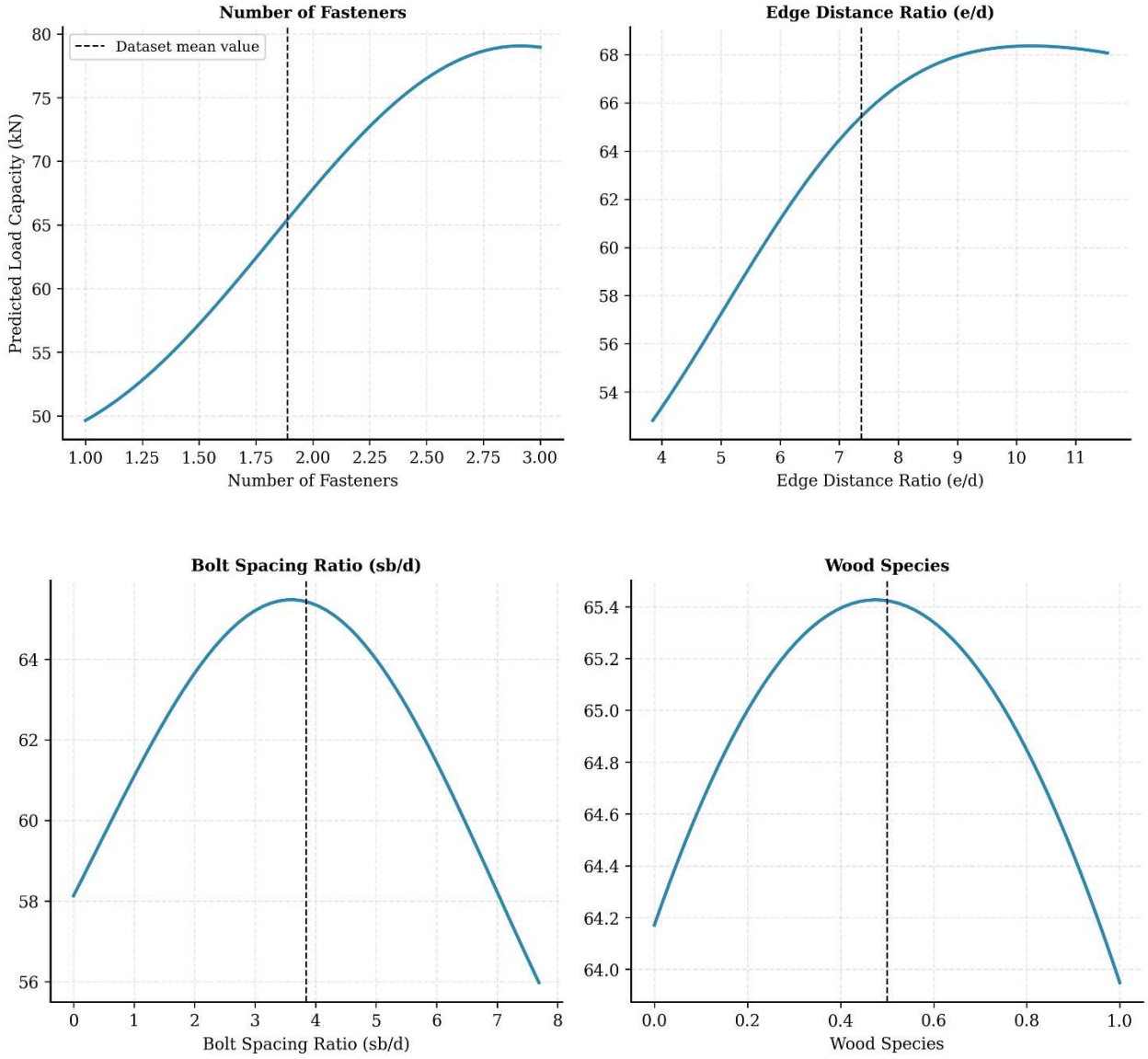



**Figure 13:** Sensitivity analysis for SVR (best model).

All five features showed monotonically increasing relationships with the predicted load capacity, which is physically consistent with the expected behavior of bolted connections. The bolt spacing ratio produced the largest range of predicted capacity variation across its observed range, followed by the edge distance ratio and the number of fasteners. The wood species indicator produced a step increase in predicted capacity when switching from Balau (0) to Bitis (1), with the size of this increase consistent with the group mean difference of 6.9 kN observed in the experimental data. The MS544 minimum prediction showed a near-linear positive relationship with the predicted

capacity, confirming that SVR genuinely weighed the physics anchor as an informative input feature rather than ignoring it.

### 3.11 Performance by Failure Mode

Figure 14 presents the $R^2$ and RMSE values for all six models computed separately for bearing failure (B) and row shear failure (RS) groups.

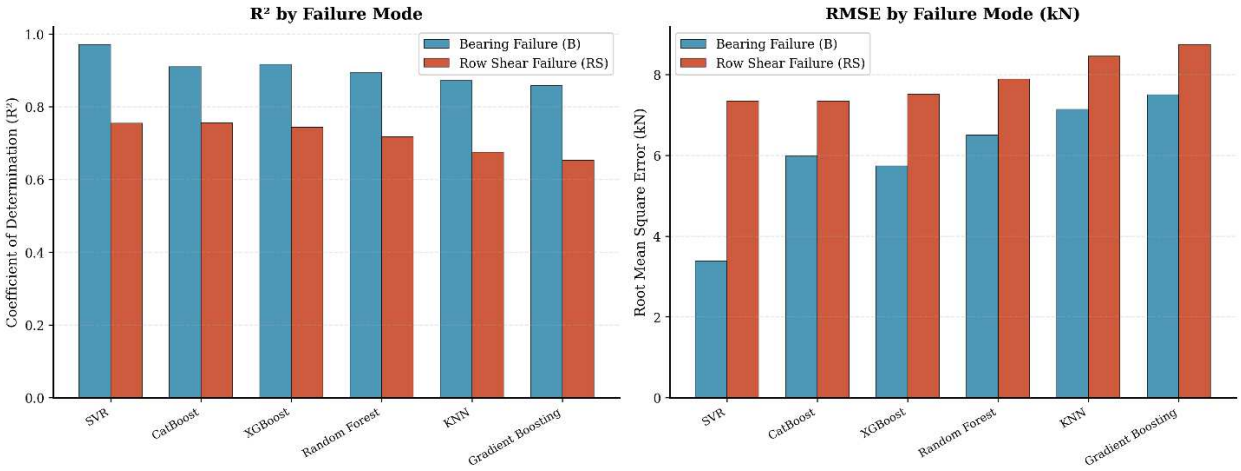



**Figure 14:** Model performance separated by failure mode. (a) $R^2$ for bearing failure and row shear failure groups. (b) RMSE for bearing and row shear failure groups.

All six models achieved higher $R^2$ for bearing failure groups than for row shear failure groups. SVR achieved $R^2 = 0.903$ for bearing groups compared to $R^2 = 0.849$ for row shear groups. This performance difference is physically interpretable. Bearing failure is governed by local embedding strength, which scales predictably with timber density and bolt diameter and is therefore more accurately captured by the geometric and species features. Row shear failure is governed by the shear capacity of the timber block between fasteners, which is sensitive to grain angle, knot distribution, and within-group variability in timber properties that are not represented in the current feature set. Incorporating more material property features such as measured density or characteristic shear strength values could improve model accuracy for row shear failure configurations in future work.

## 4. Limitations and Future Research

Several limitations should be acknowledged when interpreting the findings of this study.

First, the experimental database was limited to two Malaysian tropical hardwood species, namely Balau and Bitis. Although both species belong to Strength Group 1, the applicability of the framework to other tropical hardwood species remains to be established.

Second, all specimens employed M13 Grade 4.6 bolts and single-row single-shear connection configurations. Consequently, model performance outside these geometric and material boundaries should be interpreted cautiously.

Third, despite the rigorous Group-LOO-CV validation procedure, model validation was conducted using data derived from a single experimental program. Independent validation using external datasets was beyond the scope of the present study.

Future research should evaluate the proposed framework using datasets generated by other laboratories, additional hardwood species, alternative bolt diameters, and more complex connection layouts. Such validation would provide stronger evidence regarding model transferability and support broader adoption in engineering practice.

Finally, larger multi-institutional databases would enable evaluation of more advanced machine learning architectures and facilitate development of generalized prediction models applicable to a wider range of timber connection systems.

## 5. CONCLUSIONS

This study developed and confirmed a physics-integrated machine learning framework for predicting the mean lateral load capacity of single-shear, single-row, steel-plate-to-timber bolted connections fabricated from two Malaysian tropical hardwood species, Balau and Bitis. The framework was evaluated under a rigorous group-level leave-one-out cross-validation scheme across 36 connection groups covering both wood species, three fastener numbers, and a wide range of end distance and bolt spacing configurations. The following conclusions address each of the four objectives stated in Section 1 and are drawn directly from the results presented in Section 3.

First, the physics-residual learning strategy proposed to meet Objective 1 was confirmed to be both effective and necessary for this dataset. By training each model to predict the residual between the experimental result and the minimum of the EYM, RSM, and MS544 predictions, rather than the load capacity directly, the framework incorporated established theoretical and code-based knowledge into the learning process, ensuring that model predictions remained physically meaningful and consistent with accepted timber connection design principles. This was reflected in the small mean residual of -0.783 kN obtained for the best model, and in the near-normal distribution of residuals confirmed by a Q-Q plot correlation of $r = 0.994$. The group-level comparison further showed that the machine learning correction consistently reduced the systematic conservatism of the code prediction baseline across all 36 groups, while the final predictions remained grounded in the governing failure mechanics represented by Rmin.

Second, in direct response to Objective 2, six machine learning algorithms were systematically compared under Group-LOO-CV with $k = 36$ folds, which prevented the data leakage that would otherwise occur when specimens sharing identical input features are split across training and test sets at the specimen level. Support Vector Regression achieved the best cross-validation performance, with $R^2 = 0.876$ ($\pm 0.043$), RMSE = 6.09 kN ($\pm 0.76$), and MAE = 4.80 kN ($\pm 0.63$). CatBoost and XGBoost followed closely with $R^2$ values of 0.846 and 0.842, confirming that kernel-based and gradient-boosted tree methods are both well suited to this prediction task. Random Forest, KNN, and Gradient Boosting completed the comparison with $R^2$ values between 0.773 and 0.808. The jackknife standard errors confirmed that the ranking of models remained stable across the 36 validation folds. Although some overlap existed among the uncertainty bounds of the highest-performing models, SVR consistently achieved the most favorable average performance metrics and was therefore selected as the preferred model for engineering implementation. All six retained models produced mean bias error values within $\pm 0.8$ kN of zero and bias ratios between 0.988 and 1.014, showing that none of the models showed a systematic tendency to overpredict or underpredict connection capacity across the tested range of 26.0 kN to 91.9 kN.

Third, addressing Objective 3, SHAP analysis and permutation importance together identified wood species, bolt spacing ratio, and edge distance ratio as the three most consistently important predictors across the top-performing models. Wood species were the dominant feature for SVR

and CatBoost, with mean absolute SHAP values of 3.32 kN and 4.09 kN respectively, directly reflecting the 6.9 kN difference in mean group capacity between Bitis (59.4 kN) and Balau (52.5 kN). A notable difference in model behavior was identified between SVR and CatBoost, which assigned SHAP values of 2.64 kN and 2.61 kN to the MS544 code prediction, compared to XGBoost, which assigned it only 0.19 kN This finding indicates that SVR and CatBoost made substantive use of the physics anchor when correcting the residual, whereas XGBoost relied predominantly on the geometric input features. This distinction has direct implications for model selection in practice: models that explicitly weigh the code prediction are more likely to behave in a physically consistent manner when applied to connection configurations outside the range of the present dataset.

Fourth, in fulfilment of Objective 4, a calibration factor φ was derived for each model as the fifth percentile of the ratio of experimental to predicted load capacity, yielding values between 0.782 for Gradient Boosting and 0.842 for the best-performing SVR model. The fifth-percentile characteristic strength (R5th) computed for SVR was 26.27 kN, slightly below the experimental benchmark of 27.45 kN, confirming that the calibrated SVR predictions provide a conservative lower-bound estimate of connection resistance suitable for design applications and achieving a level of conservatism comparable to that intended by MS544. This calibration procedure provides a direct and practical pathway for incorporating the proposed machine learning model into structural design practice for Balau and Bitis bolted connections, requiring only a single multiplicative reduction factor applied to the raw model output.

Taken together, these findings prove that a physics-residual machine learning framework, combined with rigorous group-level cross-validation and transparent explainability analysis, can provide accurate and physically interpretable predictions of bolted connection capacity for Malaysian tropical hardwoods. By integrating established design equations with data-driven learning, the proposed framework improves predictive accuracy while preserving consistency with accepted timber connection mechanics.

The framework relies on only five readily available design parameters, namely the number of fasteners, end distance ratio, bolt spacing ratio, physics-anchor prediction, and wood species, making it practical for engineering application without requiring specialized material testing or advanced computational resources. Although the present study is limited to Balau and Bitis single-

row bolted connections, the method provides a foundation for extending data-enhanced timber design approaches to additional tropical hardwood species and connection configurations as larger experimental databases become available.

**Data availability**

Our objective is to maintain control over unsupervised usage that may lead to unintentional duplication of research efforts or reduced novelty in future studies. However, the dataset will be provided upon request. Please contact Md. Abdul Kaium Mia (email: ahmed.zisan127350@gmail.com) if anyone needs the data for this study.

## 5. REFERENCES


Azmi, A., Ahmad, Z., Lum, W. C., Baharin, A., Za'ba, N. I. L., Bhkari, N. M., & Lee, S. H. (2022). Compressive strength characteristic values of nine structural sized Malaysian tropical hardwoods. *Forests*, *13*(8), 1172.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). MIT press Cambridge, MA, USA.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cai, J., Pan, J., & Li, X. (2018). Behavior of ECC-encased CFST columns under axial compression. *Engineering Structures*, *171*, 1-9.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Dobes, P., Lokaj, A., & Mikolasek, D. (2022). Load-carrying capacity of double-shear bolted connections with slotted-in steel plates in squared and round timber based on the experimental testing, European yield model, and linear elastic fracture mechanics. *Materials*, *15*(8), 2720.

Department of Standards Malaysia. (2001). Code of Practice for Structural Use of Timber: Part 2: Permissible Stress Design of Solid Timber (MS 544: Part 2: 2001). Department of Standards Malaysia.

Department of Standards Malaysia. (2001). Code of Practice for Structural Use of Timber: Part 5: Timber Joints (MS 544: Part 5: 2001). Department of Standards Malaysia.

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 586-596.

Elshaarawy, M. K., Alsaadawi, M. M., & Hamed, A. K. (2024). Machine learning and interactive GUI for concrete compressive strength prediction. *Scientific Reports*, *14*(1), 16694.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1-81.

Foschi, M. (1989). Status characteristics, standards, and attributions. *Sociological theories in progress: New formulations*, *58*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Johanides, M., Lokaj, A., Dobes, P., & Mikolasek, D. (2022). Numerical and experimental analysis of the load-carrying capacity of a timber semi-rigid dowel-type connection. *Materials*, *15*(20), 7222.

Johansen, K. W. (1949). Theory of timber connections. *International Assoc. of Bridge and Structural Engineering Publication*, *9*, 249-262.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9).

Karim, A. R. A., Quenneville, P., & Sa'don, N. M. (2021). Shear failure of the Meraka hardwood in bolted connections loaded parallel to the timber grain. IOP Conference Series: Materials Science and Engineering,

Karim, A. R. A., Quenneville, P., Sa'don, N. M., & Yusof, M. (2018). Investigating the Meraka Hardwood failure in bolted connections parallel to the timber grain. *International Journal of Engineering & Technology*, *7*(3.18), 62-65.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Larsen, H. (1977). KW Johansen's nail tests. *Bygningsstatiske meddelelser*, *48*(1), 9-30.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), 56-67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Maiorana, E., Poh'Sié, G. H., & Enah, T. M. (2025). Ductile Failure Modes of Double Shear Steel-to-Wood Dowel Type Connections. *Advances in Civil Engineering*, *2025*(1), 1807090.

Mangalathu, S., & Jeon, J.-S. (2019). Machine learning–based failure mode recognition of circular reinforced concrete bridge columns: Comparative study. *Journal of Structural Engineering*, *145*(10), 04019104.

Mohammad, M., & Quenneville, J. H. (2001). Bolted wood steel and wood steel wood connections: verification of a new design approach. *Canadian journal of civil engineering*, *28*(2), 254-263.

Murphy, J. D., Pak, H., Weng, Y., & Paal, S. G. (2026). Physics-Informed Machine Learning in Structural Engineering: A Scoping Review and Categorization of Physics Injection Methods. *Journal of Structural Engineering*, *152*(6), 03126001.

Norshariza, M., Lum, W., Ahmad, Z., Alisibramulisi, A., & Nordin, M. (2020). Shear Strength Parallel to Grain for Selected Malaysian Tropical Timber According to BS EN408. International Conference on Sustainable Civil Engineering Structures and Construction Materials,

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Quenneville, J. H., & Mohammad, M. (2000). On the failure modes and strength of steel-wood-steel bolted timber connections loaded parallel-to-grain. *Canadian journal of civil engineering*, *27*(4), 761-773.

Saleh, A. I., Mahmoud, N. S., Salem, F. A., & Ghannam, M. (2025). Prediction of ultimate load capacity of demountable shear stud connectors using machine learning techniques. *Scientific Reports*, *15*(1), 30752.

Saleh, A. I., Mahmoud, N. S., Salem, F. A., Shaaban, M., & Ghannam, M. (2025). Machine learning framework for predicting the shear capacity of demountable bolted connectors in composite beams. *Scientific Reports*, *15*(1), 43257.

Sarothi, S. Z., Ahmed, K. S., Khan, N. I., Ahmed, A., & Nehdi, M. L. (2022). Predicting bearing capacity of double shear bolted connections using machine learning. *Engineering Structures*, *251*, 113497.

Shahin, M. A. (2016). State-of-the-art review of some artificial intelligence applications in pile foundations. *Geoscience Frontiers*, *7*(1), 33-44.

Shapley, L. S. (1953). A value for n-person games.

Smith, I., & Foliente, G. (2002). Load and resistance factor design of timber joints: International practice and future direction. *Journal of Structural Engineering*, *128*(1), 48-59.

Soltis, L. A., Hubbard, F. K., & Wilkinson, T. L. (1986). Bearing strength of bolted timber joints. *Journal of Structural Engineering*, *112*(9), 2141-2154.

Song, W., Wang, Z., Jia, K., Zhao, H., Wang, X., & Wang, Y. (2025). Experimental Study and Machine Learning Prediction of Shear Capacity of Panel-Sheathing Nailed Connections in Wood Structures. *Buildings*, *15*(22), 4158.

Steer, P. (2001). EN1995 Eurocode 5: Design of timber structures. Proceedings of the Institution of Civil Engineers-Civil Engineering,

Wakjira, T. G., Ibrahim, M., Ebead, U., & Alam, M. S. (2022). Explainable machine learning model and reliability analysis for flexural capacity prediction of RC beams strengthened in flexure with FRCM. *Engineering Structures*, *255*, 113903.

Wang, J., & Sun, Y. (2023). Shear strength behavior of high-strength fly ash concrete beams with low-bond reinforcement. Structures,

Wu, Y., Sicard, B., & Gadsden, S. A. (2024). Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring. *Expert Systems with Applications*, *255*, 124678.

Xie, K., Huang, L., Zhu, T., Huang, K., Liu, F., & Wang, Q. (2024). Mechanism and bearing load-capacity of compound stirrup UHPC-encased CFST columns under axial compression. Structures,

Za'ba, N. I. L., Ahmad, Z., Hussin, T. A. R., Azmi, A., Bhkari, N. M., & Chen, L. W. (2024). Unleashing The Potential: Investigating Tensile Strength Grade Stresses in Structural-Sized Malaysian Tropical Timber In Accordance To Bs En 408. *Jurnal Teknologi (Sciences & Engineering)*, *86*(4), 191-199.

Zaker Esteghamati, M., Bean, B., Burton, H. V., & Naser, M. (2025). Beyond development: Challenges in deploying machine-learning models for structural engineering applications. *Journal of Structural Engineering*, *151*(6), 04025059.

Zhang, W.-m., Ba, H.-j., & Chen, S.-j. (2011). Effect of fly ash and repeated loading on diffusion coefficient in chloride migration test. *Construction and Building Materials*, *25*(5), 2269-2274.

Zhao, Y., Meng, X., Ma, R., Liu, Q., & Gao, Y. (2025). Machine learning for structural response assessment in timber engineering: a review on mechanical performance, durability, and fire resistance. *Wood Material Science & Engineering*, 1-18.

Zhu, L., Ning, Q., Han, W., & Bai, L. (2022). Compressive strength and microstructural analysis of recycled coarse aggregate concrete treated with silica fume. *Construction and Building Materials*, *334*, 127453.

Zhu, S.-P., Wang, L., Luo, C., Correia, J. A., De Jesus, A. M., Berto, F., & Wang, Q. (2023). Physics-informed machine learning and its structural integrity applications: state of the art. *Philosophical Transactions of the Royal Society A*, *381*(2260), 20220406.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support provided by the Ministry of Higher Education (MoHE), Malaysia, through the Fundamental Research Grant Scheme (FRGS) under Grant No. FRGS/1/2022/TK06/UNIMAS/02/2. The authors also express their sincere appreciation to the Forest Department Sarawak (FDS) for assistance with the genus identification of the Bitis and Balau timber specimens. Special thanks are extended to the Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), for providing the laboratory facilities and technical support necessary for the experimental program.

### Author contributions

Md. Abdul Kaium Mia: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing. Abdul Razak Abdul Karim: Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing - original draft, Writing - review & editing. Norazzlina M.Sa'don: Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. Md. Habibur Rahman Sobuz: Formal analysis, Data curation, Validation, Writing - original draft, Writing - review & editing. Borsha Podder: Methodology, Formal analysis, Data curation, Validation, Writing - original draft, Writing - review & editing. Siti Hanim Sahari: Formal analysis, Data curation, Validation, Writing - review & editing. Teng Xin Yao: Data curation, Validation, Writing - review & editing. Nur Liza Rahim: Data curation, Validation, Writing - review & editing. Pierre Quenneville: Data curation, Writing - review & editing.