Explainable Boosting Machine for Structural Health Assessment: An Interpretable Approach to Data-Driven Structural Assessment

MIR MOHAMMAD SHAMSZADEH, KRISHNA KUMAR, ANCA-CRISTINA FERCHE, OGUZHAN BAYRAK and SALVATORE SALAMONE

ABSTRACT

Machine learning models used in structural health monitoring often act as "black boxes," offering predictions without justifying their logic. This lack of transparency undermines trust in safety-critical infrastructure assessments. To solve this, we propose the Explainable Boosting Machine, an interpretable method that explicitly links input variables (e.g., sensor data, and structural parameters) to predictions, enabling engineers to validate results against engineering principles. Real-world structural health monitoring and assessment struggles with sparse data, structural complexity, and hidden biases. Explainable Boosting Machine addresses these challenges by prioritizing transparency and physically meaningful insights. We apply it to predict the shear load-carrying capacity as a percentage of the ultimate load, based on the maximum diagonal crack widths observed on the surface of reinforced concrete beams—a critical metric for shear failure risk. Our results show that the model achieves an RMSE of 10.40% on the test dataset while identifying the influence of key predictors (e.g., beam depth, shear and skin reinforcement ratios). For instance, the model reveals that, for the same maximum diagonal crack width observed in two beams, a structure with a larger depth is farther from failure compared to the one with a smaller depth, enabling engineers to audit model logic and enhance structural assessment. This work advances trustworthy AI in structural health monitoring by bridging data-driven innovation and engineering accountability. Interpretability of explainable boosting machine ensures models remain consistent with physical laws, actionable for decision-making, and adaptable to real-world constraints. We advocate for machine learning frameworks that prioritize transparency as rigorously as predictive performance.

Mir Mohammad Shamszadeh[1], Krishna Kumar[1], Anca-Cristina Ferche[1], Oguzhan Bayrak[1], Salvatore Salamone[1]

[1]Department of Civil, Architectural, and Environmental Engineering, University of Texas at Austin, Austin, Texas, U.S.A.

INTRODUCTION

Despite the growing use of machine learning (ML) models for structural health monitoring (SHM) and damage assessment, their predictions do not always reflect the underlying physical behavior of the monitored systems. Such a disconnect reduces their practical applicability and raises concerns about their credibility in real-world SHM applications. In high-stakes engineering problems, where data is typically limited, understanding a model's reasoning is essential to trust its generalizability. Furthermore, when models fail, it is crucial to identify the cause and contributing factors, a task hindered by black-box models. This makes it difficult to assess their reliability, especially beyond the training domain. The issue of model transparency and accountability is crucial because not all model-generated approximations are physically meaningful. In engineering applications, many models may fit the data, but only those consistent with physical laws and principles are meaningful. Scientific knowledge plays a key role in model selection by helping to identify and eliminate physically inconsistent solutions, thereby minimizing model variance [1].

The trade-off between model complexity and interpretability often dictates the choice of algorithm for a given application. On one end of the spectrum, complex models such as deep neural networks offer high accuracy but suffer from being black boxes, with their decisions often not easily decipherable. On the other extreme, simpler models like linear regression and decision trees provide high interpretability through their easily traceable decision-making processes. However, they may lack the necessary accuracy for complex datasets and fail to capture intricate patterns.

To address the trade-off between model complexity and interpretability in predictive modeling and to ensure the consistency of the learned model with the physics of the problem, this study proposes the use of Explainable Boosting Machine (EBM) [2,3] as a modeling approach tailored for tabular datasets commonly encountered across many SHM applications. EBM provides a transparent modeling framework that not only offers interpretable predictions but also maintains high accuracy. This dual capability is crucial for determining when the model aligns well with the underlying physics of the problem and when it requires cautious interpretations. This approach tackles key predictive modeling challenges, improving reliability and interpretability.

To the best of the authors' knowledge, this research introduces the pioneering use of EBM in SHM and damage assessment. While prior applications of EBM in structural engineering have focused on predicting outcomes like strength directly from input features that are inherently related to those outcomes, this work addresses a setting where health index or damage index is predicted and some features act as moderator variables. This necessitates careful construction of model terms and thoughtful interpretation. It also presents practical ideas for interpreting results and new insights that could be obtained from the data regarding the damage behavior.

EXPLAINABLE BOOSTING MACHINE

Explainable Boosting Machine (EBM) is a glass-box model that is built on the Generalized Additive Model (GAM) framework [4]. GAM predicts the response variable as the additive combination of nonlinear functions, one for each feature, which reflects the relationship between the feature and the response variable. While GAM

effectively captures the influence of individual features, it does not account for interactions between them. To address this, Generalized Additive Models plus Interaction (GA^2M) was developed [2], which adds a small number of pairwise interaction terms to the univariate terms. This enables GA^2M models to achieve higher accuracy, often outperforming more complex models. EBM is a fast and parallelizable implementation of GA^2M, developed in C++ and Python [3].

Let $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote a dataset of size N. Each record includes a feature vector $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)}\right)^T \in \mathbb{R}^d$ and a corresponding response variable $y^{(i)} \in \mathbb{R}$ for a regression problem. Here, $\boldsymbol{x}^{(i)}$ represents the feature vector of the i-th record, and x_j specifically refers to the j-th feature in the feature space. EBM models the predicted response, $\hat{y}$, in an additive form as:

$$\hat{y} = g(\mathbb{E}[y|\boldsymbol{x}]) = f_0 + \sum_{j=1}^{d} f_j(x_j) + \sum_{\substack{1 \le j < k \le d \\ (j,k) \in \mathfrak{T}}} f_{jk}(x_j, x_k) \tag{1}$$

where f_0 is the intercept term, $f_j(x_j)$ is a univariate shape function which represents the main effect of the feature x_j on the response variable, $\mathfrak{T}$ represents the set of feature pairs, $f_{jk}(x_j, x_k)$ is a bivariate shape function which represents the pairwise interaction between features x_j and x_k on the response variable, and $g(.)$ is the link function.

In the EBM training process (see Figure 1), the model incrementally learns and refines its predictions through an iterative boosting framework. In each boosting round, the model is constructed by sequentially training an ensemble of bagged shallow decision trees (DTs) for each univariate shape function, $f_j(x_j)$, focusing on minimizing the residuals left by the previous trees. The model iteratively refines these individual terms using a low learning rate. Once the main effects are established, the model shifts focus to capturing interactions between pairs of features, $f_{jk}(x_j, x_k)$, using a similar boosting procedure. Once the training is completed, each feature (either univariate or bivariate) has its own set of DTs from all boosting iterations, which are then used to construct the corresponding function that models the feature's effect on the response variable. All the DTs could then be discarded. These functions are then summed in an additive manner to form the final EBM model, leaving behind an interpretable model.

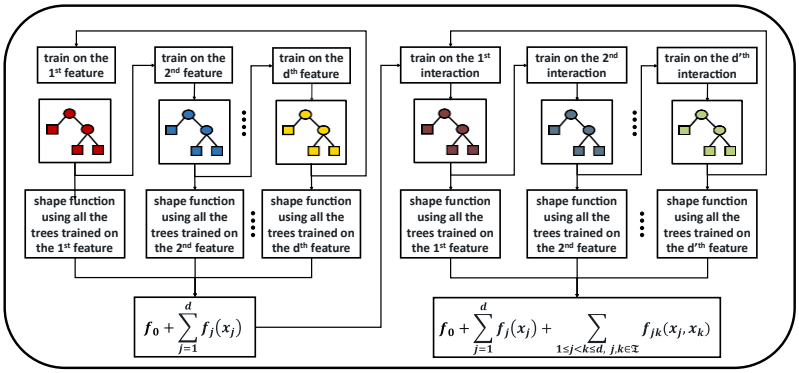

Figure 1. Explainable Boosting Machine's algorithm.

EBM is interpretable because each univariate shape function's relationship with the response variable can be visualized through a plot of $f_j(x_j)$ versus x_j. Moreover, pairwise interactions can be rendered as a heatmap of $f_{jk}(x_j, x_k)$ on the x_j-x_k plane.

Instead of allowing EBM to learn all possible main effects and interactions in a purely data-driven manner, in many SHM applications, we should impose domain-informed constraints, selecting only the terms where engineering knowledge suggests physically meaningful main effects or interactions. When the objective is to predict structural health metrics, such as a health index or damage index, the primary predictors are observed system response features, often obtained from sensors—such as deformation, crack width, or acoustic emission characteristics—which are directly related to the target variable (health metric). Accordingly, univariate shape functions $f_j(x_j)$ are associated with the observed system response features. System parameters, which are defined broadly as inherent properties or contextual conditions of the system, such as material properties, geometry, or environmental and biological factors, moderate how the observed response relates to the health metric and are incorporated through interaction terms, i.e., $f_{jk}(x_j, x_k)$.

CASE STUDY: EBM FOR CRACK WIDTH ASSESSMENT IN RC BEAMS

Current methods in the SHM of reinforced concrete (RC) structures primarily rely on visual inspections and surface crack width measurements. As a use case of EBM for an SHM task, we focus on the prediction of the percentage of the available to ultimate shear load-carrying capacity, referred to here as health index, in RC beams based on the observed maximum diagonal crack width (MDCW). In this study, RC beams with stirrups that fail in shear are investigated. We focus solely on maximum diagonal crack width (MDCW) as the primary damage feature for assessing structural health.

Dataset and Model Specification

The dataset consists of laboratory-tested RC beams, subjected to monotonically increasing loads. The data included in the dataset are obtained from [5–11]. It consists of 620 records associated with 93 beam specimens. Each record captures the MDCW at a specific load level, allowing for multiple observations per beam. Random effects arising from the hierarchical structure of the data were ignored in this study.

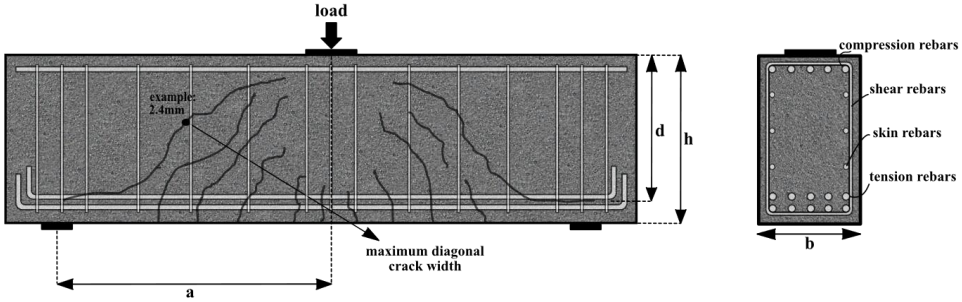

Figure 2. Schematic representation of an RC beam with rectangular section showing the developed crack pattern along with key design parameters.

Input features used in the ML model consist of beam design parameters including effective depth (d), web width to effective depth ratio (b/d), shear span-to-effective depth ratio (a/d), the percentages of tensile (ρ_s), skin (ρ_h), and shear (ρ_v) reinforcements, shear reinforcement yield stress (f_{yv}), and concrete compressive strength (f_c'), along with the observed MDCW (w) (see Figure 2). The response variable is the health index corresponding to the observed MDCW. Because the health index ranges from 0 to 100, a logit link function is employed to map model predictions to the [0, 1] interval during training. Final outputs are scaled by multiplying by 100 to express results as percentages. Moreover, MDCW is the only predictor directly related to the health index, hence, the only univariate shape function $f_j(x_j)$ is associated with MDCW. All beam design features only moderate how observed MDCW relates to the health index. Thus, they are included as pairwise interactions with MDCW, i.e., $f_{jk}(x_j, x_k)$ terms represent the interactions between MDCW and beam design features.

Model Training and Evaluation

A nested 5-fold cross-validation approach is employed to evaluate the performance of EBM on the dataset. In the nested cross-validation, the data is split into training, validation, and test sets. The outer loop creates the test set, while the inner loop splits the remaining data into training and validation sets. Hyperparameter tuning is performed using Bayesian Optimization in the inner loop by training the model on the training set and validating it on the validation set to select the best hyperparameters. The model is then evaluated on the test set in the outer loop for each fold.

Results

After training the model, the model achieved the root mean squared error (RMSE) of 6.13%, 9.99%, 10.40%, the mean absolute error (MAE) of 4.74%, 7.6%, 7.85%, and the correlation of determination (R^2) of 0.92, 0.79, 0.77 for training, validation, and test sets, respectively. The scatter plot of the model predictions versus observed values, for an example outer fold, is shown in Figure 3(a) and the distribution of residual errors for that fold is shown in Figure 3(b).

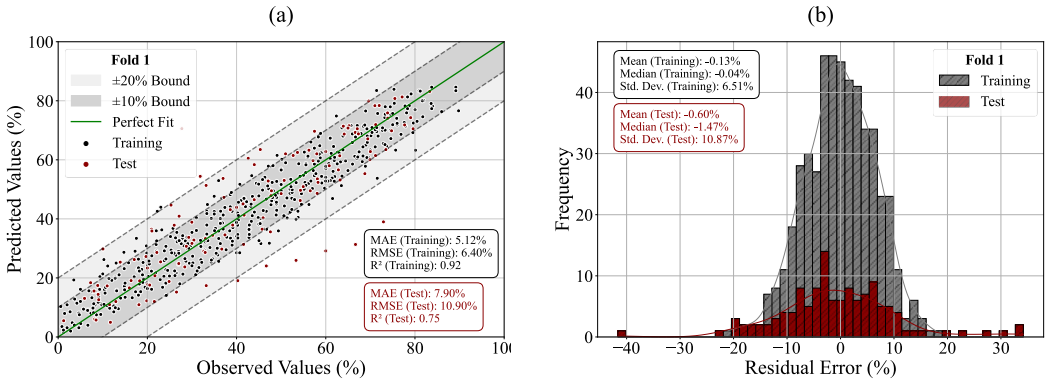

Figure 3. (a) Scatter plot of the model predictions vs observed values, and (b) the distribution of residual errors, for an example outer fold.

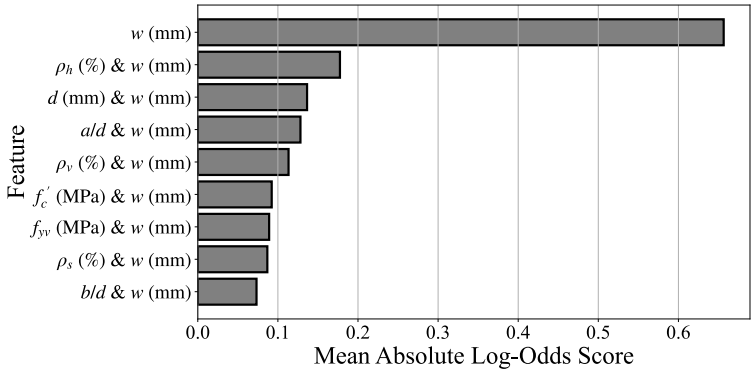

Figure 4. Global feature importance.

Using inherently interpretable models diminishes the reliance on purely accuracy-based metrics when assessing the model's generalizability, as long as the model's learned patterns and decision-making process are thoroughly inspected. One approach to understanding model behavior is through feature importance, which ranks the overall contribution of each feature to the model's prediction based on the mean absolute contribution of each term (Figure 4). Observing the feature importance plot, MDCW is the most important term in predictions, which is consistent with the fact that it is the sole univariate feature in the model directly related to the damage. Following MDCW, the order of importance of moderating variables is coherent with the physical behavior of the system. For example, the reason that the shear reinforcement ratio is not the most important moderating variable in predictions is that the dataset consists only of beams with stirrups, making the skin reinforcement ratio more important than shear reinforcement ratio in the model's predictive process, as the model implicitly recognizes the presence of stirrups based on the dataset it was trained on. One usage of feature importance is that terms with lower mean absolute score can be discarded from the model. During field inspections, the absence of data for less important features should not be a major concern, as their impact on the overall analysis is minimal.

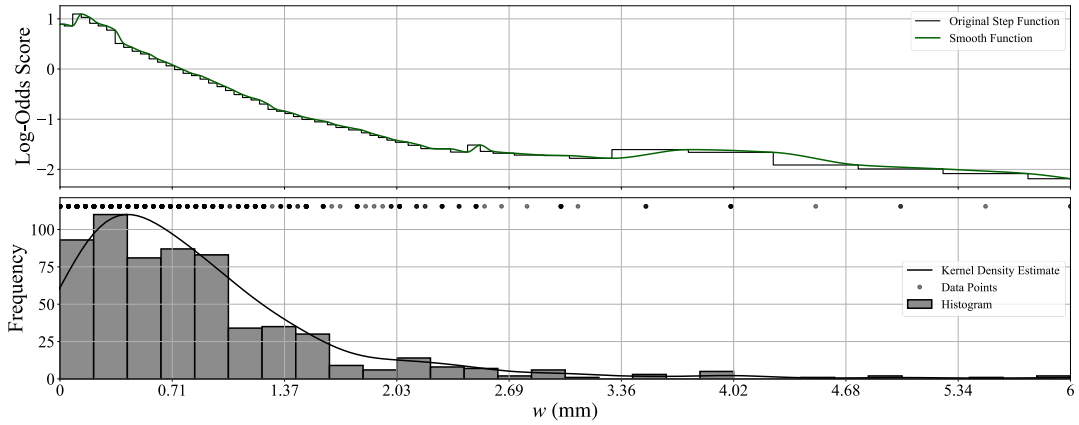

Figure 5. Univariate shape function plot for MDCW (w), and the histogram of MDCW (w).

In addition to feature importance, the functional form of each shape function in the EBM model can be visualized. The top panel in Figure 5 depicts the only univariate shape function in the trained model, MDCW (w), where the y-axis represents the log-odds score, indicating the effect of the MDCW on the model's prediction. The bottom panel shows the histogram of this feature. The decreasing trend observed in the shape function shows larger MDCWs correspond to more damage, leading to a lower predicted health index. Hence, the model captures this trivial trend.

Rendering $f_{jk}(x_j, x_k)$ versus (x_j, x_k) pairs using interaction heatmaps provides insights into how (x_j, x_k) pairs influence the prediction across the pairwise interaction's domain. However, reading heatmaps could be particularly difficult. We propose visualizing the heatmap of damage-moderator variable as cross-sections at different fixed values of the damage feature (MDCW). Figure 6 presents three panels for an example pairwise interaction in the EBM model, the interaction between the effective depth (d) and MDCW (w): the top panel shows the heatmap of the health index versus (w, d) pairs, the middle one displays the cross-sectional plots extracted from the corresponding heatmap at specific values of w, and the bottom one provides the histogram of the corresponding moderator variable (d). Considering a fixed value for the observed MDCW (w), the health index generally tends to increase with increasing effective depth. This indicates that, for two beams exhibiting the same MDCW, the beam with greater effective depth is likely subjected to a lower proportion of its ultimate load—hence, farther from failure. This may be because larger beam depths are typically associated with wider crack spacing, which could result in larger crack widths. Whether this observation aligns with physical principles or is simply the result of data artifacts or bias is something that could be explored further. Similar insights can be drawn from other interaction heatmaps. Each shape function offers a transparent visual interpretation of how feature values influence the model's predictions, enhancing the interpretability of the decision-making process.

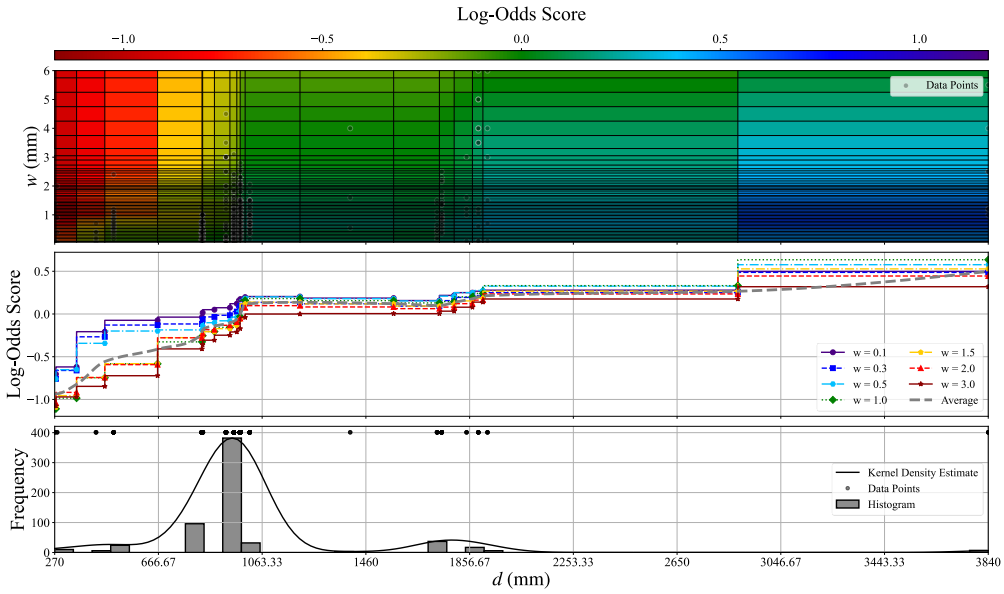

Figure 6. Bivariate shape function for pairwise interaction term between d and w along with the cross-section plots and histograms of the corresponding moderator variable d.

CONCLUSIONS

This study demonstrated the application of an interpretable machine learning approach using Explainable Boosting Machine in the field of SHM. The Case study focused on the damage assessment of shear-reinforced RC beams to assess their available shear load-carrying capacity based on observed surface crack characteristics—specifically the maximum diagonal crack width. Using structural design parameters alongside MDCW as input features, the model offers transparent predictions and insights into how crack behavior relates to structural performance. The ability to assess each feature's importance individually and visualize univariate and pairwise interactions between the features within the model results in its interpretability. This makes Explainable Boosting Machine a powerful tool for building reliable SHM solutions and ensures the model aligns with the underlying physics of the problem.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the Federal Highway Administration (FHWA) of the United States Department of Transportation, which made this research possible.

REFERENCES

1. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, et al. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng 2017;29:2318–31. https://doi.org/10.1109/TKDE.2017.2720168.
2. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2013;Part F128815:623–31. https://doi.org/10.1145/2487575.2487579.
3. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A Unified Framework for Machine Learning Interpretability 2019.
4. Hastie T, Tibshirani R. Generalized Additive Models. Statistical Science 1986;1:297–310. https://doi.org/10.1214/SS/1177013604.
5. Yoshida Y. Shear reinforcement for large lightly reinforced concrete members 2000.
6. Sherwood EG. One-way shear behaviour of large, lightly-reinforced concrete beams and slabs. University of Toronto: 2008.
7. Podgorniak-Stanik BA. The influence of concrete strength, distribution of longitudinal reinforcement, amount of transverse reinforcement and member size on shear strength of reinforced concrete members. 1998.
8. Angelakos D. The influence of concrete strength and longitudinal reinforcement ratio on the shear strength of large-size reinforced concrete beams with, and without, transverse reinforcement. 1999.
9. Larson NA, Fernández Gómez E, Garber DB, Bayrak O, Ghannoum WM. Strength and Serviceability Design of Reinforced Concrete Inverted-T Beams (FHWA/TX-13/0-6416-1). Center for Transportation Research, University of Texas at Austin, Austin, TX.: 2013.
10. Birrcher D, Tuchscherer R, Huizinga M, Bayrak O, Wood SL, Jirsa JO. Strength and Serviceability Design of Reinforced Concrete Deep Beams (FHWA/TX-09/0-5253-1). Center for Transportation Research, University of Texas at Austin, Austin, TX.: 2009.
11. Bracci JM, Hueste MBD, Keating PB. Cracking in RC Bent Caps (FHWA/TX-01/1851-1). Texas A&M Transportation Institute: 2000.