This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference

M.Z. Naser, PhD, PE
Glenn Department of Civil Engineering, Clemson University, Clemson, SC 29634, USA
AI Research Institute for Science and Engineering (AIRISE), Clemson University, Clemson, SC 29634, USA
E-mail: mznaser@clemson.edu, Website: www.mznaser.com

1.0 Abstract

While artificial intelligence (AI), and by extension machine learning (ML), continues to be adopted in parallel engineering disciplines, the integration of AI/ML into the structural engineering domain remains *minutus*. This resistance towards AI and ML primarily stems from two folds: 1) the fact that coding/programming is not a frequent element in structural engineering curricula, and 2) these methods are displayed as blackboxes; the opposite of that often favored by structural engineering education and industry (i.e., testing, empirical analysis, numerical simulation, etc.). Naturally, structural engineers are reluctant to leverage AI/ML during their tenure as such technology is viewed as opaque. In the rare instances of engineers adopting AI/ML, a clear emphasis is displayed towards chasing goodness metrics to imply "viable" inference. However, and just like the notion of correlation does not infer causation, forced goodness is prone to indicate a false sense of inference. To overcome this challenge, this paper advocates for a modern form of AI, one that is humanly explainable; thereby eXplainable Artificial Intelligence (XAI) and interpretable machine learning (IML). Thus, this work dives into the inner workings of a typical analysis to demystify how AI/ML model predictions can be evaluated and interpreted through a collection of agnostic methods (e.g., feature importance, partial dependence plots, feature interactions, SHAP (SHapley Additive exPlanations), and surrogates) via a thorough examination of a case study carried out on a comprehensive database compiled on reinforced concrete (RC) beams strengthened with fiber-reinforced polymer (FRP) composite laminates. In this case study, three algorithms, namely: Extreme Gradient Boosted Trees (ExGBT), Light gradient boosted trees (LGBT), and Keras Deep Neural Networks (KDNN), are applied to predict the maximum moment capacity of FRP-strengthened beams and the propensity of the FRP system to fail under various mechanisms. Finally, a philosophical engineering perspective into future research directions pertaining to this domain is presented and articulated.

Keywords: Explainable artificial intelligence (XAI); Interpretable machine learning (IML); Structural engineering; Concrete; FRP.

2.0 Introduction

Common methods used for the analysis and design of structures are a reflection of our understanding of the physics of structural engineering phenomena and describe such understanding, or causal effects, through symbols fitted into expressions [1]. In a way, such methods are primarily developed as a result of mathematical derivation, empirical or statistical analysis of experimental findings, or numerical simulations that infer the relationship between predictor variable(s) and response variable(s) [2,3]. Despite their origin, these methods share a few

characteristics in common; 1) they are transparent as they comprise of formulae, 2) they establish a procedure with iterative steps, 3) they are universal and mostly region-independent, 4) they have built-in reliability to favor conservativeness, and 5) they are often accepted via a community effort (i.e., voting committees) and tend to be regularly updated and disseminated as codal provisions.

Given the above, structural engineers are trained to appreciate the transparency of commonly adopted methods for analysis and design. And hence, any deviation from this norm is expected to be faced with inertia, which explains why the construction industry is often reluctant to adopt new technologies [4,5]. A prime example of such technology is artificial intelligence (AI) and machine learning (ML). AI/ML capitalizes on novel algorithms to map features governing a phenomenon to the outcome of interest to that particular phenomenon [6]. As one can see, a typical analysis fundamentally comprises of three stages: input observations → mapping → deploy on new data [7–10]. While the first and last stages are easy to follow as they require little effort to visualize, most structural engineers are not well versed with the "mapping" stage despite being tackled by various works over the past years [11–17].

However, a key component within this stage that continues to be vague revolves around answering the following questions, 1) why does a model predict the way it does? 2) what to make out of a typical model's predictions? And, 3) how to trust a model's predictions? These questions are elemental to adopt AI/ML into engineering fields that traditionally favor transparent methods, which most importantly, allow cross-checking, authentication, and contestability [18].

In the realm of computer science, explainability and interpretability can often be used interchangeably [19]. However, there are subtle differences between these two concepts. Simply put, explainability is generally coined for *"models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions,"* while interpretability is *"loosely defined as the science of comprehending what a model did (or might have done)"* [20]. In a way, explainability entails relating the inner mechanisms of a model and their influence upon a model's prediction, while interpretability implies a determination of cause and effect. Ultimately, an explainable model represents a complex function that may not be understood on its own but instead requires additional methods or techniques to be understood. On the other hand, a model can be interpretable if humans can understand it without aid (see Table 1). Other definitions and philosophical arguments concerning explainability, interpretability, transparency, justifiability, and XAI are discussed in Sec. 2.0 and can be found elsewhere [21–24].

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

Table 1 Level of explainability in common models.

Model type	Can humans easily understand this model?	Types of mechanisms used in a model.	Need explainability/interpretably methods?
Linear/Logistic Regression	Yes.	Mathematical based.	Simple/basic models that are inertly explainable.
Tree-like (Decision Tree, Random Forest)	Yes, by displaying a tree formation.	Rule-based that display how the decisions are taken at each step.	Not necessary.
k-Nearest Neighbors (k-NN)	Yes.	Tackles a large number of variables via mathematical/statistical representations.	Mostly visually explainable.
General Additive Models (GAM)	Mostly, and especially if interactions between features are smoothened.	Needs mathematical/statistical representation.	No, unless interactions turn complex.
Genetic Algorithms (GA)	Yes, since they resemble a tree-like structure.		Mostly readable.
Tree Ensembles (ExGBT, LGBT)	Unlikely.	Blenders of weaker models.	Very complex models that require further analysis via agnostic or specific methods.
Support Vector Machines (SVM)		Categorize data via hyperplanes in N-dimensional space.	
Neural Network (NNs)		Layers containing neurons and transformation functions	

A closer look into existing literature shows that publications utilizing AI/ML continue to rise in this domain and are expected to continue to do so [25]. This implies that the structural engineering community is interested in this technology. Building upon trends in parallel engineering fields extrapolated to a few years from now, AI-based methods will comprise a considerable portion of the developments within this area. As such, facilitating the integration of AI/ML is of our utmost importance [26]. From this perspective, a survey of literature shows that AI/ML has been used in a collection of structural engineering problems [27–31]. At the writing of this work, these technologies have been applied to explore properties of construction materials [27,28,32], prediction of structural responses of elements such as walls [33], bridges [34], beams [14], among others [35–37]. In addition, AI/ML has also been used to examine the quality of constructions and inspections [38–44].

A deep dive into the open literature shows that the bulk of AI/ML works applied these techniques to: 1) tie a group of features to a needed outcome (e.g., link concrete mix ingredients to expected compressive strength, etc.), and 2) be primarily designed to attain a certain level of goodness; thereby declared to having a permissible prediction capability, and generalizability. However, little is ever mentioned to that of how or why a model predicts the way it does. It is the view of this author that satisfying a selected performance metrics does not effectively imply that an AI/ML model captures the physics behind a phenomenon but rather infers the suitability of such model to predict the outcome described within the examined database, which is assumed to present a holistic view into the physics behind the phenomenon being investigated [45]. In other words, correlation does not always suggest causation, for which causation requires a display of a deep level of understanding that may go beyond satisfying selected measures of goodness [46].

From this perspective, understanding why a model predicts the way it does can open up new and exciting opportunities to structural engineers. For instance, and going back to our earlier example of tying concrete mix ingredients to the expected compressive strength, developing an AI/ML model that can truly capture this phenomenon might, in fact, be beneficial as it may infer hidden relation(s) between the examined ingredients (i.e., features) that are new to us. This concept, once extended beyond this particular example, can help engineers discover new mechanics to some of their common problems and may indeed open up new solutions to existing or long-lasting problems [47].

The above brings in the important question of how to infer a model's understanding of a phenomenon? Or better yet, of an engineering phenomenon? Despite XAI/IML being a relatively new research area, a few methods exist to explore the reasoning behind an AI/ML model predictions [48,49]. These methods vary in terms of *nature* (e.g., intrinsic [simple models i.e., trees] vs. post hoc [complex models i.e., deep networks]), *type* (model-agnostic [i.e. applied to any AI/ML model] vs. model-specific [i.e. particular to family of models]), and *scale* (local [i.e. explain individual predictions] vs. global [i.e. explain model behavior as a whole]) [50].

Of interest to this work are methods that have been usually grouped under "model agnostic" and "model specific". On one hand, model agnostic methods can be used across algorithms and platforms (which provides the user (a structural engineer) with an advantage of comparing different models following a consistent approach), while model-specific methods are only applicable to a specific AI/ML model due to the nature behind such methods' developments [51,52]. Model agnostic methods include feature importance, partial dependence plots, feature interactions, SHAP (SHapley Additive exPlanations), surrogates, and local interpretable model-agnostic explanations (LIME). On the other hand, model-specific methods include: generalized linear models and tree-based ensemble models [50,53,54].

This work advocates for implementing a humanly plausible (i.e., explainable) form of AI; thereby eXplainable Artificial Intelligence (XAI), and by extension, Interpretable Machine Learning (IML). To provide the reader with an arsenal to demystify how model predictions can be

interpreted, this paper examines a collection of agnostic methods (e.g., partial dependence plots, feature importance, feature interactions, SHAP, and surrogates), together with commonly accepted performance metrics. A case study on reinforced concrete (RC) beams strengthened with fiber-reinforced polymer (FRP) composite laminates is selected for demonstration. In this case study, the Extreme Gradient Boosted Trees (ExGBT), Light gradient boosted trees (LGBT), and Keras Deep Neural Networks (KDNN) are applied to predict the maximum moment capacity of FRP-strengthened beams and the propensity of the FRP system to fail under a variety of modes (e.g., debonding, steel yielding, concrete cover separation, mixed-mode, and FRP fracture).

3.0 A Philosophical Engineering Perspective into XAI and IML

Every profession is molded by its people. As such, structural engineers get to shape how their domain transforms in the coming years. It is only natural for a domain to be influenced by advances occurring in parallel domains. Since the design of structures is an involved process that requires interaction with engineers from other fields (presumably those who are leveraging AI/ML at the moment (e.g., mechanical engineers, etc.)), then it is only a matter of time before such interactions bring in the notion of AI/ML into our domain. Noting the tremendous improvements (work quality, revenue generation, etc.) such technologies are attributing to other fields, then perhaps we ought to consider adopting AI/ML into ours.

In this pursuit, the author hopes that we do not have to go through the same cycles of integration other domains have gone through, thereby avoiding the pain of trials and errors associated with re-inventing the wheel. A more thought of integration of AI/ML that learns from experiences in other fields is expected to ensure safe and attractive integration of these technologies into our domain. This work sheds some light on the future of XAI and IML in structural engineering. A deep dive into the big ideas behind XAI and IML shows that they fall in line with principles relevant to the structural engineering practice and industry, much more so than that of traditional AI and ML. However, such ideas are not properly articulated from an engineering perspective, and hence these are conveyed herein.

Engineers need to ***trust*** their methods and tools – especially those to be deployed in real scenarios (which in our domain involves the lives and well beings of occupants and the surrounding environment, etc.). Since XAI and IML strive to deliver a high level of trust between engineers and AI/ML, then these technologies do provide an improvement over those of traditional AI-nature. Still, one must be cautious of the degree of explainability provided by a model, as the degree of "required" explainability can be different in practical scenarios (i.e., design of a high-rise building vs. design of a storage unit), as well as to that anticipated by the practicing engineer/stakeholder/building official [55]. In other words, explainability is fluid and is highly context-dependent. This particular idea of establishing a "required" degree of explainability mirrors the rules of thumbs practiced in structural design. For example, an engineer trusts that following codal provisions will lead to a favorable performance under stressful conditions.

Building on the above, the notion of trusting AI is complicated since it requires to first define what trust is to structural engineers? As trust can be subjective, this notion does bring the idea of

transparency [56]. This concept implies that an engineer has some understanding of the means by which an AI/ML model operates (e.g., a decision tree model splits the examined dataset into a tree-like format, etc.) in a similar manner to her/his understanding of how local buckling checks examine the propensity of a structural shape to buckle.

In addition, a transparent model can also be thought of as a favorite for stakeholders since they can easily visualize its working mechanisms. Besides, the ability to easily visualize the works of a model makes it easy for those of limited technical background to understand AI/ML predictions, or at the very least see why a model predicts the way it does. A parallel to trust and transparency is *confidence*. In engineering terms, attaining confidence can be achieved by developing reliable models; those with the capability to disclose a quantifiable level of confidence in line with their predictions (e.g., the expected shear capacity of a W-shaped steel beam with specific properties and under a confidence level of 95% is 100 kN ± 5 kN and).

One of the fundamental premises for AI/ML is that it can provide us with data-driven valuable insights that existing methods fail to provide or may not realize due to assumptions used in deriving such methods or simply due to their limited extrapolatibity. A structural engineer is primarily interested in identifying causes and effects (say, how does a load-bearing configuration resists a particular load condition). While traditional analysis and design methods for common elements (e.g., beams, columns, frames) can capture such effects with high accuracy, one is reminded that such methods took decades of experimentation to arrive at this level of accurately capturing the aforenoted *causation* [57]. Both XAI and IML have the potential to realize such a degree of inference, not only for simple elements but also for more complex designs, and hence these are of merit to structural engineers. If the stars align, XAI and IML can at the very least elucidate areas of high merit for exploration between features pertaining to realizing true causation within the realm of the structural design.

Oftentimes, data is limited [58]. That does not imply that we do not have data from years of engineering practice, but rather the data we have access to and can use to train AI/ML models is limited. This brings in a limitation not only to train models but also to train models capable of proper *generalization* beyond their training data points. As such, it will be foolish to solely rely on such a model as it may lead to serious consequences. An XAI/IML can be helpful in which it could be designed to provide examples as to why it made its predictions (e.g., in a given connection design that falls beyond the range of codal provisions, a model may layout a preliminary connection type with an estimated number of bolts, etc. while attributing its decision to an analogous connection of similar properties to the one on hand but used at a sister project). In this instance, while the model may not have arrived at a "prediction" per se, it did, however, direct us to a sort of solution that can jump-start our design. This is often referred to as *analog informativeness* or simply explainability by examples.

A look into engineers' role in society shows that one of their primary goals is to address social problems [59]. From a structural engineering perspective, this goal can be achieved by creating safe, suitable, and affordable structures for all. In most settings, certain decisions can be undertaken without regard to social status since, at the end of the day, a structural design is to satisfy an array

of conditions. Applying an AI/ML model to such scenarios is expected to be naturally straightforward ("math is math, and physics is physics"). However, in instances where stakeholders belong to weaker/forgotten sections of our society, then in addition to satisfying codal design provisions, we must also adhere to additional "social attributes" to realize socially just structures. XAI and IML can be integrated with *fairness* and *inclusivity* to identify affordable alternatives to new constructions, eco-friendly building materials, etc. The same can also generate designs that leverage novel initiates such as green constructions, adaptability, and circular economy [60,61].

The reader should note that the philosophical discussion on XAI and IML grows beyond the scope of this work to that which may cover autonomy, biasness, accessibility, interactivity, privacy, the trade-off between accuracy and explainability, and others [55,62].

4.0 Model Agnostic Methods for Explainability and Interpretability

This section briefly describes five commonly used model agnostic techniques in detail. The methods described herein can be used to explain a model's predictions on the global (all dataset) or local (a portion of the dataset) levels. These methods are applied to post the development of a model and completion of an AI/ML analysis and can be plugged into any AI/ML model – hence becoming useful on a larger scale. The reader is to note that information regarding the history, mathematical derivation, and the background of each technique can be found in their perspective references and in [48]. All these techniques will be examined via the presented case study in a later section.

4.1 Feature Importance

A typical AI/ML model comprises multiple features, wherein each feature makes a unique and likely quantifiable contribution towards the response (prediction) of such a model. As a result, a model can be interpreted by understanding the influence of each of its own features. Herein where feature importance comes in handy. Feature importance is a generic term for the degree to which an AI/ML model relies on a particular feature in its prediction, and hence this method measures the extent to which a given feature influences the outcome of an AI/ML model. This can be measured by evaluating the increase of a model's prediction error after permuting the involved features systematically [63]. In a way, the concept behind feature importance is to measure the entropy in the change of predictions, given a perturbation of a given feature. In this process, a feature is declared "important" if permuting its values increases the model error, effectively indicating that the model relied on this particular feature to arrive at a good prediction. Similarly, a feature is deemed "not important" if its permutation maintains the prediction error unchanged.

4.2 Partial Dependence Plots

The partial dependence plot (PDP) depicts an individual feature's marginal effect, or group of features, on the prediction of an AI/ML model while holding other features constant within the same model [64]. Such a plot can also be used to infer the type of relationship between the feature(s) and response(s), e.g., linear, nonlinear, or complex. Overall, the PDP accounts for all observations in a database to give inference to the global relationship of a feature on the predicted

outcome. A PDP helps determine the transition in a model's predictive performance to the change in the feature(s). The result of partial dependence describes the impact of a feature on a model's prediction (similar to how a coefficient reflects weight in a regression model). Unlike in regression, when a PDP is applied to a classification problem, this method gives the probability for a particular class given different values for feature(s) in a database. One should note a fundamental assumption in PDP relates to feature(s) of interest not being correlated with other features. An extension to PDP is known as Individual Conditional Expectation (ICE) plot [65]. An ICE plot displays how a prediction changes when a feature changes. Unlike a PDP, ICE does not average the relationship between features and predicted responses.

4.3 Feature Interaction

Feature interaction involves two or more features influencing each other (or interacting) during an AI/ML analysis. It is due to the existence of such interaction that the overall prediction performance of an AI/ML model is not equal to the simple sum of all features. For example, in a given AI/ML model, two features are said to interact when the effect of one feature on the response of the model is not constant but also depends on the value of the second feature [66]. In a simple AI/ML model that has two features (*X1* and *X2*), a given prediction from such a model can be broken down into four component terms; a constant term, a term for the first feature, a term for the second feature, and a fourth term that contains an interaction between the two original features [67]. Feature interactions can be evaluated via methods such as correlation matrix[1], association matrix, Cramer's Phi, or heat maps [68]. This method can also be used to select features during the processing stage.

4.4 SHAP (SHapley Additive exPlanations)

The SHAP method is considered as a unified agnostic method that can be applied to explain individual responses (i.e., output/predictions) of any AI/ML model [52,69]. SHAP is based on a game theory approach to additively accumulate the contribution of all features involved in a model. As such, this method assigns each feature an "importance value" within a set of conditional expectations for a particular prediction. The results of this additive procedure are called "SHAP values". These values can be spread across from a "base value" (which represents the average of the observations). Then, the SHAP method would graphically list the contribution of all features to the SHAP value, identifying which feature(s) contributed to such value being larger or smaller than the "base value" in order. As one can see, since the SHAP method accounts for all features and randomness of their order, this method can be quite computationally expensive for large models, yet necessary to understand the logic behind AI/ML models.

[1] For example, the Pearson correlation matrix is a visual aid in the form of a table that lists the "linear" correlation coefficients between features. Such matrix summarizes the degree of correlation between features and can be used to explain the selection of highly independent features and deselection of dependent features or those of limited impact. Still, one of the limitations of such matrix is that is relies on "linear" correlation between variables which may not be of high merit if the relationship between variable is of a complex nature – as such the use of other feature selection methods become more beneficial.

4.5 Surrogates

In this method, an external secondary AI/ML model is used to proxy (or simply explain) the prediction of an existing model. The external model is often simple in nature, or one that we have a clear understating of its inner mechanisms (e.g., a linear model); thereby, and once augmented into the complex model, the surrogate may be used to give us insights into why the complex model predicts the way it does [70]. In practice, the choice of a surrogate model is decoupled from the complex model; as technically, all that a user needs is a model that is simple in nature and can capture the predictions of the original model with good accuracy (such as additive models, genetic algorithms, tree-like models [71–73]). Note that accuracy, in this context, could refer to attaining an adequate performance metric (e.g., coefficient of determination).

5.0 Description of Database

This section describes the examined database to be used in this work as a case study.

5.1 Database on Reinforced Concrete (RC) Beams Strengthened with Fiber Reinforced Polymer (FRP) Composite Laminates

The adopted database herein was compiled in a companion work [74] and collected comprehensive data from 103 experimental tests carried out on RC beams strengthened with FRP composites. This database includes full information on compressive strength of concrete (f_c), yield strength of steel reinforcement (f_y), ratio of steel reinforcement (r_s), FRP ratio (r_f), modulus of FRP (E_{fr}), moment arm (a), and strengthening type (T). All features were collected from the following works [75–92] and are compiled into this database. The outcome/response of this database resembles the magnitude of moment capacity, and failure mode observed in each corresponding test (i.e., steel yielding/concrete crushing, debonding, concrete cover separation, mixed-mode, and FRP rupture). In a way, this database can be used in a regression problem (to predict the magnitude of moment capacity at failure), or in binary- (failure through FRP system vs. steel yielding), or multi-classification problem (failure due to a collection of causes such as debonding, steel yielding, concrete cover separation, mixed-mode, and FRP fracture).

A deep examination of Table 2 shows that the compiled database and range of features used cover practical scenarios in which FRP-strengthened members are often used and agree with that identified in design building codes [93–96]. It can then be inferred that the developed database represents scenarios a structural engineer can face in practice and hence can be used with confidence. Table 2 and Fig. 1 list the selected features and their ranges. One should note that 42 beams failed via steel yielding, and 59 beams failed due to FRP system failure (debonding: 18 beams, concrete cover separation: 18 beams, mixed-mode: 16 beams, and FRP rupture: 7 beams).

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

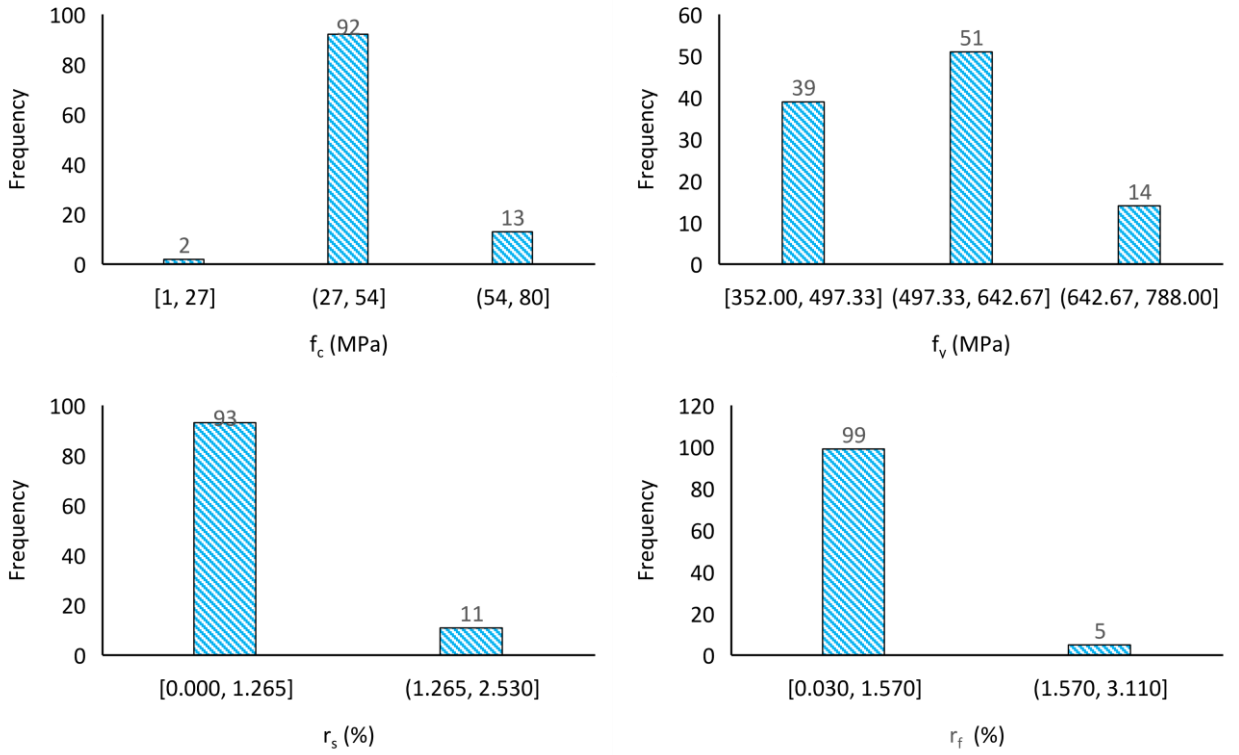
Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction.* https://doi.org/10.1016/j.autcon.2021.103821

322 Table 2 Sample database for FRP-reinforced concrete beams for moment capacity and failure
323 mode identification

	Inputs							**Output**	
	Compressive strength of concrete (f_c) – MPa	Yield strength of steel reinforcement (f_y) – MPa	Ratio of steel reinforcement (r_s) – %	Modulus of FRP rebars (E_{fr}) – GPa	FRP plate/sheet/NSM ratio (r_f) – %	Moment arm (a) – mm	Strengthening type (T) – 1 for plate/sheet, 2 for Near-surface mounted (NSM)	Moment capacity (M) – kN.m	Failure mode (F) – steel yielding, debonding, concrete cover separation, mixed-mode, FRP rupture
Range of features	31-80	352-788	0.01-2.5	0.03-3.1	11-240	300-1830	1, 2	-	-

324



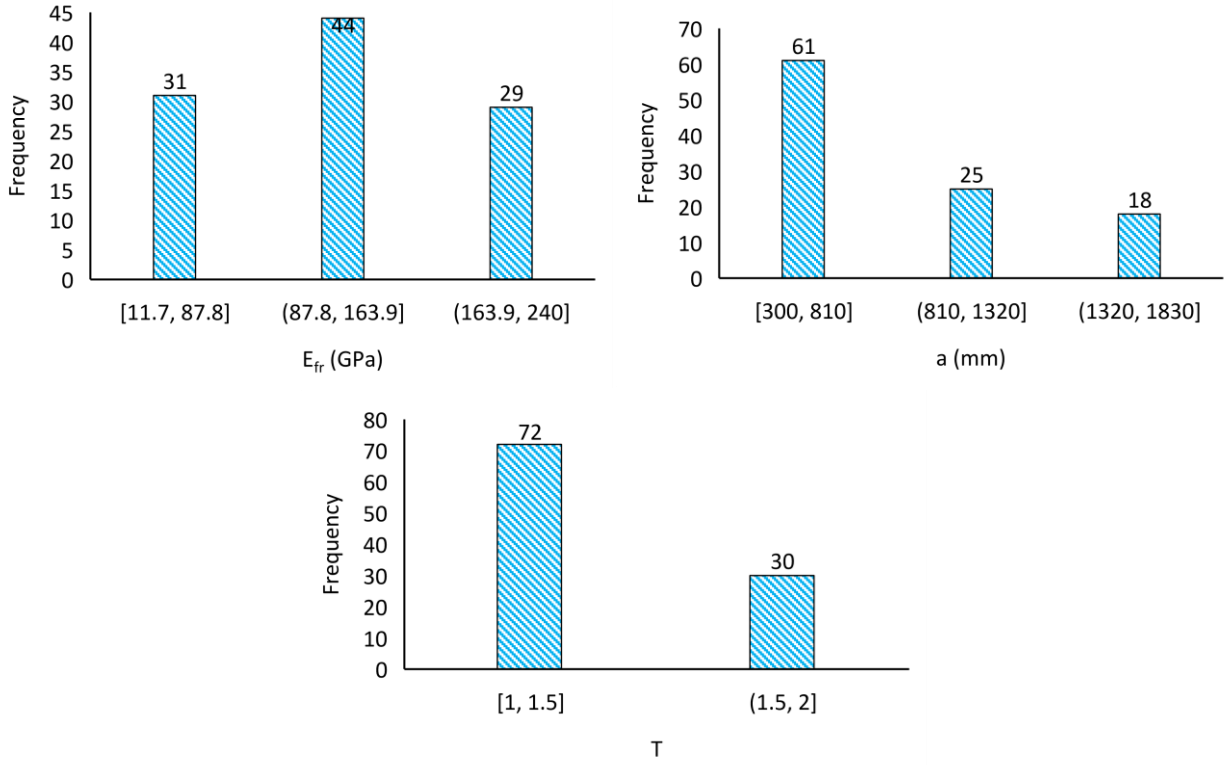


Fig. 1 Frequency of identified features in the compiled database

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

Table 3 shows further statistical insights into the compiled database. For example, the minimum and maximum compressive and yield strength of the compiled beams range between 31-80 MPa and 352 and 788 MPa, respectively. The steel and FRP ratios have a maximum range of 2.5% and 3.1%, respectively, typical of that used in RC beams design. The modulus of FRP ranges from low stiffness (11.7 GPa) to high stiffness (240 GPa). The same table also shows a sensitivity analysis to identify the correlation between all features compiled in this database. The outcome of this analysis shows that moment arm has a strong positive correlation with the moment at failure and also shows that beams' material properties and geometric features seem to have a medium correlation with the observed moment at failure.

Table 3 Key statistics.

Section	*Features*	f_c *(mm)*	f_y *(mm)*	r_s *(%)*	r_f *(%)*	E_{fr} *(GPa)*	*a (mm)*	*T*	*M (kN.m)*
FRP-strengthened RC beams	Min	31.0	352.0	0.0	0.0	11.7	300.0	1.0	6.0
	Max	80.0	788.0	2.5	3.1	240.0	1830.0	2.0	585.0
	Average	43.0	530.6	0.7	0.4	129.0	881.3	1.3	119.9
	Standard deviation	8.9	116.5	0.4	0.5	65.6	444.6	0.5	142.7
	Median	0.7	0.8	1.5	3.0	-0.2	1.1	0.9	2.0
	Skewness	31.0	352.0	0.0	0.0	11.7	300.0	1.0	6.0

Parameter	f_c	f_y	r_s	r_f	E_{fr}	*a*	*T*	*M*
f_c	1.000							
f_y	0.315	1.000						
r_s	-0.007	-0.399	1.000					
r_f	-0.102	-0.256	0.039	1.000				
E_{fr}	0.180	0.277	-0.092	**-0.550**	1.000			
a	-0.489	-0.596	0.396	0.139	-0.174	1.000		
T	-0.181	0.230	-0.274	-0.291	0.052	-0.082	1.000	
M	-0.471	-0.457	0.430	0.041	-0.083	**0.907**	-0.169	1.000

6.0 Selected Machine Learning Algorithms

As mentioned earlier, the primary goal of this work is to showcase the application of the AI/ML explainable methods described in an earlier section. In this pursuit, three algorithms are selected for showcasing the applicability of these methods, namely, Extreme Gradient Boosted Trees (ExGBT), Light Gradient Boosted Trees (LGBT), and Keras Deep Residual Neural Network (KDNN)), and these are briefly discussed herein with the full description being found in their respective references, as well as in [97,98].

6.1 Extreme Gradient Boosted Trees (ExGBT)

The ExGBT algorithm re-samples the collected data points into a tree-like format, where each tree sees a boostrap portion (a sampled dataset with replacement) of the database in each iteration [99].

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

ExGBT fits each successive tree to previous residual errors obtained from previous trees, thereby focusing the prediction effort of each iteration on the most challenging responses to predict, which becomes a good practice for the algorithm to yield high prediction accuracy (see Eq. 1). The ExGBT brings two techniques to improve the performance of a GBT; a weighted quantile sketch (an approximation algorithm for determining how to make splits candidate in a tree) and the sparsity-aware split finding (which works on sparse data, as well as data with missing values). The ExGBT uses a pre-sorted algorithm and a histogram-based algorithm for computing the best split [54].

$$Y = \sum_{k=1}^{M} f_k(x_i)\,, f_k \in F = \left\{ f_x = w_{q(x)}, q: R^p \to T, w \in R^T \right\} \quad (1)$$

Where, M is additive functions, T is the number of leaves in the tree, w is a leaf weights vector, w_i is a score on i-th leaf, and $q(x)$ represents the structure of each tree that maps an observation to the corresponding leaf index [100]. The code of the used ExGBT can be found online at [101,102]. This algorithm incorporates the following pre-tuned settings of learning rate of 0.015, maximum tree depth of 3, subsample feature of 0.8, and 500 for the number of boosting stages.

6.2 Light Gradient Boosted Trees (LGBT)

Light gradient boosted trees is an algorithm that requires little processing and resembles that of the random forest algorithm (which contains a series of tree-like elements) [103]. The LGBT successively fits the trees and fits the residual errors from all the previous trees combined [104]. This is advantageous, as the model focuses each iteration on the most challenging examples to predict. Similar to the ExGBT, the LGBM algorithm introduces two new techniques to further improve its the performance. These techniques are gradient-based one-side sampling (which identifies the most informative observations and skips those less informative), and exclusive feature bundling (which groups features in a near-lossless way) [105]. The used algorithm can be found at [106] with the following default settings: learning rate = 0.05, maximum depth = "none", number of boosting stages = 500.

6.3 Keras Deep Residual Neural Network (KDNN)

Keras is a high-level library for developing neural networks [107]. In a residual network, a direct connection exists linking data points to the outputs. Such a connection smoothens out the loss function and enables better optimization of the network. In the used KDNN, default settings of a learning rate of 0.03 was used, along with a *Prelu* activation function, one layer containing 64 neurons. KDNN can be readily found at [108].

7.0 Model Performance

A proper AI/ML analysis aims to minimize flaws within selected models. This is often handled by randomly shuffling and splitting the database into three sets (T: training, V: validation, and S:

testing). A model is then trained and validated on the first two sets, respectively, and then independently checked against the last set (since it was not involved in the training and validation procedure). In all cases, 10-fold cross-validation was also employed.

In addition to the above, and in order to verify the adequacy of the selected models, model predictions were first cross-checked against performance metrics. Such metrics pertain to mathematical constructs intended to measure test measurements' closeness to that predicted by a model [109–111]. In this work, metrics from two domains, regression, and classification, are selected (see Table 4). These listed metrics are frequently used within the structural engineering domain, among others [7,12,112,113]. The regression metrics include Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Briefly, MAPE measures the error between continuous variables as a percentage, while RMSE measures the standard deviation of residuals and describes the errors in a scale-independent order. R^2 is a scale-free score that measures the degree of association between observed and predicted values.

For classification, three metrics are also presented, including; Balanced accuracy (BACC), Area under the ROC curve (AUC), and Log Loss Error (LLE). The BACC is useful in scenarios involving imbalanced features and multi-classes. The AUC measures the area under the Receiver Operating Characteristic (ROC) curve, with a value of unity indicating an accurate prediction. The LLE measures a classification model's performance whose output is a probability value between 0 and 1, with values approaching zero inferring perfect performance.

As one can see from the results listed in Table 4 and Fig. 2, the ExGBT algorithm seems to outperform that of the LGBT and KDNN in most comparisons and for all selected performance metrics. This implies that this algorithm did indeed capture the two examined phenomena provided by the database (i.e., regression to predict moment capacity at failure and classification to predict mode of failure). As such, only the ExGBT is augmented with methods of explainability to understand its reasoning behind its predictions better.

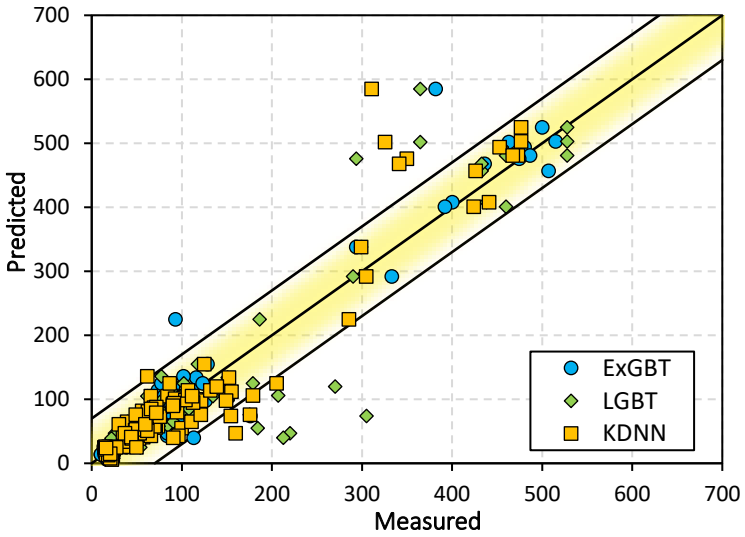


Fig. 2 Comparison between ExGBT, LGBT, and KDNN

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

Table 4 List of selected performance metrics.

Problem	Name	Metric	ExGBT			LGBT			KDNN		
			T	V	S	T	V	S	T	V	S
Regression	Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\lvert E_i\rvert/\lvert A_i\rvert$	**35.29**	**21.79**	27.56	54.03	38.6	50.59	51.45	38.56	**26.97**
Regression	Root Mean Squared Error (RMSE)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} {E_i}^2}{n}}$	**21.35**	**26.16**	41.19	49.30	40.86	67.17	33.39	49.50	**39.41**
Regression	Coefficient of Determination (R^2)	$R^2 = 1 - \sum_{i=1}^{n}(P_i - A_i)^2 / \sum_{i=1}^{n}(A_i - A_{mean})^2$	**0.98**	**0.96**	**0.91**	0.91	0.95	**0.91**	0.96	0.85	0.90
Classification (Top: Binary, Bottom: Multiclass)	Balanced accuracy (BACC)	$BACC = \frac{1}{M}\sum_{m=1}^{M}\frac{r_m}{n_m}$ *Where, M = number of classes, $n_{m=}$ data size belongs to class m, $r_{m=}$number of data accurately predicted belonging to class m.*	**0.68**	**0.44**	0.52	0.48	0.36	**0.61**	0.42	0.37	0.49
			0.84	**0.85**	0.49	0.61	0.72	**0.5**	0.51	0.57	0.53
	Area under the ROC curve (AUC)	$AUC = \sum_{i=1}^{N-1}\frac{1}{2}(FP_{i+1} - FP_i)\,(TP_{i+1} - TP_i)$	**0.83**	**0.80**	0.84	0.81	0.75	**0.87**	0.81	0.78	0.70
			0.92	**0.93**	0.75	0.81	0.86	0.75	0.75	0.78	**0.76**
	Log Loss Error (LLE)	$LLE = -\sum_{c=1}^{M} A_i log P,$ *where, M: number of classes, c: class label, y: binary indicator (0 or 1) if c is the correct classification for a given observation.*	**1.03**	**1.09**	**1.11**	1.21	1.25	1.13	1.28	1.28	1.81
			0.42	**0.36**	0.61	0.55	0.49	**0.53**	0.52	0.49	0.56

A: actual measurements, *P*: predictions, *n*: number of data points, *E = A-P, P (denotes the number of real positives), N (denotes the number of real negatives), TP (denotes true positives), and FP (denotes false positives).*

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

8.0 Evaluation of Explainability and Interpretability

In this section, the five previously described model agnostic methods are applied to the ExGBT model.

8.1 Feature Importance

Feature importance showcases the degree to which an AI/ML model relies on a particular feature and hence measures the extent to which a given feature influences the outcome of an AI/ML model. In the developed regression model, the feature importance of moment arm (a) was shown to be dominant, followed by the ratio of steel reinforcement (16.67%), FRP ratio (10.62%), yield strength of steel reinforcement (10.25%), modulus of FRP (5.76%), compressive strength of concrete (5.06%), and strengthening type (2.35%).

On the binary classification front, the importance of features varies from that observed on the regression front. In this instance, the dominant feature is the ratio of steel reinforcement, followed by moment arm (54.47%), yield strength of steel reinforcement (41.77%), modulus of FRP (25.91%), compressive strength of concrete (24.14%), strengthening type (15.38%), and FRP ratio (13.57%). Then, on the multi-classification front, the dominant feature is the ratio of steel reinforcement, followed by the yield strength of steel reinforcement (78.81%), FRP ratio (52.24%), moment arm (38.77%), compressive strength of concrete (34.66%), modulus of FRP (28.61%), and strengthening type (17.39%). This analysis shows that despite using the same database and algorithm, the type of analysis can affect the importance distribution of selected features involved in the database.

Figure 3 shows a graphical distribution of the importance of all features in the above three analyses. The same figure also shows the considerable variation in feature importance between moment prediction (i.e., regression problem) and failure prediction (classification problem). Simply put, the reasoning behind ExGBT prediction of the moment capacity of a given beam is primarily governed by the length of the moment arm. Similarly, a model's prediction of classifying failure of beams is influenced by the degree of reinforcement ratio, followed by the yield strength of reinforcement and FRP level of strengthening. Besides, a close examination of this figure shows a general agreement between the importance of features in the case of binary and multi-classification of the failure phenomenon in FRP-strengthened RC beams, with the exception of yield strength of steel reinforcement and FRP ratio.

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

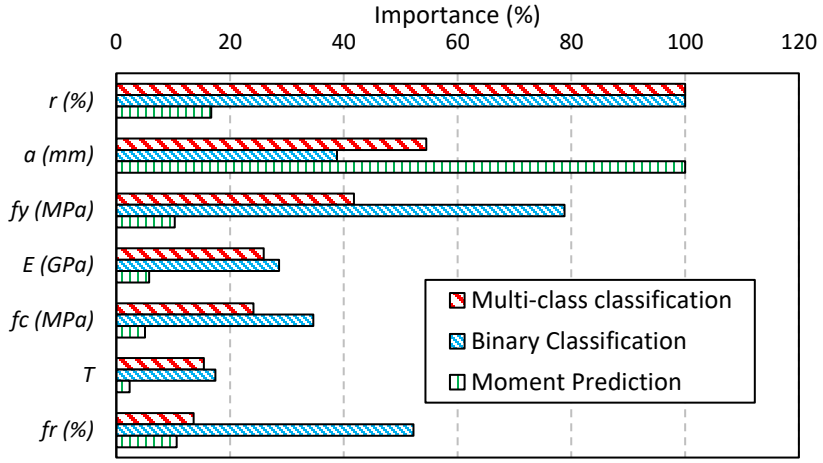


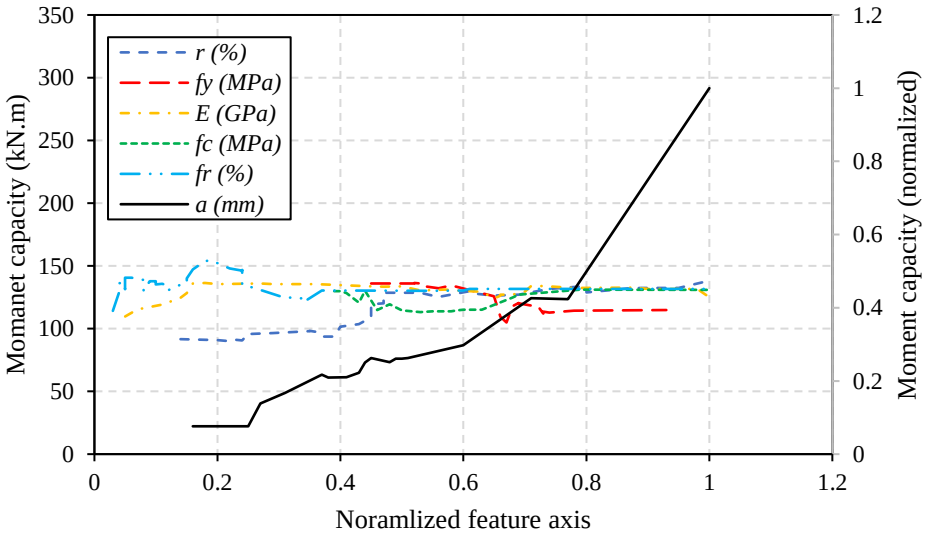
Fig. 3 Explainability through feature importance values

The reader is reminded that this analysis does not aim to justify the model's predictions or to declare validity, but rather this analysis aims to explain why a model is behaving the way it does. From a structural engineering perspective, and for identical beams, a larger moment arm is expected to generate a high moment magnitude (in a simply supported configuration), thereby significantly influencing the attained moment. Furthermore, in classifying a failure mechanism, a balance between steel and FRP materials primarily dictates the expected failure, as noted by a number of researchers [79,114]. From a practical view, if an engineer aims to identify important features in a given problem, then perhaps a good practice is to compare feature importance results across a collection of parallel models to identify features highly ranked by different models.

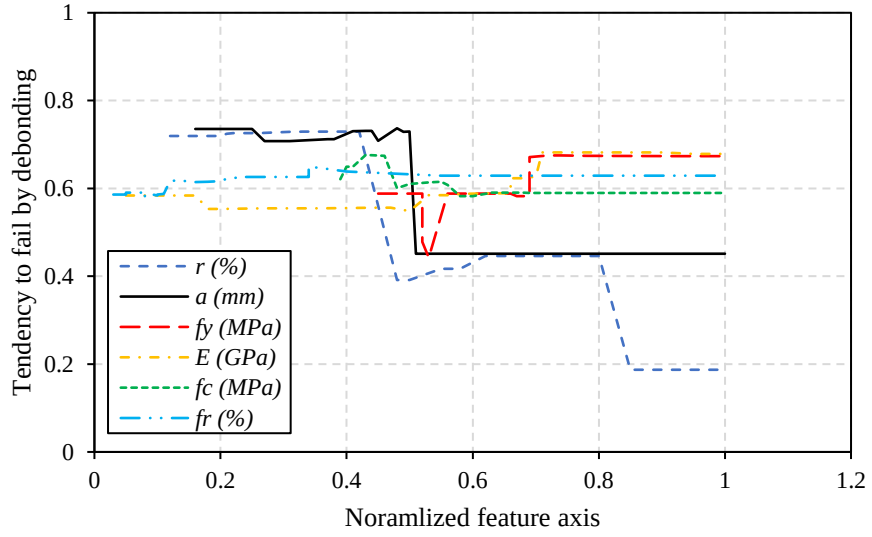
8.2 Partial Dependence Plots

As discussed earlier, a PDP depicts the marginal effect of an individual feature, or group of features, on the prediction of an AI/ML model while holding other features constant within the same model [64]. Thus, Fig. 3 shows PDPs for all features used in the moment capacity prediction analysis, as well as the binary prediction of failure (as developing PDPs for multi-classification problems can be cumbersome, yet can be found elsewhere [115]).

In Fig. 4a, and with regard to moment prediction, one can see that the dominance of the moment arm feature towards the prediction of a larger moment at failure exceeds all other features for moments exceeding 135 kN.m. In other words, larger moment arms of 1280 mm (computed as 0.70×1830 mm) tend to generate higher moments, and hence the model's reasoning for tying this feature with larger values of moments at failure. Furthermore, smaller moment arms of less than 900 mm do not seem to affect model prediction as compared to the other features significantly. Additional insights into the impact of each of these features on the increase in a moment (when all other features remain constant) can also be drawn. For example, this figure shows how all other features seem to slightly influence the model's decision-making process, which meshes with that observed in the low degree of feature importance these features display.



(a) Moment capacity



(b) Binary classification

Fig. 4 Explainability through partial dependence plots (PDP) plots

On a similar note, Fig. 4b plots the variation in partial dependence for features used in the classification AI/ML analysis. This figure shows that lesser values of moment arm, ratio of steel reinforcement, and steel yield strength are directly related to an increase of failure through FRP debonding. This also agrees with experimental observations wherein shorter moment arms tend to intensify shear effects, and smaller reinforcement amplifies the utilization of FRP systems under loading, making it vulnerable to debonding [116]. Also, a clear transition phase at around 40-60% of the maximum values for moment arm (1830 mm) and steel reinforcement ratio (2.5%) is

apparent. This figure can become a crucial element to minimizing premature failure of FRP systems by avoiding scenarios that may trigger debonding failure.

8.3 Feature Interaction

The correlation matrix of the used database in this case study has already been presented and discussed in Table 2. Thus, Table 5 lists the interaction between features using mutual information association which measures how much information the presence/absence of a term contributes to making the correct prediction. This table shows the existence of a strong association between moment arm and material properties (compressive strength, yield strength, and modulus of FRP), as well as a slightly strong association with FRP ratio and steel ratio. Other noteworthy associations also exist between other features such as yield strength and FRP modulus, and compressive strength, and yield strength. The moment capacity seems to have a good association with all features (>0.500), with the exception of the strengthening method. Similar observations can also be made for both cases of classification. A complimentary discussion on feature interaction is revisited in the surrogate section.

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

Table 5 Explainability through association matrix

Regression	*M*	*a*	f_c	f_y	E_{fr}	f_f	*r*	*T*
M	1.000							
a	0.551	1.000						
f_c	0.517	**0.737**	1.000					
f_y	0.560	**0.854**	**0.762**	1.000				
E_{fr}	0.534	**0.735**	0.684	**0.762**	1.000			
f_f	0.560	0.596	0.560	0.624	0.651	1.000		
r	0.538	0.649	0.608	0.669	0.581	0.493	1.000	
T	0.160	0.389	0.203	0.245	0.219	0.207	0.243	1.000
Binary classification	*Debonding*	*r*	*T*	f_c	*a*	E_{fr}	f_r	f_y
Debonding	1.000							
r	0.238	1.000						
T	0.019	0.238	1.000					
f_c	0.211	0.600	0.202	1.000				
a	0.198	0.628	0.350	**0.721**	1.000			
E_{fr}	0.167	0.575	0.219	0.681	**0.721**	1.000		
f_r	0.123	0.486	0.209	0.544	0.583	0.652	1.000	
f_y	0.177	0.679	0.293	**0.787**	**0.862**	**0.779**	0.637	1.000
Multi classification	*Debonding*	f_c	*r*	f_y	E_{fr}	*T*	f_r	*a*
Debonding	1.000							
f_c	0.302	1.000						
r	0.305	0.600	1.000					
f_y	0.316	**0.787**	0.679	1.000				
E_{fr}	0.330	0.681	0.575	**0.779**	1.000			
T	0.055	0.202	0.238	0.293	0.219	1.000		
f_r	0.273	0.544	0.486	0.637	0.652	0.209	1.000	
a	0.279	**0.721**	0.628	0.862	**0.721**	0.350	0.583	1.000

8.4 SHAP (SHapley Additive exPlanations)

The "base value" for the SHAP method was calculated at 111.73 kN.m. To showcase why did the ExGBT algorithm arrives at a particular prediction, one prediction for a given beam is examined herein. This beam (P3) was tested by Kotynia [92] and was made from concrete with a compressive strength of 44 MPa, yield strength of steel reinforcement of 541 MPa, ratio of steel reinforcement of 0.502, with FRP ratio of 0.24, modulus of FRP of 163 GPa, moment arm of 1400 mm and NSM strengthening type. For this particular beam, the ExGBT predicts a moment at failure of 112.6 kN.m (with a measured moment of 106 kN.m). Using the SHAP method, we can then see that the model predicts 112.6 kN.m as a result of the SHAP values calculated for the contribution of individual features. These contributions add up at +0.432 for moment arm, +0.026 for

strengthening type, -0.253 for compressive strength of concrete, -0.103 for ratio of steel reinforcement, -0.056 for yield strength of steel reinforcement, and -0.038 for all other features (see Fig. 5). As expected, the algebraic summation of these values adds to the difference between the prediction and the "base value". As such, the summation of the contributions with the prediction values adds up to Zero.

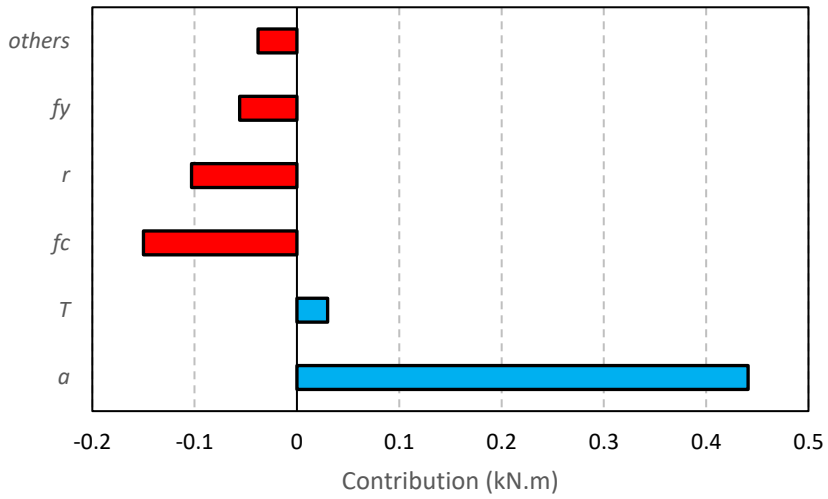


Fig. 5 Explainability through the SHAP method (Note: causes behind difference in model prediction (112.6 kN.m) and SHAP base value (111.73 kN.m))

8.5 Surrogates

A surrogate model is one that is simple, and is then used to augment predictions from a more complex model. Once the surrogate model passes the training procedure with good performance, then this model is said to be able to explain the predictions from the complex model. In this study, two models are used, one that is the generalized additive model (GAM) and another that is based on genetic algorithm (GA), to surrogate the ExGBT model. The GAM model was able to augment the ExGBT model with the following metrics; MAPE (37.16/2.03/41.35 kN.m), RMSE (32.45/31.40/52.69 kN.m), and R^2 (0.96/0.94/0.85) for training/validation/testing, respectively. These metrics show the validity of the GAM model.

Due to its simplicity, the GAM model can be represented by a linear formula. This formula applies the inverse of the link function used in the GAM model (i.e., exponential function), being multiplied to the sum of all standardized features multiplied by computed coefficient (see Eq. 2) in addition to an intercept (if any). Thus, Eq. 2 now augments the rationale of the ExGBT model and hence can be used directly without the need to re-run the ExGBT model for each moment capacity prediction. Equation 3 shows a logistic regression model (a simple model used to surrogate a complex classification-based model) for the case of binary classification [117]. This model achieved the following metrics BACC (0.75/0.87/0.68), AUC (0.68/0.85/0.58), and LLE (0.60/0.47/0.71).

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction*. https://doi.org/10.1016/j.autcon.2021.103821

$$Prediction_{GAM} = link\ function\ \Big[(a_{standardized} \times 63.31) + (r_{standardized} \times 16.84) + \Big(f_{y_{standardized}} \times -5.83\Big) + \Big(f_{c_{standardized}} \times -20.44\Big) + \Big(f_{r_{standardized}} \times -1.22\Big) + \Big(E_{fr_{standardized}} \times -2.28\Big) + (T_{standardized} \times -7.28) + 94.10\Big] \quad (2)$$

$$Prediction_{LR} = logit\ \Big[(a_{standardized} \times -1.03) + (r_{standardized} \times -1.74) + \Big(f_{y_{standardized}} \times 0.24\Big) + \Big(f_{c_{standardized}} \times -0.0002\Big) + \Big(f_{r_{standardized}} \times 0.82\Big) + \Big(E_{fr_{standardized}} \times 1.06\Big) + (T_{standardized} \times 0.10) + 1.39\Big] \quad (3)$$

Another surrogate that can be used to augment an AI/ML model is genetic algorithms (GA). In this surrogate, and similar to GAMs, a string of representations can be formed [118]. In this analysis, a GA model was developed and achieved the following performance in MAPE (66.86/35.72/29.91 kN.m), RMSE (50.38/37.38/48.52 kN.m), and R^2 (0.91/0.92/0.87) for training/validation/testing, respectively. The generated representation is shown in Eq. 4. As one can see, this GA representation only contains five features, as the rest of the features were dropped as they were not deemed to be of importance. Interestingly, this expression also shows the direct but minor interaction between yield strength and FRP ratio, as well as additivity between all other features[2]. Also, another GA model for the binary classification model is shown in Eq. 5 with the following metrics BACC (0.82/0.82/0.78), AUC (0.85/0.84/0.75), and LLE (0.52/0.48/0.55). Further details with regard to using GA in this database can be found elsewhere [74]. One should note that there is an ongoing debate on the benefits vs. drawbacks of surrogates [21], and additional discussion on surrogates is provided in the next section.

$$Prediction_{GA_{Moment}} = exp\ (3.48 + 0.312r + 0.0015a - 0.0006f_y - 0.002E_{fr} - 0.0007f_y f_f) \quad (4)$$

$$Prediction_{GA_{Binary}} = logistic\ (1.2 \times lesser\ of[r, 0.74] - 0.056 - 0.0003 \times a) \quad (5)$$

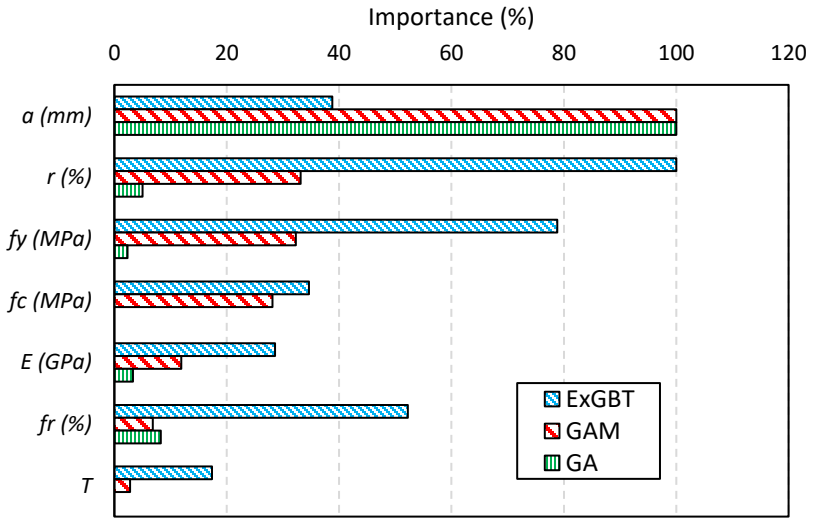
9.0 Insights into Explainability Methods

The first four discussed explainability methods provide insights into the working of complex models such as ExGBT. These methods tend to be informative and showcase why a model predicts the way it does. While the use of surrogates may also help explain such models, this method may not be as neat as that observed from other methods [119].

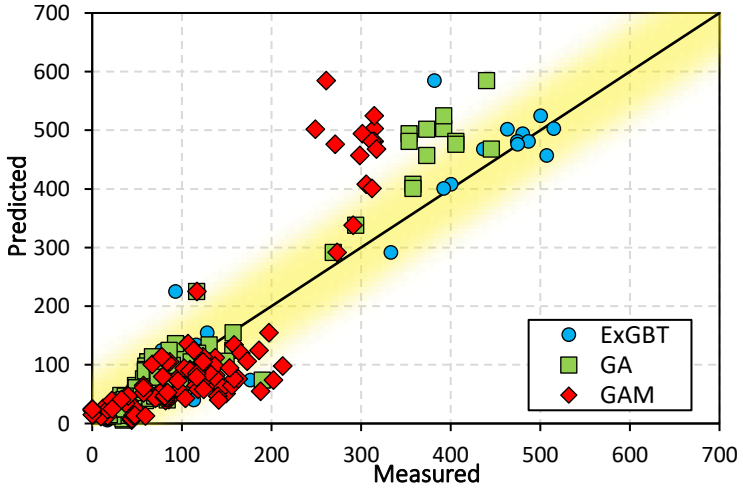
As shown in Fig. 6a, extracted feature importance of GAM and GA are not identical to that obtained from ExGBT. This is expected as GAM and GA are unique models on their own and do

[2] A more complex representation was also identified as:
$Prediction_{GA_{Moment}} = exp\ (7.64 + 0.68ff_f + 0.0065rE_{fr} + 0.0014f_c^2 + 6.09 \times 10^{-5} f_c a + 1.38E_{fr}^2 + 5.45 \times 10^{-6} f_y E_{fr} - 0.012E_{fr} - 0.165f_c - 0.65f_f - 1.21f_y a - 0.40r^2)$
This representation was not further examined as the goal of this work is to articulate simple surrogates. However, this representation shows more complex interaction between features than that displayed in Eq. 4.

not aim to validate ExGBT, but rather to explain its behavior. From this view, explaining the ExGBT's behavior does not necessarily rely on having the same importance for all features, but rather it aims to use the simplicity of the GAM/GA models to understand ExGBT predictions. When evaluating both surrogates herein, one must keep in mind; 1) surrogates are applied to a dataset that has the same features as that of the original dataset, but with an outcome (response) obtained from ExGBT predictions (i.e., in a way, the new dataset is a derivative of the original one), and 2) GAM and GA have their own inner working mechanisms and hence do not necessarily need to oblige or follow that of the ExGBT.



(a) Feature importance comparison between GAM, GA, and ExGBT

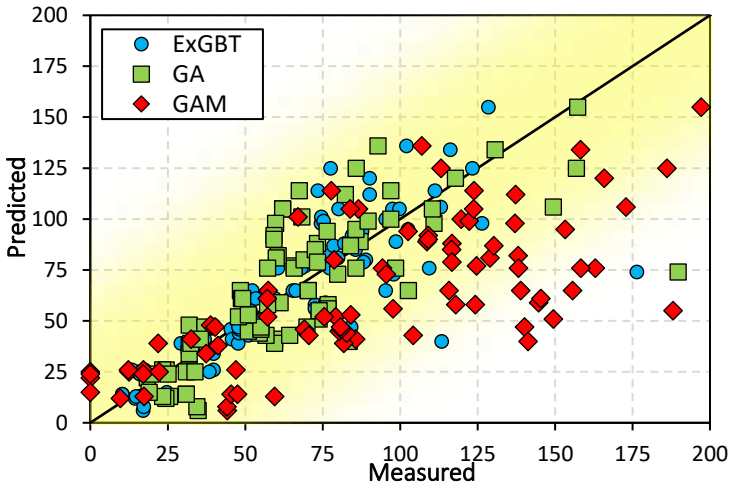


(b) Predictions vs measured observations (Range: 0-700 kN.m)

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.autcon.2021.103821

Please cite this paper as:

Naser M.Z. (2021). "An Engineer's Guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating Causality, Forced Goodness, and the False Perception of Inference." *Automation in Construction.* https://doi.org/10.1016/j.autcon.2021.103821



(c) Predictions vs measured observations (Range: 0-200 kN.m) [shaded area represents 10% error range with 700 kN.m as maximum moment]

Fig. 6 Comparison between ExGBT, GAM, and GA predictions

To further emphasize the above notion, the plots shown in Figs. 6b and 6c are thought to be interesting to discuss as they present GAM's and GA's response when compared to the measured values, as well as to ExGBT's predictions. These figures complement the performance of both surrogates, which also attained good performance metrics [GAM; MAPE (37.16/2.03/41.35 kN.m), RMSE (32.45/31.40/52.69 kN.m), and R^2 (0.96/0.94/0.85), and GA; MAPE (66.86/35.72/29.91 kN.m), RMSE (50.38/37.38/48.52 kN.m), and R^2 (0.91/0.92/0.87)]. A closer look into these two figures also shows how GAM and GA seem to capture the phenomenon on hand well, especially in beams with moments reaching around 200 kN.m. Beyond this range, both GAM and GA deviate their predictions (unlike that observed by ExGBT, the original model used in this study). This case study shows the need for a deeper dive into surrogates and their behavior worth further investigation across the range of all datasets. This is best suited for a future work.

In lieu of the observations as triggered by Fig. 6, a question then arises, why would a user not apply simple models like GAM or GA directly instead of more complex models? The answer is simple. A user can indeed apply such models. In fact, and *ceteris paribus*, a user is advised to try to apply simple models first. If such models fail or do not achieve the required degree of goodness, then a user may opt to apply more complex models [120]. At the time of this work, the use of multi-search models, ensembles, and others seem to be worthy of exploring. One should be cognizant of the fact that selecting a model to explore a given phenomenon via AI/ML does not follow a standardized procedure simply because we do not have such a procedure yet. On a more positive note, the open literature does contain a set of recommendations obtained via domain-specific and expert-guided examinations, which can be used to pave the way towards a more standardized AI/ML application procedure [121–123].

A similar notion can also be argued for the quality of explanation. While an explanation, or a method of explainability, can satisfy a user, the same degree of reasonability of explainability may not provide enough evidence to satisfy others [124–126]. In a way, there could potentially be a

tradeoff between the degree of explainability to that expected level of satisfaction a user is looking for (e.g., given their background, or need to comply with rules or societal expectations, etc.). Thus, future works in our domain are invited to explore the notion of XAI and IML and develop explainable models that best suit our problems, expectations, and norm.

10.0 Conclusions

This paper showcases the applicability of five explainability methods (feature importance, partial dependence plots, feature interactions, SHAP (SHapley Additive exPlanations), and surrogates) via a thorough examination of a case study carried out on reinforced concrete (RC) beams strengthened with fiber-reinforced polymer (FRP) composite laminates. In this case study, three algorithms, namely: Extreme Gradient Boosted Trees (ExGBT), Light gradient boosted trees (LGBT), and Keras Deep Neural Networks (KDNN), are applied to predict the maximum moment capacity of FRP-strengthened beams and the propensity of FRP systems to fail through debonding mechanisms. The following list of inferences can be drawn from the findings of this study:

- Structural engineers remain reluctant to adopt AI/ML methods as primary tools of analysis and design of structures due to a few issues related to the opaque nature of AI/ML and to the limited coding/programming educational experience in typical curricula.
- Current efforts in AI/ML can benefit from focusing on explainability and interpretability, thereby clearing the blackbox reputation of common AI/ML methods in the structural engineering domain.
- Of all algorithms showcased herein, the Extreme Gradient Boosted Trees (ExGBT) ranked the best performance in examining the maximum moment capacity and failure mechanism of FRP-strengthened beams.
- Using simple surrogate models such as GAM and GA can help augment more complex models such as ExGBT.

Data Availability

The datasets used in this paper is uploaded to the Mendeley public dataset. This database can be accessed at https://data.mendeley.com/datasets/8hv5grr49s/4 starting January 2022.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgement

I would like to thank the Editor and Reviewers for their support of this work and constructive comments that enhanced the quality of this manuscript. This manuscript is dedicated to Prof. Jack McCormack – the author of Structural Steel Design – who passed away in June 2021 at 93 years of age. Prof. McCormack taught at Clemson University for 36 years. He was kind, funny and an unparalleled engineer. Rest in Peace, "Happy Jack".

References

[1] W. Prager, J.E. Taylor, Problems of optimal structural design, Journal of Applied

Mechanics, Transactions. Vol. 35, No. 1, (1964). pp. 102-106. https://doi.org/10.1115/1.3601120.

[2] J. Rutherford, Practical Experiment Designs for Engineers and Scientists, 2002. pp. 448. Wiley. ISBN-10: 0471390542. https://doi.org/10.1198/tech.2002.s79.

[3] J. Antoy, Design of Experiments for Engineers and Scientists: Second Edition, 2014. ISBN-10: 0080994172. https://doi.org/10.1016/C2012-0-03558-2.

[4] The Economist, Can we fix it? - The construction industry's productivity problem | Leaders | The Economist, (2017). https://www.economist.com/leaders/2017/08/17/the-construction-industrys-productivity-problem (accessed August 27, 2020).

[5] R. Bogue, What are the prospects for robots in the construction industry?, Industrial Robot. Vol. 45, No. 1, (2018), pp. 1-6. https://doi.org/10.1108/IR-11-2017-0194.

[6] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects., Science. Vol. 349, (2015). pp. 255–60. https://doi.org/10.1126/science.aaa8415.

[7] W.Z. Taffese, E. Sistonen, Machine learning for durability and service-life assessment of reinforced concrete structures: Recent advances and future directions, Automation in Construction. Vol. 77, (2017), pp. 1-14. https://doi.org/10.1016/j.autcon.2017.01.016.

[8] C.V. Dung, L.D. Anh, Autonomous concrete crack detection using deep fully convolutional neural network, Automation in Construction. Vol. 99, (2019), pp. 52-58, https://doi.org/10.1016/j.autcon.2018.11.028.

[9] E. Valero, A. Forster, F. Bosché, E. Hyslop, L. Wilson, A. Turmel, Automated defect detection and classification in ashlar masonry walls using machine learning, Automation in Construction. Vol. 106, (2019). https://doi.org/10.1016/j.autcon.2019.102846.

[10] I.H. Witten, E. Frank, M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2011. Morgan Kaufmann; 3rd edition. pp. 664. ISBN-10: 0123748569

[11] M. Naser, H. Hostetter, A. Daware, AI modelling & mapping functions: a cognitive, physics- guided, simulation-free and instantaneous approach to fire evaluation, in: The 11th International Conference on Structures in Fire, The University of Queensland, Brisbane, Australia, 2020. pp. 590-599. https://doi.org/10.14264/a0b3b36

[12] A.H. Alavi, A.H. Gandomi, M.G. Sahab, M. Gandomi, Multi expression programming: A new approach to formulation of soil classification, Engineering with Computers. Vol. 26 (2010), pp. 111–118. https://doi.org/10.1007/s00366-009-0140-7.

[13] A. Seitlllari, M.Z. Naser, Leveraging artificial intelligence to assess explosive spalling in fire-exposed RC columns, Computers and Concrete. Vol. 24, (2019), pp. 271-282. https://doi.org/10.12989/cac.2019.24.3.271.

[14] R. Solhmirzaei, H. Salehi, V. Kodur, M.Z. Naser, Machine learning framework for

predicting failure mode and shear capacity of ultra high performance concrete beams, Engineering Structures. Vol. 224. (2020). https://doi.org/10.1016/j.engstruct.2020.111221.

[15] H. Huang, H. V. Burton, Classification of in-plane failure modes for reinforced concrete frames with infills using machine learning, Journal of Building Engineering. Vol. 25. (2019). https://doi.org/10.1016/j.jobe.2019.100767.

[16] A. Dexters, R.R. Leisted, R. Van Coile, S. Welch, G. Jomaas, Testing for Knowledge: Maximising Information Obtained from Fire Tests by using Machine Learning Techniques, Proceedings of Interflam 2019. Egham, United Kingdom. http://hdl.handle.net/1854/LU-8622485.

[17] A. Kaveh, M.Z. Kabir, M. Bohlool, Optimum design of three-dimensional steel frames with prismatic and non-prismatic elements, Engineering with Computers. Vol. 36, (2019), pp. 1011–1027, https://doi.org/10.1007/s00366-019-00746-9.

[18] E. Tjoa, C. Guan, Quantifying Explainability of Saliency Methods in Deep Neural Networks, (2020). https://arxiv.org/pdf/2009.02899v1.pdf

[19] F.K. Dosilovic, M. Brcic, N. Hlupic, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018. https://doi.org/10.23919/MIPRO.2018.8400040.

[20] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2019. Turin, Italy. https://doi.org/10.1109/DSAA.2018.00018.

[21] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence. Vol. 1, (2019), pp. 206–215. https://doi.org/10.1038/s42256-019-0048-x.

[22] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.Z. Yang, XAI-Explainable artificial intelligence, Science Robotics. Vol. 4, (2019). https://doi.org/10.1126/scirobotics.aay7120.

[23] B. Boehmke, B. Greenwell, B. Boehmke, B. Greenwell, Interpretable Machine Learning, in: Hands-On Machine Learning with R, 2020. Chapman and Hall/CRC. ISBN: 9780367816377, https://doi.org/10.1201/9780367816377-16.

[24] The Royal Society, Explainable AI: the basics – Policy Briefing, 2019. ISBN: 978-1-78252-433-5.

[25] B. D'Amico, R.J. Myers, J. Sykes, E. Voss, B. Cousins-Jenvey, W. Fawcett, S. Richardson, A. Kermani, F. Pomponi, Machine Learning for Sustainable Structures: A Call for Data, Structures. Vol. 19, (2019), pp. 1-4.

https://doi.org/10.1016/j.istruc.2018.11.013.

[26] X. Zhu, Machine teaching: An inverse problem to machine learning and an approach toward optimal education, in: AAAI15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. pp. 4083–4087. https://dl.acm.org/doi/10.5555/2888116.2888288.

[27] A. Behnood, E.M. Golafshani, Machine learning study of the mechanical properties of concretes containing waste foundry sand, Construction and Building Materials. Vol. 243. (2020). https://doi.org/10.1016/j.conbuildmat.2020.118152.

[28] A. Behnood, E. Mohammadi Golafshani, Predicting the dynamic modulus of asphalt mixture using machine learning techniques: An application of multi biogeography-based programming, Construction and Building Materials. Vo. 266, (2021). https://doi.org/10.1016/j.conbuildmat.2020.120983.

[29] S. Mangalathu, H. Jang, S.H. Hwang, J.S. Jeon, Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls, Engineering Structures. Vol. 208, (2020). https://doi.org/10.1016/j.engstruct.2020.110331.

[30] S. Mangalathu, S.H. Hwang, E. Choi, J.S. Jeon, Rapid seismic damage evaluation of bridge portfolios using machine learning techniques, Engineering Structures. Vo. 201, (2019). https://doi.org/10.1016/j.engstruct.2019.109785.

[31] M.Z. Naser, Heuristic machine cognition to predict fire-induced spalling and fire resistance of concrete structures, Automation in Construction. Vol. 106, (2019). https://doi.org/10.1016/J.AUTCON.2019.102916.

[32] M.Z. Naser, V.A. Uppala, Properties and material models for construction materials post exposure to elevated temperatures, Mechanics of Materials. Vol. 142, (2020). https://doi.org/10.1016/j.mechmat.2019.103293.

[33] A. Siam, M. Ezzeldin, W. El-Dakhakhni, Machine learning algorithms for structural performance classifications and predictions: Application to reinforced masonry shear walls, Structures. Vol. 22, (2019). https://doi.org/10.1016/j.istruc.2019.06.017.

[34] S. Mangalathu, G. Heo, J.S. Jeon, Artificial neural network based multi-dimensional fragility development of skewed concrete bridge classes, Engineering Structures. Vol. 162, (2018), pp. 166-167. https://doi.org/10.1016/j.engstruct.2018.01.053.

[35] A.A.H. Alwanas, A.A. Al-Musawi, S.Q. Salih, H. Tao, M. Ali, Z.M. Yaseen, Load-carrying capacity and mode failure simulation of beam-column joint connection: Application of self-tuning machine learning model, Engineering Structures. Vol. 194, (2019). pp. 220-229. https://doi.org/10.1016/j.engstruct.2019.05.048.

[36] S. Inkoom, J. Sobanjo, A. Barbu, X. Niu, Prediction of the crack condition of highway pavements using machine learning models, Structure and Infrastructure Engineering. Vol.

15, (2019), pp. 940-953. https://doi.org/10.1080/15732479.2019.1581230.

[37] M.H. Rafiei, H. Adeli, A novel machine learning-based algorithm to detect damage in high-rise building structures, Structural Design of Tall and Special Buildings. Vol. 26, (2017), pp. 1-11. https://doi.org/10.1002/tal.1400.

[38] Y. Ai, K. Xu, Feature extraction based on contourlet transform and its application to surface inspection of metals, Optical Engineering. Vol. 51, (2012), pp. 1-8. https://doi.org/10.1117/1.oe.51.11.113605.

[39] J.K. Park, B.K. Kwon, J.H. Park, D.J. Kang, Machine learning-based imaging system for surface defect inspection, International Journal of Precision Engineering and Manufacturing - Green Technology. Vol. 3, (2016), pp. 303–310. https://doi.org/10.1007/s40684-016-0039-x.

[40] G.S. Bobadilha, C.E. Stokes, D.J. Verly Lopes, Artificial neural networks modelling based on visual analysis of coated cross laminated timber (CLT) to predict color change during outdoor exposure, Holzforschung. Vol. 73, (2020). https://doi.org/10.1515/hf-2020-0193.

[41] H. Kim, E. Ahn, M. Shin, S.H. Sim, Crack and Noncrack Classification from Concrete Surface Images Using Machine Learning, Structural Health Monitoring. Vol. 18, (2019), pp. 725-738. https://doi.org/10.1177/1475921718768747.

[42] M. Zhang, C.N. Sun, X. Zhang, P.C. Goh, J. Wei, D. Hardacre, H. Li, High cycle fatigue life prediction of laser additive manufactured stainless steel: A machine learning approach, International Journal of Fatigue. Vol. 128, (2019). https://doi.org/10.1016/j.ijfatigue.2019.105194.

[43] N. Lu, M. Noori, Y. Liu, Fatigue Reliability Assessment of Welded Steel Bridge Decks under Stochastic Truck Loads via Machine Learning, Journal of Bridge Engineering. Vol. 22, (2017), pp. 1-12. https://doi.org/10.1061/(asce)be.1943-5592.0000982.

[44] W. Ben Chaabene, M. Flah, M.L. Nehdi, Machine learning prediction of mechanical properties of concrete: Critical review, Construction and Building Materials. Vol. 260, (2020). https://doi.org/10.1016/j.conbuildmat.2020.119889.

[45] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access. Vol. 6, (2018), pp. 52138 - 52160. https://doi.org/10.1109/ACCESS.2018.2870052.

[46] J. Pearl, Causal inference in statistics: An overview, Statistics Surveys. Vol. 3, (2009), pp. 96 - 146. https://doi.org/10.1214/09-SS057.

[47] D. Morgan, R. Jacobs, Opportunities and Challenges for Machine Learning in Materials Science, Annual Review of Materials Research. Vol. 50, (2020), pp. 71-103, https://doi.org/10.1146/annurev-matsci-070218-010015.

[48] C. Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable., Lulu (2019), pp. 320. https://christophm.github.io/interpretable-ml-book/ (accessed on June 15, 2021).

[49] C.C.S. Liem, M. Langer, A. Demetriou, A.M.F. Hiemstra, A.S. Wicaksana, M.P. Born, C.J. König, Explainable and interpretable models in computer vision and machine learning, 2018. ISBN: 978-3-319-98131-4, pp. 299. https://doi.org/10.1007/978-3-319-98131-4

[50] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Communications of the ACM. Vol. 63, (2020), pp. 68–77, https://doi.org/10.1145/3359786.

[51] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016, https://doi.org/10.1145/2939672.2939778.

[52] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, https://dl.acm.org/doi/10.5555/3295222.3295230.

[53] T.J. Hastie, D. Pregibon, Generalized linear models, in: Statistical Models in S, 2017, Routledge, ISBN: 9780203738535, https://doi.org/10.1201/9780203738535.

[54] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[55] Z.C. Lipton, The mythos of model interpretability, Communications of the ACM. Vol. 61, (2018), pp. 36–43. https://doi.org/10.1145/3233231.

[56] J.B. Lyons, K.S. Koltai, N.T. Ho, W.B. Johnson, D.E. Smith, R.J. Shively, Engineering Trust in Complex Automated Systems, Ergonomics in Design. Vol. 24, (2016), pp. 13-17. https://doi.org/10.1177/1064804615611272.

[57] J. Heyman, The Science of Structural Engineering, 1999. Imperial College Press, ISBN-10: 1860941893, https://doi.org/10.1142/p163.

[58] X. Zhu, C. Vondrick, C.C. Fowlkes, D. Ramanan, Do We Need More Training Data?, International Journal of Computer Vision. Vol. 119, (2016), pp. 76–92. https://doi.org/10.1007/s11263-015-0812-2.

[59] B. Cohen, The Rise of Engineering's Social Justice Warriors — The James G. Martin Center for Academic Renewal, The James G. Martin Center for Academic Renewal . (2018). https://www.jamesgmartin.center/2018/11/the-rise-of-engineerings-social-justice-warriors/ (accessed February 22, 2021).

[60] B.E. Ross, D.A. Chen, S. Conejos, A. Khademi, Enabling Adaptable Buildings: Results of a Preliminary Expert Survey, in: Procedia Engineering, Vol. 145, (2016), pp. 420-427. https://doi.org/10.1016/j.proeng.2016.04.009.

[61] K.T. Adams, M. Osmani, T. Thorpe, J. Thornback, Circular economy in construction: Current awareness, challenges and enablers, in: Proceedings of the Institution of Civil Engineers - Waste and Resource Management. Manag., Vol. 170, (2017), pp. 15-24. https://doi.org/10.1680/jwarm.16.00011.

[62] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence. Vol. 267, (2019), pp. 1-38. https://doi.org/10.1016/j.artint.2018.07.007.

[63] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: A corrected feature importance measure, Bioinformatics. Vol. 26, (2010), pp. 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

[64] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics. Vl. 29, (2001), pp. 1189-1232. https://doi.org/10.1214/aos/1013203451.

[65] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation, Journal of Computational and Graphical Statistics. Vol. 24, (2015), pp. 44-65. https://doi.org/10.1080/10618600.2014.907095.

[66] K. Goyal, S. Dumancic, H. Blockeel, Feature Interactions in XGBoost, (2020), https://arxiv.org/ftp/arxiv/papers/2007/2007.05758.pdf.

[67] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers and Electrical Engineering. Vol. 40, (2014), pp. 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

[68] F.N. David, H. Cramer, Mathematical Methods of Statistics., Biometrika. Vol. 34, (1947), pp. 347. https://doi.org/10.2307/2332454.

[69] S.M. Lundberg, G.G. Erion, S.I. Lee, Consistent individualized feature attribution for tree ensembles, (2018). https://arxiv.org/pdf/1802.03888.pdf.

[70] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C.W. Rosenbrock, G. Csányi, D.W. Wingate, G.L.W. Hart, Machine-learned multi-system surrogate models for materials prediction, npj Computational Materials. Vol. 5, (2019). https://doi.org/10.1038/s41524-019-0189-9.

[71] D. Gorissen, I. Couckuyt, P. Demeester, T. Dhaene, K. Crombecq, A surrogate modeling and adaptive sampling toolbox for computer based design, Journal of Machine Learning Research. Vol. 11, (2010), pp. 2051-2055. http://jmlr.org/papers/v11/gorissen10a.html.

[72] M.Z. Naser, A. Seitllari, Concrete under fire: an assessment through intelligent pattern

recognition, Engineering with Computers. Vol. 36, (2020), pp. 1915–1928. https://doi.org/10.1007/s00366-019-00805-1.

[73] M.Z. Naser, Observational Analysis of Fire-Induced Spalling of Concrete through Ensemble Machine Learning and Surrogate Modeling, Journal of Materials in Civil Engineering. Vol. 33, (2021), pp. 1-11. https://doi.org/10.1061/(ASCE)MT.1943-5533.0003525.

[74] M.Z. Naser, Machine learning assessment of fiber-reinforced polymer-strengthened and reinforced concrete members, ACI Structural Journal. Vol. 117, (2020), pp. 237-251. https://doi.org/10.14359/51728073.

[75] P. Alagusundaramoorthy, I.E. Harik, C.C. Choo, Flexural behavior of R/C beams strengthened with carbon fiber reinforced polymer sheets or fabric, Journal of Composites for Construction. Vol. 7, (2003), pp. 292-301. https://doi.org/10.1061/(ASCE)1090-0268(2003)7:4(292).

[76] T.H. Almusallam, Y.A. Al-Salloum, Ultimate strength prediction for RC beams externally strengthened by composite materials, Composites Part B:Engineering. Vol. 32, (2001), pp. 609-619. https://doi.org/10.1016/S1359-8368(01)00008-7.

[77] H. Rahimi, A. Hutchinson, Concrete beams strengthened with externally bonded FRP plates, Journal of Composites for Construction. Vol. 5, (2001), pp. 44-56. https://doi.org/10.1061/(ASCE)1090-0268(2001)5:1(44).

[78] R.A. Hawileh, M.Z. Naser, J.A. Abdalla, Finite element simulation of reinforced concrete beams externally strengthened with short-length CFRP plates, Composites Part B: Engineering. Vol. 45, (2013), pp. 1722–1730. https://doi.org/10.1016/j.compositesb.2012.09.032.

[79] A.K. Al-Tamimi, R. Hawileh, J. Abdalla, H.A. Rasheed, Effects of Ratio of CFRP Plate Length to Shear Span and End Anchorage on Flexural Behavior of SCC RC Beams, Journal of Composites for Construction. Vol. 15, (2011), pp. 908–919. https://doi.org/10.1061/(ASCE)CC.1943-5614.0000221.

[80] C. Sabau, C. Popescu, G. Sas, J.W. Schmidt, T. Blanksvärd, B. Täljsten, Strengthening of RC beams using bottom and side NSM reinforcement, Composites Part B: Engineering. Vol. 149, (2018), pp. 82-91. https://doi.org/10.1016/j.compositesb.2018.05.011.

[81] I.A. Sharaky, L. Torres, J. Comas, C. Barris, Flexural response of reinforced concrete (RC) beams strengthened with near surface mounted (NSM) fibre reinforced polymer (FRP) bars, Composite Structures. Vol. 109, (2014), pp. 8-22. https://doi.org/10.1016/j.compstruct.2013.10.051.

[82] J.G. Teng, L. De Lorenzis, B. Wang, R. Li, T.N. Wong, L. Lam, Debonding failures of RC beams strengthened with near surface mounted CFRP strips, Journal of Composites

for Construction. Vol. 10, (2006), pp. 92-105. https://doi.org/10.1061/(ASCE)1090-0268(2006)10:2(92).

[83] T.C. Triantafillou, N. Plevris, Strengthening of RC beams with epoxy-bonded fibre-composite materials, Materials and Structures. Vol. 25, (1992), pp. 201–211. https://doi.org/10.1007/BF02473064.

[84] W. Wenwei, L. Guo, Experimental study and analysis of RC beams strengthened with CFRP laminates under sustaining load, International Journal of Solids and Structures. Vol. 43, (2006), pp. 1372-1387. https://doi.org/10.1016/j.ijsolstr.2005.03.076.

[85] M. Arduini, A. Di Tommaso, A. Nanni, Brittle failure in FRP plate and sheet bonded beams, ACI Structural Journal. Vol. 43, (1997), pp. 363-370. https://doi.org/10.14359/487.

[86] F. Ceroni, Experimental performances of RC beams strengthened with FRP materials, Construction and Building Materials. Vol. 24, (2010), pp. 1547-1559. https://doi.org/10.1016/j.conbuildmat.2010.03.008.

[87] V.S. Kuntal, M. Chellapandian, S.S. Prakash, Efficient near surface mounted CFRP shear strengthening of high strength prestressed concrete beams – An experimental study, Composite Structures. Vol. 180, (2017), pp. 16-28. https://doi.org/10.1016/j.compstruct.2017.07.095.

[88] S.M. Daghash, O.E. Ozbulut, Flexural performance evaluation of NSM basalt FRP-strengthened concrete beams using digital image correlation system, Composite Structures. Vol. 176, (2017), pp. 748-756. https://doi.org/10.1016/j.compstruct.2017.06.021.

[89] S.J.E. Dias, J.A.O. Barros, W. Janwaen, Behavior of RC beams flexurally strengthened with NSM CFRP laminates, Composite Structures. Vol. 201, (2018), pp. 363-376. https://doi.org/10.1016/j.compstruct.2018.05.126.

[90] P.J. Fanning, O. Kelly, Ultimate response of RC beams strengthened with CFRP plates, Journal of Composites for Construction. Vol. 5, (2001), pp. 122-127. https://doi.org/10.1061/(ASCE)1090-0268(2001)5:2(122).

[91] F. Al-Mahmoud, A. Castel, R. François, C. Tourneur, RC beams strengthened with NSM CFRP rods and modeling of peeling-off failure, Composite Structures. Vol. 92, (2010), pp. 1920-1930. https://doi.org/10.1016/j.compstruct.2010.01.002.

[92] R. Kotynia, Analysis of the flexural response of NSM FRP-strengthened concrete beams, in: Proceedings of the 8th International Symposium on Fiber Reinforced Polymer Reinforcement for Reinforced Concrete Structures (FRPRCS-8), 2007: pp. 16–18.

[93] Canadian Standards Association (CSA), CSA S806, Design and construction of building components with fibre-reinforced polymers, 2002.

https://standards.globalspec.com/std/10073547/csa-s806 (Accessed on 6/15/2021).

[94] BRI, Guidelines for Structural Design of FRP Reinforced Concrete Building Structures, Building Research Institute, Japan, 1997. http://www.jsce.or.jp/committee/concrete/e/newsletter/newsletter01/recommendation/FRP-bar/document.htm (Accessed on 6/15/2021)

[95] ACI committee 440, Guide for the Design and Construction of Structural Concrete Reinforced with Fiber-Reinforced Polymer (FRP) Bars, 2015. ISBN: 9781942727101.

[96] CNRDT-203, Guide for the design and construction of concrete structures reinforced with fiber-reinforced polymer bars, Italy, 2007. https://www.cnr.it/en/node/2639 (Accessed on 6/15/2021)

[97] E.R. Ziegel, The Elements of Statistical Learning, Technometrics. Vol. 45, (2003), pp. 267. https://doi.org/10.1198/tech.2003.s770.

[98] N. Ketkar, N. Ketkar, Introduction to Keras, in: Deep Learn. with Python, 2017. Apress, Berkeley, CA, ISBN: 978-1-4842-2765-7. https://doi.org/10.1007/978-1-4842-2766-4_7.

[99] Y. Freund, R.E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences. Vol. 55, (1997), pp. 119-139. https://doi.org/10.1006/jcss.1997.1504.

[100] Gradient boosted tree (GBT), (2019). https://software.intel.com/en-us/daal-programming-guide-details-24 (accessed April 9, 2021).

[101] Scikit, sklearn.ensemble.GradientBoostingRegressor — scikit-learn 0.24.1 documentation, (2020). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed February 9, 2021).

[102] XGBoost Python Package, Python Package Introduction — xgboost 1.4.0-SNAPSHOT documentation, (2020). https://xgboost.readthedocs.io/en/latest/python/python_intro.html#early-stopping (accessed February 10, 2021).

[103] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm, Proceedings of the 13th International Conference on Machine Learning. (1996). https://doi.org/10.1.1.133.1040. https://dl.acm.org/doi/10.5555/3091696.3091715.

[104] M.Z. Naser, Mechanistically Informed Machine Learning and Artificial Intelligence in Fire Engineering and Sciences, Fire Technology. (2021) pp. 1–44. https://doi.org/10.1007/s10694-020-01069-8.

[105] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, Proceedings of the 31st International

Conference on Neural Information Processing Systems. 2017, pp. 3149–3157, https://dl.acm.org/doi/10.5555/3294996.3295074.

[106] LightGBM, Welcome to LightGBM's documentation! — LightGBM 3.1.1.99 documentation, (2020). https://lightgbm.readthedocs.io/en/latest/ (accessed February 9, 2021).

[107] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, in: Advances in Neural Information Processing Systems., 2018. https://arxiv.org/abs/1712.09913

[108] Keras, GitHub - keras-team/keras: Deep Learning for humans, (2020). https://github.com/keras-team/keras (accessed February 9, 2021).

[109] M.D. Schmidt, H. Lipson, Age-fitness pareto optimization, in: Proceedings of the 12th annual conference on Genetic and evolutionary computation, 2010, Portland, Oregon, USA, pp. 543–544. https://doi.org/10.1145/1830483.1830584.

[110] P. Cremonesi, Y. Koren, R. Turrin, Performance of Recommender Algorithms on Top-N Recommendation Tasks Categories and Subject Descriptors, Proceedings of the fourth ACM conference on Recommender systems. (2010), New York, USA, pp. 9–46. https://doi.org/10.1145/1864708.1864721

[111] M. Laszczyk, P.B. Myszkowski, Survey of quality measures for multi-objective optimization: Construction of complementary set of multi-objective quality measures, Swarm and Evolutionary Computation. Vol. 48, (2019), pp. 109–133. https://doi.org/10.1016/J.SWEVO.2019.04.001.

[112] M.Z. Naser, A. Seitllari, Concrete under fire: an assessment through intelligent pattern recognition, Engineering with Computers. Vol. 36, (2019), pp. 1–14. https://doi.org/10.1007/s00366-019-00805-1.

[113] V. V. Degtyarev, Neural networks for predicting shear strength of CFS channels with slotted webs, Journal of Constructional Steel Research. Vol. 177, (2021). https://doi.org/10.1016/j.jcsr.2020.106443.

[114] L. De Lorenzis, J.G. Teng, Near-surface mounted FRP reinforcement: An emerging technique for strengthening structures, Composites Part B: Engineering. Vol. 38, (2007), pp. 119-143. https://doi.org/10.1016/j.compositesb.2006.08.003.

[115] Scikit, 4.1. Partial Dependence and Individual Conditional Expectation plots — scikit-learn 0.24.1 documentation, (2021). https://scikit-learn.org/stable/modules/partial_dependence.html (accessed February 20, 2021).

[116] H.Y. Zhang, H.R. Lv, V. Kodur, S.L. Qi, Performance comparison of fiber sheet strengthened RC beams bonded with geopolymer and epoxy resin under ambient and fire conditions, Journal of Structural Fire Engineering. Vol. 9, (2018), pp. 174–188.

https://doi.org/10.1108/JSFE-01-2017-0023.

[117] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: A methodology review, Journal of Biomedical Informatics. Vol. 35, (2002), pp. 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0.

[118] Discipulus, Discipulus Professional - G6G Directory of Omics and Intelligent Software, (2004). https://www.g6g-softwaredirectory.com/ai/genetic-programming/20047RMLTechnolDiscipulusProfess.php (accessed January 23, 2019).

[119] A. Blanco-Justicia, J. Domingo-Ferrer, Machine Learning Explainability Through Comprehensible Decision Trees, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2019, pp. 15-26. ISBN: 978-3-030-29725-1, https://doi.org/10.1007/978-3-030-29726-8_2.

[120] M.Z. Naser, Observational Analysis of Fire-Induced Spalling of Concrete through Ensemble Machine Learning and Surrogate Modeling, Journal of Materials in Civil Engineering. Vol. 33, (2021). https://doi.org/10.1061/(ASCE)MT.1943-5533.0003525.

[121] M. a. Hall, L. a. Smith, Practical feature subset selection for machine learning, Computer Science. Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth, 4-6 February, 1998, pp. 181-191. Berlin: Springer (1998). https://hdl.handle.net/10289/1512

[122] C. Kaibel, T. Biemann, Rethinking the Gold Standard With Multi-armed Bandits: Machine Learning Allocation Algorithms for Experiments, Organizational Research Methods. Vol. 24, (2021), pp. 78-103. https://doi.org/10.1177/1094428119854153.

[123] S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning, ArXiv. (2018). https://arxiv.org/abs/1811.12808

[124] F. Emmert-Streib, O. Yli-Harja, M. Dehmer, Explainable artificial intelligence and machine learning: A reality rooted perspective, Data Mining and Knowledge Discovery. (2020). https://doi.org/10.1002/widm.1368.

[125] R.R. Hoffman, G. Klein, S.T. Mueller, Explaining explanation for "explainable AI, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 62, (2018), pp, 197-201. https://doi.org/10.1177/1541931218621047.

[126] P. Hase, M. Bansal, Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (2020), pp. 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491.