Journal of Materials Science: Materials Theory

RESEARCH Open Access

A machine-learning-based model for predicting crack arrest fracture toughness through literature data mining

Gabriel Correa[1], Christos E. Athanasiou[2], Ting Zhu[3*] and Xing Liu[1*]

*Correspondence:
Ting Zhu
ting.zhu@me.gatech.edu
Xing Liu
xing.liu@njit.edu
[1]Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA
[2]Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[3]George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

Machine learning (ML) is transforming fracture mechanics research by offering unprecedented capabilities for analyzing high-throughput, multi-modal, and multi-fidelity data. Recent work has generated large amounts of new fracture data through experiments and simulations and extracted useful insights from them. However, the vast body of fracture data accumulated over the past century remains largely untapped, primarily because it is scattered throughout the literature without systematic organization. Our central hypothesis is that these historical data contain valuable insights that can be unlocked using ML. To test this hypothesis, we present a case study on crack arrest fracture toughness, a critical yet poorly understood property in fracture mechanics. We compiled a comprehensive dataset of crack arrest toughness for a wide range of structural steels through a thorough literature review and analyzed it using neural network (NN)-based methods, including feature-importance assessment and high-dimensional regression. Our analysis shows that elements such as carbon and manganese exert a stronger influence on crack arrest toughness than others such as copper, and that temperature also plays a critical role. We further developed an NN-based model capable of predicting crack arrest toughness from these factors with an error of 11.8%. This study demonstrates the substantial opportunities for advancing fracture mechanics by mining the vast body of historical literature data, while also highlighting the challenges associated with their fragmented and multi-fidelity nature.

Keywords Crack arrest fracture toughness, Machine learning, Structural steels

Introduction

Fracture is ubiquitous in natural and engineered materials, and understanding it has long been essential for innovations in material design and manufacturing. Since the landmark theory of Griffith in 1921 (Griffith 1921), the field of fracture mechanics has seen tremendous advances, including the statistical theory of fracture by Weibull (Weibull 1951), the path-independent integrals by Rice (Rice 1968), the HRR field by Hutchinson, Rice, and Rosengren (Hutchinson 1968; Rice and Rosengren 1968), and the probabilistic theory of quasi-brittle fracture by Bažant (Bažant and Le 2017), among others. More

recently, machine learning (ML) has opened new opportunities for advancing fracture mechanics even further.

The impact of ML now spans the entire cycle of fracture research. For fracture detection, ML has been used to analyze mechanical signals, acoustic emissions, and imaging data for nondestructive evaluation and high-resolution reconstruction of crack nucleation and growth (Zhang et al. 2022; Shukla et al. 2020; Niu and Srivastava 2022; Shen et al. 2021; Jones et al. 2020; Müller et al. 2021). For fracture analysis and prediction, ML has accelerated the extraction of fracture properties from experiments (Liu et al. 2020; Athanasiou et al. 2022), uncovered governing criteria for small-fatigue-crack propagation (Rovinelli et al. 2018), and enabled topological analysis of dynamic fracture networks (Cacas et al. 1990; Hyman et al. 2015; Valera et al. 2018; Srinivasan et al. 2018). At the atomistic scale, ML models informed by molecular simulations have achieved rapid prediction of dynamic crack growth (Lew et al. 2021; Hsu et al. 2020). For fracture manipulation, ML has accelerated the exploration of the vast material design space and enabled the discovery of tougher composites and super-stretchable graphene kirigami (Gu et al. 2018a, 2018b; Guo and Buehler 2020; Hanakata et al. 2018). Collectively, these developments demonstrate the transformative potential of ML for advancing fracture mechanics, particularly when sufficiently rich datasets are available.

A notable common thread among these advances is that they are all propelled by the surge in generating new "Big Fracture Data" through high-throughput experiments and simulations. However, the vast body of fracture data accumulated over the past century remains largely untapped. Although historical fracture data are scattered across journal articles, technical reports, and industry documents and often exist in diverse formats with wide variations in quality and fidelity, valuable fracture mechanics knowledge is embedded within them. Our central hypothesis is that unlocking this underutilized reservoir of knowledge through ML represents an important frontier for the field.

To test the above hypothesis, we present a case study on crack arrest fracture toughness, aiming to develop a predictive model from the data available in the literature. Crack arrest toughness is a critical yet understudied material property that characterizes a material's ability to stop a running crack. While crack initiation toughness has dominated research and design practice, crack arrest toughness plays an equally important role in ensuring structural safety, as arresting a crack before catastrophic failure can be the last line of defense in large-scale structures such as pipelines, ships, reactors, and energy storage devices (Taylor et al. 2020). Despite its importance, the underlying factors that dictate a material's crack arrest toughness remain poorly understood.

In this case study, we compiled a comprehensive dataset of crack arrest fracture toughness through an extensive review of historical literature. The dataset includes measured crack arrest toughness for a wide range of structural steels, along with the corresponding chemical compositions, processing conditions, test methods, and temperatures. Using ML techniques, we assessed the relative importance of these factors and found that alloying elements such as C and Mn play a dominant role in determining crack arrest toughness. Temperature was also identified as a critical governing factor. Building on these insights, we developed an ML-based model that predicts crack arrest toughness from composition and temperature with a reasonable prediction error of 11.8%, despite the substantial variability and multi-fidelity nature of the underlying literature data.

Data collection from a comprehensive literature review

The first documented experimental measurement of crack arrest fracture toughness was conducted by Nordell and Hall in 1965 on thick, wide plates of welded steel (Nordell and Hall 1965). In the following decades, various test procedures were explored, and efforts toward standardization evolved along two main paths: wide-plate configurations (Nordell and Hall 1965) and compact-specimen configurations (Gehlen et al. 1979). Thanks to the work of pioneers such as Crosley, Ripling, Hahn, Hoagland, and Irwin (Crosley and Ripling 1981; Crosley et al. 1983; Barker et al. 1988), a standardized test method, ASTM E1221, was established in 1988 based on a compact-specimen design (ASTM International 2023). Parallel developments using wide-plate specimens were carried out primarily in Japan, culminating in two equivalent standards: WES 2815 in 2014 (The Japan Welding Engineering Society 2815) and ISO 20064 in 2019 (International Organization for Standardization 2019). A comprehensive historical review of these developments was provided by Gudas (Gudas 1987) and Yanagimoto (Yanagimoto et al. 2025). Our survey of a substantial body of technical reports and publications indicates that most reported crack arrest toughness data were generated using ASTM-like or ISO-like test methodologies.

Standard methods for measuring crack arrest fracture toughness

ASTM E1221 specifies a laboratory-scale test method for determining plane-strain crack arrest fracture toughness of ferritic steels at a given temperature using an edge-notched compact specimen (Fig. 1a). The crack-starter notch is loaded through a wedge-loading system, where the applied wedge-force is cyclically ramped until a fast-running, unstable

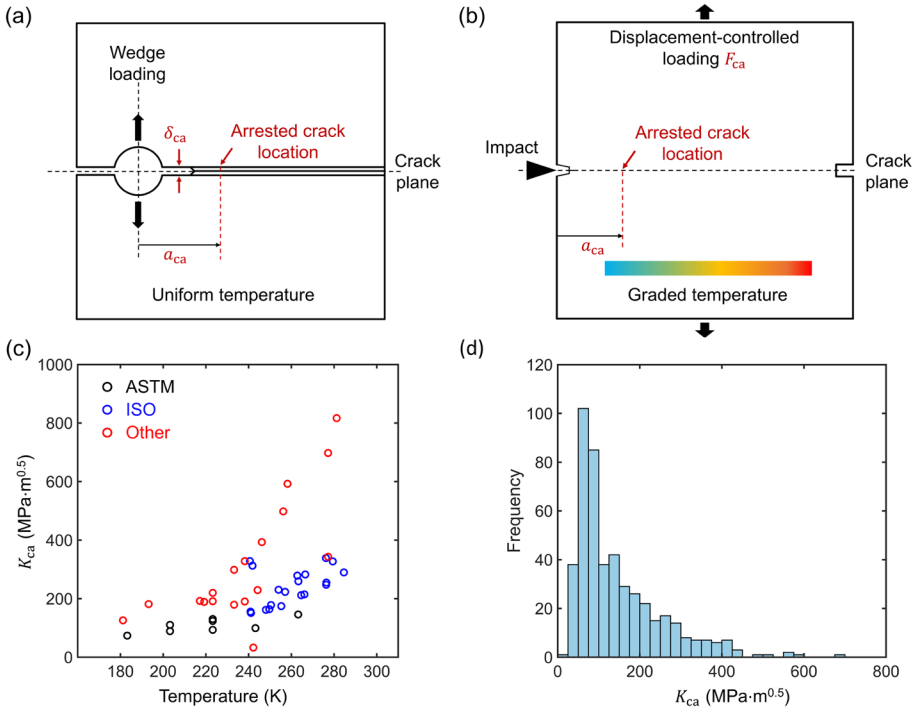

Fig. 1 Experimental measurement of crack arrest fracture toughness. **a** Schematic of the compact specimen configuration specified in ASTM E1221. **b** Schematic of the standard wide-plate specimen configuration specified in ISO 20064. **c** Crack arrest toughness of EH47 steels measured using different test methods over a range of temperatures. **d** Long-tailed distribution of crack arrest toughness values for a wide range of steels in the compiled dataset

crack initiates. Upon crack initiation, the wedge force drops as the crack-mouth opening widens, causing rapid crack arrest and producing a run-arrest segment of crack extension. During the test, the specimen is maintained at a predetermined temperature without a temperature gradient. The wedge force and the crack-month opening displacement (CMOD) at a prescribed offset from the load line are recorded, from which the instant of crack arrest is identified. The crack arrest fracture toughness is calculated from the critical CMOD, δ_{ca}, measured with appropriate corrections shortly after crack arrest, and the final arrested crack size, a_{ca}, determined by breaking the specimen and examining the exposed fracture surfaces.

ISO 20064 provides a large-scale test method for determining the crack arrest fracture toughness of steels using edge-notched wide plates with a temperature gradient (Fig. 1b). At the beginning of the test, the specimen is loaded to a predetermined stress level and held through a displacement-controlled load cell, while a predetermined temperature gradient is established across the specimen. An impact is then applied to the notch via a wedge mounted on it to initiate a brittle crack and produce a run-arrest segment of crack extension. The force applied to the specimen by the load cell at the time of crack initiation, F_{ca}, is recorded. The arrested crack length, a_{ca}, is measured from the photographs of fracture surfaces and crack propagation path, and the arrest temperature at the location of crack arrest, T_{ca}, is recorded. The crack arrest fracture toughness at the temperature, T_{ca}, is calculated from F_{ca} and a_{ca}.

Both test standards impose stringent requirements on specimen dimensions and post-test validation to ensure compliance with linear elastic fracture mechanics (LEFM) analysis and plane-strain conditions. ISO 20064 generally requires wider and thicker specimens to evaluate the arrest toughness of steel plates under service conditions, while ASTM E1221 employs more compact specimens and is considered to provide the lower-bound toughness. Both standards produce a crack run-arrest event in the specimen, through impact loading in ISO 20064 and a progressively increasing load in ASTM E1221. Special consideration must be given to the control parameters in these crack-initiation approaches to minimize their influence on the arrest process. Both standards require temperature control: ASTM E1221 maintains a uniform temperature across the specimen, while ISO 20064 establishes a moderate temperature gradient along the crack extension direction but is intended to obtain toughness values comparable to those measured without a temperature gradient. Both standards rely on static LEFM solutions to calculate the crack arrest toughness. However, the validity of such estimations requires further investigation, such as through dynamic finite element analysis.

Our review of the two test standards clearly shows that measuring crack arrest fracture toughness is both costly and labor intensive. On the one hand, the specially designed test apparatuses integrate sophisticated thermal and mechanical components for controlled loading and monitoring and demand substantial capital investment, especially when large-specimen tests are required. On the other hand, the comprehensive post-test examination, which involves numerous validation criteria, limits the overall success rate of these tests. The trial runs conducted prior to obtaining a valid test further increase the overall expense of crack arrest toughness measurements. Therefore, the crack arrest toughness data reported in the literature are highly valuable and should be utilized effectively.

Dataset compilation

Our central criterion in compiling the dataset is to maximize data quantity without sacrificing quality. Therefore, we retained all the data in the collected documents, including crack arrest fracture toughness, chemical composition (*i.e.*, weight fractions of alloying elements: C, Mn, Si, P, S, Cu, Ni, Cr, Nb, Ti, Al, V, Mo, N, and B), processing condition (*e.g.*, heat treatment), test method, and arrest temperature, except those obtained from non-standard tests that deviate substantially from ASTM or ISO standards. These data appeared in various forms, such as numerical values, text descriptions, tables, and figures, and were all numerically encoded before being included in the dataset. Incomplete entries were supplemented with referenced information or domain knowledge whenever possible. Selected data are presented in Table 1, which summarizes the 18 features $\{x_i, i = 1,2,\ldots,18\}$ and the label $y = K_{\mathrm{ca}}$ in the dataset. The statistics of the features and label are presented in Table 2. The full dataset consists of 473 data points drawn from 21 sources and is available online along with the corresponding references (Liu and Gabriel 2026).

We made several notable observations while compiling the dataset. First, significant variations were found in crack arrest toughness measurements for the same material when different test methods were used, and even among repeated measurements of the same material using the same method (Fig. 1c). Second, the measured arrest toughness values span a wide range and exhibit a long-tailed distribution (Fig. 1d). These observations indicate that uncertainties on the order of 100% in measured or predicted crack arrest fracture toughness are not uncommon, and that some materials can exhibit exceptionally high crack arrest toughness.

Table 1 Selected data in the dataset

Data No		1	2	3	4	5
Features	C%	0.08	0.17	0.25	0.065	0.16
	Mn%	1.27	1.07	1.25	1.495	1.5
	Si%	0.325	0.23	0.325	0.16	0.4
	P%	0.023	0.018	0.035	0.008	0
	S%	0.007	0.012	0.04	0.002	0
	Cu%	0.015	0.2	0.35	0	0
	Ni%	0.01	0.1	0.25	0	0.25
	Cr%	0.04	0.13	0.25	0	0.35
	Nb%	0.033	0	0	0.05	0.05
	Ti%	0	0	0	0.017	0.05
	Al%	0.044	0.043	0	0.035	0
	V%	0.034	0.055	0.08	0.013	0.09
	Mo%	0.005	0	0.08	0	0.4
	N%	0	0.007	0	0.0045	0
	B%	0	0	0	0	0.005
	Processing condition*	3	1	2	3	0
	Test Method**	0	0	0	1	0
	Temperature (K)	273.15	283.15	227.17	184.55	213.15
Label	K_{ca} (MPa m$^{0.5}$)	129.0	133.0	59.0	289.7	48.0

*0 = quenching, 1 = controlled rolling, 2 = normalizing, 3 = unspecified

**0 = ASTM-like, 1 = ISO-like

Table 2 Statistics of the features and label

			Range (Raw Data)	Mean	Standard Deviation	Range (After Z-Score Normalization)
Features	C%	x_1	0.0500–0.4300	0.1798	0.1077	−1.2045–2.3229
	Mn%	x_2	0.2600–1.9500	1.0965	0.4027	−2.0776–2.1196
	Si%	x_3	0.0200–0.4900	0.2677	0.0979	−2.5292–2.2695
	P%	x_4	0.0000–0.0350	0.0167	0.0107	−1.5642–1.7149
	S%	x_5	0.0000–0.0400	0.0168	0.0127	−1.3171–1.8217
	Cu%	x_6	0.0000–1.6100	0.1308	0.3690	−0.3545–4.0083
	Ni%	x_7	0.0000–9.0000	0.6603	1.7800	−0.3710–4.6853
	Cr%	x_8	0.0000–2.0000	0.3584	0.6022	−0.5952–2.7259
	Nb%	x_9	0.0000–0.0600	0.0070	0.0165	−0.4233–3.2074
	Ti%	x_{10}	0.0000–0.0500	0.0018	0.0078	−0.2250–6.1625
	Al%	x_{11}	0.0000–0.0440	0.0022	0.0087	−0.2489–4.8139
	V%	x_{12}	0.0000–0.1100	0.0147	0.0310	−0.4758–3.0771
	Mo%	x_{13}	0.0000–0.6080	0.1319	0.2122	−0.6243–2.2545
	N%	x_{14}	0.0000–0.0150	0.0009	0.0026	−0.3231–5.3674
	B%	x_{15}	0.0000–0.0050	0.0001	0.0007	−0.1468–6.7972
	Processing condition*	x_{16}	0,1,2,3	/	/	/
	Test Method**	x_{17}	0,1	/	/	/
	Temperature (K)	x_{18}	77.1500–438.7100	262.6261	72.3352	−2.5641–2.4343
Label	K_{ca} (MPa m$^{0.5}$)	y	24.1000–685.000	145.1465	105.6247	/

Knowledge extraction using ML

It is standard practice to preprocess a dataset before conducting ML analysis. Here, we applied Z-score normalization (Fisher 1925) to rescale the composition and temperature features to a consistent range, while leaving the label, K_{ca}, unscaled due to its long-tailed distribution. Given that the original dataset contains 18 features, an important question arises: do all of these features play an equally important role in determining K_{ca}? To address this question, we evaluated the importance of each feature based on the dataset.

Quantification of feature importance

We employed an iterative method, Random Subset Feature Selection (RSFS) (Breiman 2001), to quantify feature importance (Fig. 2). In each iteration, a subset of N_{RSFS} features was randomly selected from the full set of 18 features, and the performance of this subset was evaluated using the predictive accuracy of a neural network (NN) trained on the selected features and the label. This procedure was repeated 100 times, and the importance of each feature was determined by its frequency of appearance among the top 10 high-performing subsets.

In evaluating subset performance, we used fully connected NNs with the rectified linear unit (ReLU) activation function, consisting of an input layer with N_{RSFS} nodes, two hidden layers with n_1 and n_2 nodes, and an output layer with a single node, denoted as $N_{\mathrm{RSFS}}/n_1/n_2/1$. A total of 200 NNs with $n_1 = n_2 = 8$ were trained in TensorFlow/2.20 (Abadi, et al. 2015) using the "mean squared error" loss function and the Nadam algorithm. Training was performed on 70% of the full dataset, with the remaining dataset split into validation (15%) and test sets (15%). Each NN was trained for 5000 training iterations with a learning rate of 0.01, followed by another 5000 fine-tuning iterations with a reduced learning rate of 0.001. This two-stage training protocol has proven effective for training simple fully connected NNs (Liu et al. 2021). Overfitting was assessed by

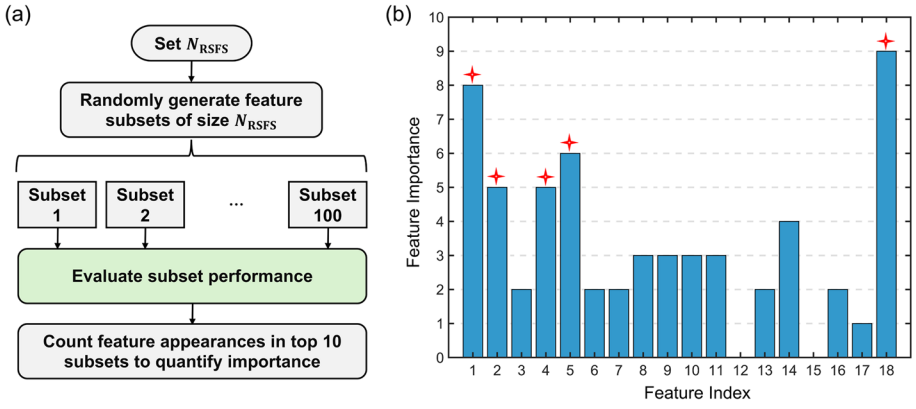

Fig. 2 RSFS-based quantification of feature importance. **a** Flowchart of the RSFS algorithm, integrated with NN-based evaluation of subset performance (highlighted in green). **b** Feature importance was quantified based on the frequency of each feature's appearance among the top-performing subsets of size $N_{\mathrm{RSFS}} = 6$. The five most important features are identified as x_1, x_2, x_4, x_5, and x_{18} (highlighted by red crosses), corresponding to the weight fractions of C, Mn, P, S, and temperature

monitoring prediction errors on the training and validation datasets throughout training and by comparing the final training error with that obtained on the independent test dataset. No evident overfitting was observed; therefore, an early stopping mechanism was not employed. The NN achieving the lowest overall prediction error was selected, and its prediction error served as the performance measure for the corresponding feature subset. Note that, throughout this work, the mean absolute percentage error (MAPE) was used as the metric for evaluating model accuracy.

Starting with $N_{\mathrm{RSFS}} = 6$, we found that x_1, x_2, x_4, x_5, x_{18} appeared most frequently among the top 10 high-performing subsets, corresponding to the weight fractions of C, Mn, P, S, and temperature (Fig. 2b). These 10 high-performing subsets exhibit an average prediction error of 20.9%, which is significantly lower than that of the remaining 90 subsets (32.7%). We further evaluated the subset containing only these five features and obtained an error of 20.7%, confirming that x_1, x_2, x_4, x_5, x_{18} are indeed the five most important features. Therefore, the final set of features exhibiting most significant effect on K_{ca} was identified as $\{x_1, x_2, x_4, x_5, x_{18}\}$. The weight fractions of V and B (x_{12} and x_{15}) were found to contribute least to the predictive performance.

Although simple N_{RSFS}/8/8/1 NNs were employed as the performance evaluator for random feature subsets, RSFS is inherently model-agnostic, and similar results are expected when using more complex NNs or other regression models, such as regression trees. Simple NNs were adopted instead of regression trees because of their architectural simplicity. As demonstrated in our prior work (Liu et al. 2020), simple NNs with tens of nodes can achieve predictive accuracy comparable to that of tree-based models containing thousands of leaf nodes. More complex NN architectures with additional layers or nodes were not employed in order to limit training cost, which is particularly important given the large number of random feature subsets evaluated in the RSFS procedure.

To further strengthen the RSFS results, we employed a model-free, nonparametric approach based on mutual information (MI) to re-evaluate the importance of each feature (Shannon 1948). Specifically, for each feature, we estimated its MI with the label using the *k*-nearest neighbors algorithm implemented in *scikit-learn* (Kraskov et al. 2004; Pedregosa et al. 2011). The resulting MI values serve as quantitative measures of feature importance. This analysis relies solely on the observed data and does not involve

fitting any regression models. Figure 3 presents the MI results for $k = 2$–5, which show strong agreement with the RSFS results. The features identified by RSFS as high-importance, $\{x_1, x_2, x_4, x_5, x_{18}\}$, also rank among those with the highest MI values. Conversely, the feature identified by RSFS as low-importance, x_{15} (corresponding to the weight fraction of B), exhibits the lowest MI value. This strong consistency between the two independent methods provides additional support for the validity and robustness of the RSFS-based feature-importance results.

ML-based predictive model

The next question is whether K_{ca} can be reliably predicted from the identified high-importance features. We adopted fully connected NNs as the base predictive model to learn the dependence of K_{ca} on these features, because they represent the simplest class of NN architectures. We started with networks comprising 2 hidden layers, since for a given total number of nodes, a two-layer network generally provides greater representational capacity than a single-layer network. We increased the number of nodes in each hidden layer from 8 (used for RSFS in "Quantification of feature importance" section) to 16 because the 5/8/8/1 architecture yielded unsatisfactory predictive accuracy (20.7% MAPE). A total of 1000 5/16/16/1 NNs were trained on the reduced dataset $\{x_1, x_2, x_4, x_5, x_{18}; y\}$ using the same training procedure described in "Quantification of feature importance" section. It was found that the best-performing NN still exhibits a prediction error of 14.4%. Although this falls within our tolerance of 20%, further

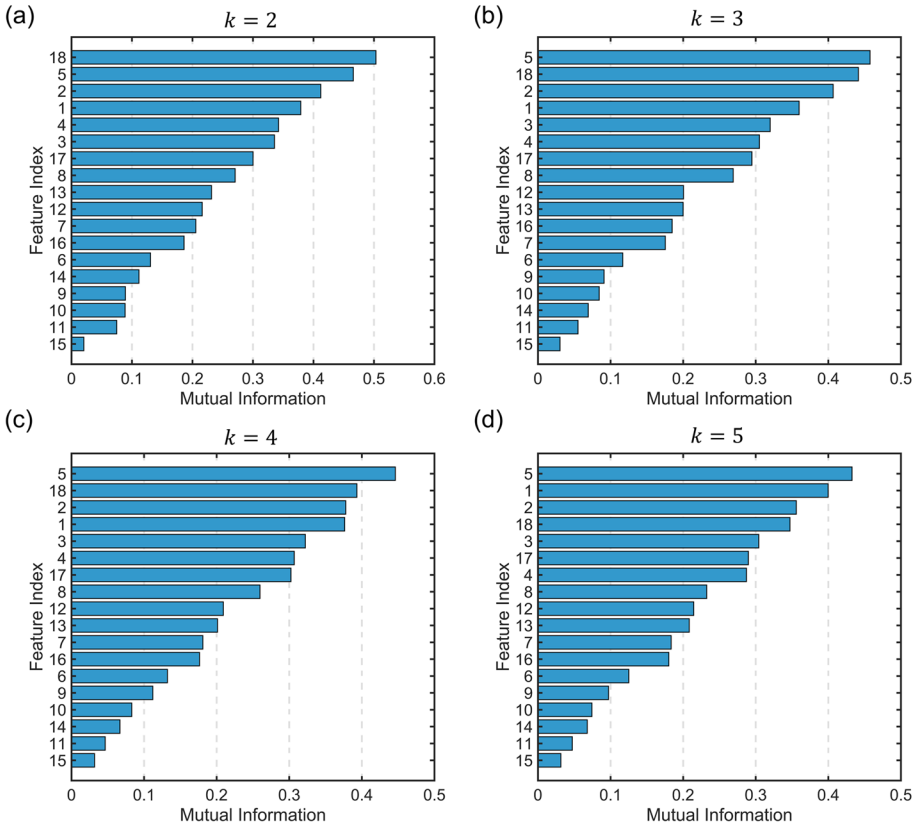

Fig. 3 MI-based feature-importance analysis. The 18 features in the compiled dataset are ranked according to their MI with the label. **a**–**d** show the results obtained using the *k*-nearest neighbors algorithm with $k = 2$, 3, 4, and 5, respectively

improvement is desirable, motivating the search for a more accurate predictive model. However, simply increasing the number of nodes within a single NN led to overfitting.

We attributed this suboptimal performance of the simple NNs to the long-tailed distribution of K_{ca}, as such models tend to bias toward the middle of the distribution and struggle to learn patterns associated with rare, extreme values. To address this issue, we combined two NNs, one specialized for the dominant moderate-K_{ca} regime and the other for the sparse high-K_{ca} tail, to jointly predict K_{ca}. These two regression NNs, together with a binary NN classifier that directs predictions to the appropriate regime, constitute a composite NN model for the final prediction (Fig. 4a).

To develop the binary classifier, we augmented the dataset with an auxiliary variable θ, where $\theta = 0$ for $K_{\mathrm{ca}} < K^*$ and $\theta = 1$ otherwise. The threshold K^* was set to the sum of the mean and standard deviation of K_{ca}, yielding $K^* = 251\mathrm{MPa} \cdot \mathrm{m}^{0.5}$. This value corresponds approximately to the upper tail of the K_{ca} distribution in the dataset (Fig. 1d) and provides a natural statistical separation between moderate- and high-K_{ca} regimes. A total of 1000 NNs with an 5/8/8/1 architecture were trained on the augmented dataset $\{x_1, x_2, x_4, x_5, x_{18}; \theta\}$ using a training procedure similar to that described in "Quantification of feature importance" section, except that the binary cross-entropy loss function was used. The best-performing classifier achieved an accuracy of 96.0%, with

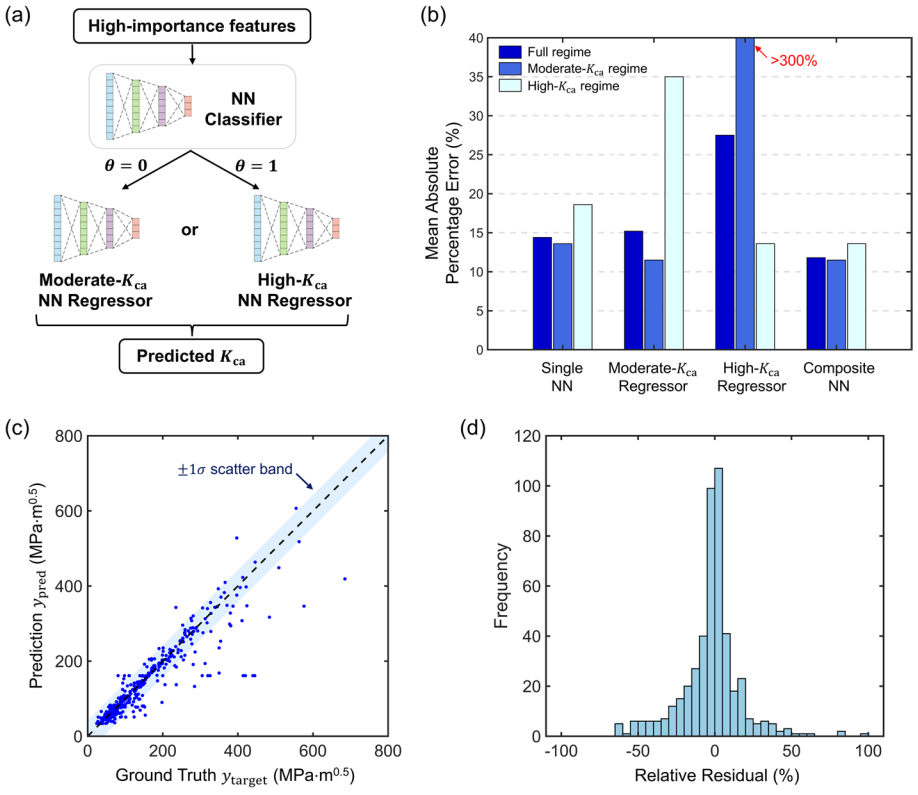

Fig. 4 Predicting crack arrest fracture toughness using a composite NN. **a** Architecture of the composite predictive model, consisting of a binary NN classifier and two NN regressors trained for the moderate-K_{ca} and high-K_{ca} regimes, respectively. **b** Comparison of the prediction performance of the single NN (trained on the full dataset without regime separation), the moderate-K_{ca} NN regressor, the high-K_{ca} NN regressor, and the composite NN, evaluated over the full regime, the moderate-K_{ca} regime, and the high-K_{ca} regime. **c** Predictions from the composite model versus ground-truth values from the compiled experimental dataset. The dashed line denotes perfect agreement, and the shaded scatter band represents the standard deviation of the residuals, $\sigma = 42.2$ MPa $\mathrm{m}^{0.5}$. **d** The distribution of relative residuals indicates negligible systematic bias

misclassifications occurring mainly near the transition boundary between the moderate-K_{ca} and high-K_{ca} regimes. The high accuracy indicates that the selected 5/8/8/1 architecture is appropriate, and increasing the number of nodes is unnecessary.

To develop the regression NN for the moderate-K_{ca} regime, we retained only the data with $\theta = 0$ and those misclassified by the NN classifier in the training, validation, and test datasets. Using the same procedure described in "Quantification of feature importance" section, we trained 1000 NNs with a 5/16/16/1 architecture and selected the best-performing one. The resulting moderate-K_{ca} regressor yields prediction errors of 11.6% and 13.1% on the training and test datasets, respectively, indicating no evidence of overfitting or underfitting. Similarly, for the high-K_{ca} regime, we developed a regression NN using the data with $\theta = 1$ along with those misclassified. The corresponding prediction errors on the training and test datasets are 14.6% and 16.0%, respectively. Finally, we assembled the two NN regressors together with the NN classifier to form a composite model.

A comparison of the performance of the single NN (trained without distinguishing between the moderate-K_{ca} and high-K_{ca} regimes), the moderate-K_{ca} regressor, the high-K_{ca} regressor, and the composite mode, is shown in Fig. 4b. The single NN exhibits prediction errors of 13.6% in the moderate-K_{ca} regime and 18.6% in the high-K_{ca} regime. Each specialized regressor performs well within its designated regime but degrades significantly when applied outside it. The moderate-K_{ca} regressor reduces the prediction error in the moderate-K_{ca} regime to 11.5%, and the high-K_{ca} regressor reduced the prediction error in the high-K_{ca} regime to 13.6%, both showing substantial improvements over the single NN. The composite model preserves the strengths of the two specialized NNs, thereby improving the overall predictive accuracy and reducing the error to 11.8%.

To further evaluate the performance of the composite model, we directly compared its predictions, y_{pred}, with the ground-truth experimental data, y_{target}, in Fig. 4c. The standard deviation of the residuals ($y_{\mathrm{pred}} - y_{\mathrm{target}}$) is 42.2 MPa $\mathrm{m}^{0.5}$. Around 88% of the data points fall within the $\pm 1\sigma$ scatter band, indicating that the composite model captures the overall trend of the experimental data with limited dispersion. Figure 4d shows the distribution of relative residuals, $\left(y_{\mathrm{pred}} - y_{\mathrm{target}}\right)/y_{\mathrm{target}}$, which is approximately symmetric about zero, with a mean relative residual of 2.2%. This small mean value indicates negligible systematic overprediction or underprediction.

Open-access deployment

We prioritized model simplicity in developing the predictive model for K_{ca}, as simplicity is essential for open-access deployment. The trained NNs were exported as open-standard JavaScript Object Notation (JSON) files and deployed through a web-based application with a graphical user interface, named "Crack Arrest Fracture Toughness Predictor" (Liu 2025). This open-access website enables users to input steel composition and arrest temperature to obtain real-time predictions of K_{ca}. By releasing our models in an accessible and extensible format, we aim to facilitate broader collaboration among researchers and engineers in the fracture mechanics community, promote transparency and reproducibility, and establish a foundation for community-driven refinement of predictive models for crack arrest fracture toughness.

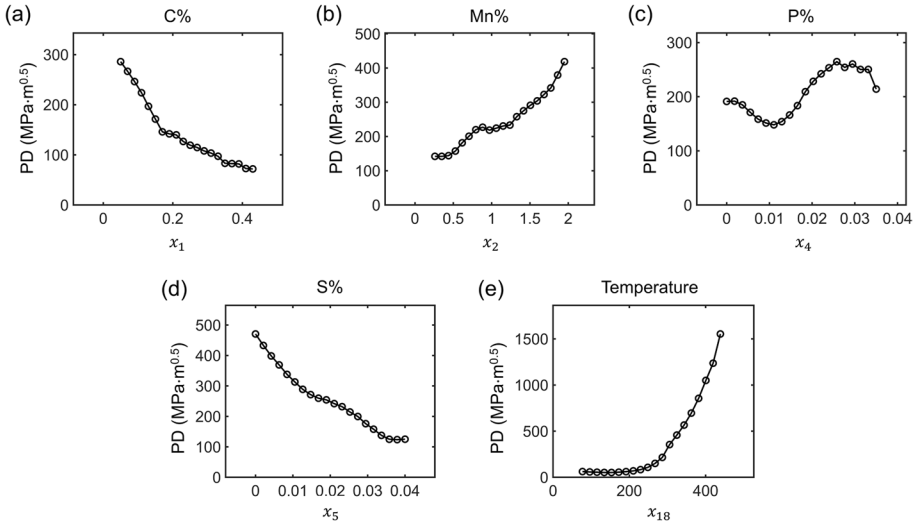

Fig. 5 Partial dependence plots. **a**–**e** illustrate the effects of C, Mn, P, and S, and temperature on K_{ca}, respectively

Discussion and concluding remarks

Opportunities in crack arrest fracture toughness

We chose crack arrest fracture toughness as our case study because, despite its long-recognized importance in the nuclear and shipbuilding sectors, it has received limited attention in recent academic efforts to develop new materials. Moreover, its distinction from the widely studied initiation toughness is rarely emphasized in contemporary fracture mechanics research. To date, no experimental or theoretical evidence supports a direct correlation between crack arrest toughness and initiation toughness, underscoring the need to investigate crack arrest toughness as an independent physical property. Most available data on crack arrest toughness remain confined to classical material systems, such as ferritic steels. Through this case study, we aim to introduce the concept of crack arrest toughness to a broader audience and motivate its extension to a wider range of materials, including complex concentrated alloys, metamaterials, and composites.

Physical interpretation of the ML-based predictive model

According to our feature-importance analysis, the crack arrest toughness of steels is highly sensitive to the fractions of certain alloying elements, including C, Mn, P, and S, as well as to temperature. Thus, adjusting the levels of these elements may provide a viable pathway for tuning crack arrest toughness. The strong temperature dependence is unsurprising, given that both plasticity and fracture are thermally activated processes. In contrast, it may appear counterintuitive that processing conditions were not identified as a highly important factor. This outcome arises from the fact that most entries (~ 80%) list processing conditions as "unspecified," preventing a statistically significant assessment of their effect. Therefore, including such microstructure-related information in future experimental and computational studies of crack arrest toughness is critical.

To interpret the dependence of K_{ca} on the selected features, we present partial dependence (PD) plots for $x_1, x_2, x_4, x_5, x_{18}$ in Fig. 5 (Hastie et al. 2009). The PD results indicate that (1) increasing C and S contents decreases K_{ca}; (2) increasing Mn content and temperature increases K_{ca}; and (3) P exhibits a nonmonotonic effect on K_{ca}. It should be noted that these trends reflect the behavior captured by the developed predictive

model and may not fully represent the underlying physics, owing to limitations in the available K_{ca} dataset.

Carbon in steels is primarily present in the form of carbides distributed within grains or along grain boundaries. These hard, brittle particles interact with dislocations, leading to precipitation strengthening and hardening. Consequently, increasing the C content generally suppresses plastic deformation and reduces ductility in steels, thereby lowering fracture toughness. Moreover, fracture may nucleate within carbides or at carbide-matrix interfaces, further degrading toughness.

Regarding the effect of Mn content, manganese is well known to enhance both strength and toughness in steels. A notable example is Hadfield steel (also known as Mangalloy). Mn mitigates the detrimental effects of S by promoting the formation of MnS instead of FeS, thereby reducing hot shortness. In addition, Mn contributes to grain refinement by lowering the austenite-to-ferrite transition temperature, which is beneficial to toughness.

Phosphorus segregates to grain boundaries, weakening intergranular cohesion and promoting cold shortness. Accordingly, increasing the P content is expected to reduce fracture toughness. The predictive model does not fully capture this trend, likely due to limitations of the compiled dataset.

Sulfur is widely regarded as a highly detrimental impurity in structural steels. The formation of low-melting-point FeS promotes the development of thin, brittle films along grain boundaries, leading to hot shortness. Therefore, increasing the S content reduces fracture toughness.

The temperature effect reflects the well-known ductile-to-brittle transition in structural steels. At high temperatures (upper shelf), steels fail predominantly through ductile fracture mechanisms, such as microvoid coalescence, and exhibit high fracture toughness. Near the transition temperature, fracture toughness decreases rapidly as temperature drops. At low temperatures (lower shelf), brittle cleavage fracture prevails and toughness is low. Figure 5(e) likely captures the lower shelf and transition regimes but does not fully represent the upper shelf regime. The model predictions at high temperatures appear overestimated, likely due to the limited availability of upper-shelf data in the compiled dataset.

Opportunities and challenges of mining historical data

This study demonstrates that mining archival fracture literature can support the sustainable and cost-effective advancement of fracture mechanics. First, historical fracture data provide a valuable foundation for new studies and help guide future research directions. In particular, our analysis of historical K_{ca} data reveals underexplored dependencies of K_{ca} on specific alloying elements. Fully understanding and harnessing such dependencies to enable K_{ca}-based materials design motivates more systematic investigations in the future. Second, leveraging existing fracture datasets reduces the need for repeated experiments, lowering both cost and labor requirements. This advantage is especially important for the present work, as experimental measurement of K_{ca} requires substantial investment in specialized equipment, specimen preparation, and personnel training (see "Standard methods for measuring crack arrest fracture toughness" section).

The challenges of using historical data are also evident, as we have limited control over data quantity, quality, availability, fidelity, and sampling structure. Substantial effort

was therefore devoted to mitigating these limitations during the compilation of the K_{ca} dataset. Regarding data quantity, historical data are often not collected in a systematic manner, and the available sample size may be insufficient to characterize complex high-dimensional relationships. To address this limitation, we conducted a comprehensive literature review to identify as many relevant data sources as possible and reduced the dimensionality of the regression task through feature-importance analysis. Regarding data quality, historical records may exhibit inconsistency in reporting standards and miss critical information, such as material composition. When possible, we supplemented incomplete entries using referenced sources or domain knowledge; otherwise, such data were excluded to maintain overall data quality. Regarding data availability, historical data may appear in scattered and diverse forms (*e.g.*, tables, charts, figures) and are not always readily accessible. We therefore manually digitized and systematically organized all available data entries for dataset construction. Regarding data fidelity, historical data may have been produced using outdated or technically limited experimental approaches, potentially affecting data fidelity. We therefore retained only data obtained using procedures consistent with ASTM or ISO standards and excluded data from non-standard testing methods. Regarding sampling structure, historical records may exhibit heterogeneous sampling of the feature space, potentially resulting in underrepresentation of certain feature dependencies. In the compiled K_{ca} dataset, although all features vary across different sources, temperature is, in most cases, the only feature that varies within an individual source. This limited within-source variability of non-temperature features may therefore lead to an underestimation of their importance compared to temperature.

In addition to these limitations, historical data may contain substantial and incompletely characterized variability, arising from both true experimental scatter (*i.e.*, variability in repeated measurements under identical conditions for the same material) and systematic inter-source variability. This variability is often difficult to quantify due to the lack of uncertainty reporting based on repeated or benchmark tests. For some individual sources in our K_{ca} dataset, groups of repeated measurements are available, allowing us to estimate within-source scatter by calculating their standard deviation. The resulting standard deviations span a wide range, from 2.1 MPa $\mathrm{m}^{0.5}$ to 141.3 MPa $\mathrm{m}^{0.5}$, suggesting the presence of inter-source variability. However, repeated measurements across different sources are largely unavailable, limiting our ability to quantify this systematic inter-source variability. As a result, the overall variability of the compiled K_{ca} dataset cannot be reliably characterized and thus cannot provide a quantitative basis for explaining the observed scatter in the model predictions (Fig. 4c, d), which may partially arise from this underlying data variability.

Collectively, these limitations constrain the intrinsic informativeness of the compiled dataset, thereby imposing an upper bound on the performance of purely data-driven predictive models unless additional data or physical insights are incorporated. In the case study of K_{ca}, we developed a composite NN to capture the dependence of K_{ca} on the weight fractions of C, Mn, P, S, and temperature, as the dataset contains richer and more consistent information on the effects of these selected features than on those of the remaining ones. The absence of other features, such as processing conditions, does not imply that they are physically insignificant; rather, it likely reflects the limited informativeness of the dataset. In other words, data-driven predictions do not necessarily

provide a complete representation of a high-dimensional nonlinear physical relationship; instead, they reflect the latent structure embedded in the available data, which may constitute a partial realization of the underlying physics.

This study highlights both the opportunities and challenges of mining archival fracture literature. On one hand, historical data contain rich and otherwise inaccessible insights that can advance fracture mechanics. On the other hand, historical data are rarely presented in an organized or machine-readable format. Their fragmented, multimodal, and multi-fidelity nature demands careful curation and preprocessing prior to knowledge extraction, underscoring the value of community-wide efforts to standardize fracture mechanics data reporting and archival practices. The reasonably good performance of the predictive model for K_{ca}, developed based on historical data, demonstrates the promise of combining ML with historical fracture data to deepen understanding of fracture phenomena in complex material systems and to pave the way for data-driven design of more fracture-resistant materials.

Abbreviations

ML	Machine learning
CMOD	Crack-month opening displacement
LEFM	Linear elastic fracture mechanics
RSFS	Random subset feature selection
NN	Neural network
ReLU	Rectified linear unit
MAPE	Mean absolute percentage error
MI	Mutual information
JSON	JavaScript object notation

Acknowledgements
Not applicable.

Authors' contributions
GC and XL conceptualized the study. XL and TZ supervised the study. GC and XL compiled the dataset. GC, XL, and CEA performed the data analysis. GC, XL, and TZ drafted the original manuscript. All authors reviewed, revised, and approved the final manuscript.

Funding
GC acknowledges the support of the NJIT Undergraduate Research and Innovation (URI) Summer Fellowship Program (2025) and the New Jersey Space Grant Consortium (NJSGC) Academic Year Internship (2025–2026). CEA acknowledges the support of the NSF CAREER Award [CMMI-2338508]. TZ acknowledges the support of the Carter N. Paden, Jr. Distinguished Chair for Innovation in Materials Science and Metals Processing at Georgia Tech.

Data availability
The datasets generated and analyzed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.17842616 (Liu and Gabriel 2026).

Declarations

Competing interests
The authors declare no competing interests.

Received: 10 December 2025 / Accepted: 9 March 2026
Published online: 02 April 2026

References

M. Abadi et al., TensorFlow: Large-scale machine learning on heterogeneous systems (2015).
ASTM International, ASTM E1221–23: Standard test method for determining plane-strain crack-arrest fracture toughness, KIa, of ferritic steels (ASTM International, West Conshohocken, PA, 2023).
C.E. Athanasiou, X. Liu, B. Zhang, T. Cai, C. Ramirez, N.P. Padture, J. Lou, B.W. Sheldon, H. Gao, Integrated simulation, machine learning, and experimental approach to characterizing fracture instability in indentation pillar-splitting of materials. J. Mech. Phys. Solids 170, 105092 (2022). https://doi.org/10.1016/j.jmps.2022.105092
D.B. Barker, R. Chona, W.L. Fourney, G.R. Irwin, A report on the round robin program conducted to evaluate the proposed ASTM test method for determining the crack arrest fracture toughness, KIa, of ferritic materials. NUREG/CR-4996 (ORNL/Sub/79–7778/4), U.S. Nuclear Regulatory Commission, Washington, DC (1988).

Z.P. Bažant, J.L. Le, *Probabilistic Mechanics of Quasibrittle Structures: Strength, Lifetime, and Size Effect* (Cambridge University Press, Cambridge, 2017)
L. Breiman, Random Forests. Mach. Learn. 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
M.C. Cacas, E. Ledoux, G. de Marsily, B. Tillie, A. Barbreau, E. Durand, B. Feuga, P. Peaudecerf, Modeling fracture flow with a stochastic discrete fracture network: calibration and validation: 1. The flow model. Water Resour. Res. 26, 479–489 (1990). https://doi.org/10.1029/WR026i003p00479
P.B. Crosley, E.J. Ripling, Development of a standard test for measuring Kla with a modified compact specimen. NUREG/CR-2294 (ORNL/Sub-81/7755/1), Materials Research Laboratory, Glenwood, IL (1981).
P.B. Crosley, W.L. Fourney, G.T. Hahn, R.G. Hoagland, G.R. Irwin, E.J. Ripling, Final report on cooperative test program on crack arrest toughness measurements. NUREG/CR-3261, U.S. Nuclear Regulatory Commission, Washington, DC (1983).
R.A. Fisher, *Statistical Methods for Research Workers* (Oliver & Boyd, Edinburgh, 1925)
P.C. Gehlen, R.G. Hoagland, C.H. Popelar, A method of extracting dynamic fracture toughness from CT tests. Int. J. Fract. 15, 69–84 (1979). https://doi.org/10.1007/BF00115910
A.A. Griffith, The phenomena of rupture and flow in solids. Philos. Trans. A Math. Phys. Eng. Sci. 221, 163–198 (1921). https://doi.org/10.1098/rsta.1921.0006
G.X. Gu, C.T. Chen, M.J. Buehler, De novo composite design based on machine learning algorithm. Extreme Mechanics Letters 18, 19–28 (2018a). https://doi.org/10.1016/j.eml.2017.10.001
G.X. Gu, C.T. Chen, D.J. Richmond, M.J. Buehler, Bioinspired hierarchical composite design using machine learning: simulation, additive manufacturing, and experiment. Mater. Horiz. 5, 939–945 (2018b). https://doi.org/10.1039/c8mh00653a
J.P. Gudas, Micromechanisms of fracture and crack arrest in two high strength steels (Report No. DTNSRDC/SME-87–20, David Taylor Naval Ship Research and Development Center, Bethesda, MD, 1987).
K. Guo, M.J. Buehler, A semi-supervised approach to architected materials design using graph neural networks. Extreme Mech. Lett. 41, 101029 (2020). https://doi.org/10.1016/j.eml.2020.101029
P.Z. Hanakata, E.D. Cubuk, D.K. Campbell, H.S. Park, Accelerated search and design of stretchable graphene kirigami using machine learning. Phys. Rev. Lett. 121, 069901 (2018). https://doi.org/10.1103/PhysRevLett.121.255304
T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd edn. (Springer, New York, 2009). https://doi.org/10.1007/978-0-387-84858-7
Y.C. Hsu, C.H. Yu, M.J. Buehler, Using deep learning to predict fracture patterns in crystalline solids. Matter 3, 197–211 (2020). https://doi.org/10.1016/j.matt.2020.04.019
J.W. Hutchinson, Singular behaviour at the end of a tensile crack in a hardening material. J. Mech. Phys. Solids 16, 13–31 (1968). https://doi.org/10.1016/0022-5096(68)90014-8
J.D. Hyman, S. Karra, N. Makedonska, C.W. Gable, S.L. Painter, H.S. Viswanathan, Dfnworks: a discrete fracture network framework for modeling subsurface flow and transport. Comput. Geosci. 84, 10–19 (2015). https://doi.org/10.1016/j.cageo.2015.08.001
International Organization for Standardization, *ISO 20064: Metallic materials-Steel-Method of test for the determination of brittle crack arrest toughness, Kca* (International Organization for Standardization, Geneva, 2019)
R.M. Jones, A. Sharma, R. Hotchkiss, J.W. Sperling, J. Hamburger, C. Ledig, R. O'Toole, M. Gardner, S. Venkatesh, M.M. Roberts, R. Sauvestre, M. Shatkhin, A. Gupta, S. Chopra, M. Kumaravel, A. Daluiski, W. Plogger, J. Nascone, H.G. Potter, R.V. Lindsey, Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit. Med. 3, 144 (2020). https://doi.org/10.1038/s41746-020-00352-w
A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information. Phys. Rev. E 69, 066138 (2004). https://doi.org/10.1103/PhysRevE.69.066138
A.J. Lew, C.H. Yu, Y.C. Hsu, M.J. Buehler, Deep learning model to predict fracture mechanisms of graphene. NPJ 2D Mater. Appl. 5, 48 (2021). https://doi.org/10.1038/s41699-021-00228-x
X. Liu, C. Gabriel, Crack Arrest Fracture Toughness of Steels. (2026). https://doi.org/10.5281/zenodo.17842615.Accessed5Feb
X. Liu, C.E. Athanasiou, N.P. Padture, B.W. Sheldon, H. Gao, A machine learning approach to fracture mechanics problems. Acta Mater. 190, 105–112 (2020). https://doi.org/10.1016/j.actamat.2020.03.016
X. Liu, C.E. Athanasiou, N.P. Padture, B.W. Sheldon, H. Gao, Knowledge extraction and transfer in data-driven fracture mechanics. Proc. Natl. Acad. Sci. U. S. a. (2021). https://doi.org/10.1073/pnas.2104765118
X. Liu, Crack Arrest Fracture Toughness Predictor (2025), https://hint1412.github.io/Crack-Arrest/Kca_Predictor_Steels/index.html. Accessed 7 Dec 2025.
A. Müller, N. Karathanasopoulos, C.C. Roth, D. Mohr, Machine learning classifiers for surface crack detection in fracture experiments. Int. J. Mech. Sci. 209, 106698 (2021). https://doi.org/10.1016/j.ijmecsci.2021.106698
S. Niu, V. Srivastava, Simulation trained CNN for accurate embedded crack length, location, and orientation prediction from ultrasound measurements. Int. J. Solids Struct. 242, 111521 (2022). https://doi.org/10.1016/j.ijsolstr.2022.111521
W.J. Nordell, W.J. Hall, Two-stage fracturing in welded mild steel plates. Weld. J. Res. Suppl. 44, 124–134 (1965)
F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011)
J.R. Rice, A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 35, 379–386 (1968). https://doi.org/10.1115/1.3601206
J.R. Rice, G.F. Rosengren, Plane strain deformation near a crack tip in a power-law hardening material. J. Mech. Phys. Solids 16, 1–12 (1968). https://doi.org/10.1016/0022-5096(68)90013-6
A. Rovinelli, M.D. Sangid, H. Proudhon, W. Ludwig, Using machine learning and a data-driven approach to identify the small fatigue crack driving force in polycrystalline materials. NPJ Comput. Mater. 4, 35 (2018). https://doi.org/10.1038/s41524-018-0094-7
C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
S.C.Y. Shen, M. Peña Fernández, G. Tozzi, M.J. Buehler, Deep learning approach to assess damage mechanics of bone tissue. J. Mech. Behav. Biomed. Mater. 123, 104761 (2021). https://doi.org/10.1016/j.jmbbm.2021.104761

K. Shukla, P.C. Di Leoni, J. Blackshire, D. Sparkman, G.E. Karniadakis, Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks. J. Nondestruct. Eval. 39, 61 (2020). https://doi.org/10.1007/s10921-020-00705-1

G. Srinivasan, J.D. Hyman, D.A. Osthus, B.A. Moore, D. O'Malley, S. Karra, E. Rougier, A.A. Hagberg, A. Hunter, H.S. Viswanathan, Quantifying topological uncertainty in fractured systems using graph theory and machine learning. Sci. Rep. 8, 116657 (2018). https://doi.org/10.1038/s41598-018-30117-1

J. Taylor, A. Mehmanparast, R. Kulka, P. Moore, L. Xu, G.H. Farrahi, Experimental study of the relationship between fracture initiation toughness and brittle crack arrest toughness predicted from small-scale testing. Theor. Appl. Fract. Mech. 110, 102799 (2020). https://doi.org/10.1016/j.tafmec.2020.102799

The Japan Welding Engineering Society, WES 2815: Method of determining brittle crack arrest toughness Kca (JWES, Tokyo, 2014).

M. Valera, Z. Guo, P. Kelly, S. Matz, V.A. Cantu, A.G. Percus, J.D. Hyman, G. Srinivasan, H.S. Viswanathan, Machine learning for graph-based representations of three-dimensional discrete fracture networks. Comput. Geosci. 22, 695–710 (2018). https://doi.org/10.1007/s10596-018-9720-1

W. Weibull, A statistical distribution function of wide applicability. J. Appl. Mech. 18, 293–297 (1951). https://doi.org/10.1115/1.4010337

F. Yanagimoto, T. He, K. Shibanuma, The state-of-art of studies on brittle crack arrest in steel. Eng. Fract. Mech. 323, 111132 (2025). https://doi.org/10.1016/j.engfracmech.2025.111132

E. Zhang, M. Dao, G.E. Karniadakis, S. Suresh, Analyses of internal structures and defects in materials using physics-informed neural networks. Sci. Adv. 8, eabk0644 (2022). https://doi.org/10.1126/sciadv.abk0644

Publisher's Note