IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Identifying the influence of surface texture waveforms on colors of polished surfaces using an explainable AI approach

Yuhao Zhong, Akash Tiwari, Hitomi Yamaguchi, Akhlesh Lakhtakia & Satish T.S. Bukkapatnam

To cite this article: Yuhao Zhong, Akash Tiwari, Hitomi Yamaguchi, Akhlesh Lakhtakia & Satish T.S. Bukkapatnam (2022): Identifying the influence of surface texture waveforms on colors of polished surfaces using an explainable AI approach, IISE Transactions, DOI: 10.1080/24725854.2022.2100050

To link to this article: https://doi.org/10.1080/24725854.2022.2100050

IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2022.2100050

Identifying the influence of surface texture waveforms on colors of polished surfaces using an explainable AI approach

Yuhao Zhong[a], Akash Tiwari[a], Hitomi Yamaguchi[b], Akhlesh Lakhtakia[c], and Satish T.S. Bukkapatnam[a]

[a]Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA; [b]Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA; [c]Engineering Science and Mechanics, The Pennsylvania State University, University Park, PA, USA

ABSTRACT

An explainable artificial intelligence approach based on consolidating the Local Interpretable and Model-agnostic Explanation (LIME) model outputs was devised to discern the influence of the surface morphology on the colors exhibited by stainless-steel 304 parts polished with a Magnetic Abrasive Finishing (MAF) process. The MAF polishing process was used to create two regions, each appearing either blue or red to the naked eye. The color distribution was microscopically heterogeneous, i.e., some red microscale patches were dispersed in blue regions, and vice versa. The surface morphology was represented in the frequency domain (using a 2D Fourier transform) to capture the harmonic surface patterns, such as the feed and lay marks from the polishing process. A Convolutional Neural Network (CNN) was employed to identify the color of the region from the frequency characteristics of the surface morphology. The CNN was able to predict the observed colors with test accuracies exceeding 99%, suggesting that the frequency characteristics of the surface morphology of the red regions are distinctly different from those of the blue regions. A LIME model was constructed around each small segment within each region of the surface to identify the frequency features that are influential for differentiating between the colors. To deal with the effect of heterogeneity, an algorithm based on the query by experts was used to reconcile the local influences and gather the global explanations of the frequency characteristics that inform the blue versus red regions. We found that the dominant morphological features in the red regions are those that capture the polishing lay patterns underlying surface structure, whereas those in the blue regions capture the non-uniform and high-frequency waveform patterns, such as those result when oxide films form due to the intense polishing conditions.

ARTICLE HISTORY
Received 30 August 2021
Accepted 29 June 2022

KEYWORDS
Convolutional neural network; explainable machine learning; local interpretable and model-agnostic explanation; magnetic abrasive finishing process; surface morphology and colors

1. Introduction

Structural colors and pigments on polished metallic surfaces have fascinated humankind for millennia, starting from the colors on wootz steel, Damascus swords, and ornaments (Srinivasan and Ranganathan 2004). In recent times, researchers have shown renewed interest in metal surface coloration, as it has a wide range of applications from simple aesthetic-decorations up to advanced information-encoding for authentication and traceability of products in supply chains (Veiko *et al.*, 2017; He *et al.*, 2019; Wang *et al.*, 2020). Due to the increasing focus on cybersecurity assurance in manufacturing supply chains, embedding watermarks in products has garnered increased interest (Mahesh *et al.*, 2020; Tiwari, Villasenor, Gupta, Reddy, Karri and Bukkapatnam, 2021). Surface colors are attractive as potential watermarks on products, as they have minimal effect on the product functionality. However, current technologies for imparting durable surface colors by controlling the surface morphology remain slow and expensive (Liu *et al.*, 2019). Hence, they are limited to embedding colors over small areas of a product. A cost-effective technology to controllably impart surface colors over large areas does not exist.

Recently, it has been reported that Magnetic Abrasive Finishing (MAF) (Yamaguchi *et al.*, 2007; Ganguly *et al.*, 2013) can produce colors on stainless-steel surfaces during the polishing process (Tiwari, Xu, Lakhtakia, Yamaguchi and Bukkapatnam, 2021). The MAF process utilizes a magnetic field to force ferromagnetic particles and abrasives against the target surface with a relative motion to finish a variety of metallic surfaces (Yamaguchi *et al.*, 2007; Ganguly *et al.* 2013). The MAF process offers the potential to impart a variety of colors over a large area, including curved and freeform surfaces.

Figure 1 shows two sample surfaces from the MAF process. Under natural daylight, three colored regions on the curved surface are visible to the naked eye. For the purpose of discussion, they are named: (i) yellow, unpolished, (ii) red, mildly polished, and (iii) blue, intensely MAF-polished regions. However, it was noted that these regions are not monochromatic, but in fact are highly heterogeneous (see Figure 2). Each region is composed of different, multi-colored patterns at the microscale, and these microscale patterns are dispersed across the blue and the red regions,

CONTACT Satish T.S. Bukkapatnam satish@tamu.edu
Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24725854.2022.2100050.

albeit at different intensities (Tiwari, Xu, Lakhtakia, Yamaguchi and Bukkapatnam, 2021).

In order to create specific surface colors using the MAF polishing process, it is necessary to first gain insight into the causes of color formation and the characteristics of the colors formed. Several studies have been conducted to investigate the effect of morphology imparted by various manufacturing and surface modification processes on surface colors. For example, Veiko *et al.* (2017) showed that the period, height, and orientation angle of laser-induced periodic structures on AISI 304 stainless steel inform surface coloration. Zheng *et al.* (2002) found that the laser-induced colors also vary with the thickness of a layer of oxides formed on the surface. The thickness of the Anodic Aluminum Oxide (AAO) films have been found to affect the color of metal-AAO-Al nanostructures (Manzano *et al.*, 2018). Tiwari, Xu, Lakhtakia, Yamaguchi and Bukkapatnam (2021) have suggested that the oxide films form preferentially along the MAF abrasion feed marks on the surface and that the observed colors result from the oxide pigments and thin-film interference. This earlier study, however, did not delineate the effect of structural morphology on the colors.

In the present work, we focus on understanding the surface morphological characteristics (i.e., the surface height distribution) that inform the major colors in the red and blue regions. It is noted that the surface height distribution may be attributed to the thickness and distribution of the oxide films, as well as the intrinsic surface structure (e.g., abrasion lay patterns). We frame the present work as an attempt to offer statistically consistent inferences for the following two questions:

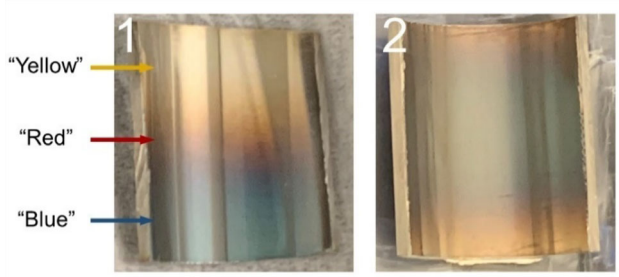

Figure 1. Stainless-steel 304 curved parts cut off from MAF-polished tubes (left: Sample 1, right: Sample 2) showing three distinctly colored regions: yellow, red and blue (images taken with a smartphone camera).

Q1: To what extent can the surface morphology inform the major surface colors?

Q2: Which morphological features influence these colors and why?

Machine learning methods can address both questions by producing models that connect output variables to input variables. For example, Baxter *et al.* (2019) employed neural networks to predict colors from laser parameters, as well as from geometric parameters such as nanoparticle spacing and radius, in a laser-machining process. In their work, the colors were represented by RGB (red, green, and blue) values converted from simulated reflectance spectra. Machine learning methods have also been applied to decode the information encrypted under patterned structural colors by classifying the optical microscope images (He *et al.*, 2019). However, the mechanism of how the complicated machine learning models, also called "black box" models, generate "good" predictions usually remains unexplained. This can raise questions as the models could have merely captured noise or artifacts in the data. Meanwhile, we might lose opportunities to discover some previously unknown patterns that the model may have learned.

Efforts during at least the past four decades on interpreting the outcomes of machine learning models (Scott *et al.*, 1977), coupled with their growing complexity, have led to the emergence of *explainable* AI (XAI) as an active research area. XAI techniques aim to produce details that make the functioning of "black box" models clearer or easier to understand for certain audiences (Gunning, 2017; Arrieta *et al.*, 2020).

Generally, XAI techniques can be classified into *ante-hoc* and *post-hoc* approaches (Murdoch *et al.*, 2019). Ante-hoc approaches aim to construct transparent models, such as linear regression, decision trees, and modified neural networks that are inherently easier for humans to understand and retrace the decision processes (Li *et al.*, 2018). For instance, Letham *et al.* (2015) built interpretable Bayesian decision lists based on carefully selected features extracted from health records to predict stroke risks. Post-hoc approaches use simpler surrogates of pre-trained complex models to generate intuitive explanations and reasoning that the complex "black box" models could not provide in the first place (Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Zhang *et al.*, 2018). In practice, the ante-hoc approaches are preferred whenever the underlying relationships are simple, and/or

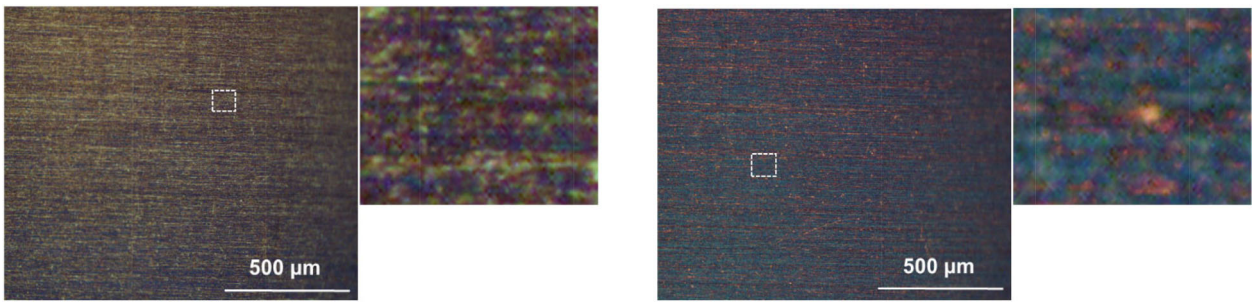

Figure 2. Micrographs of (left) red and (right) blue regions, and their zoomed local areas showing a random, anisotropic dispersion of various colors within each region (obtained using Olympus BX51 microscope under 10x magnification).

adequate domain expertise exists to determine the useful features beforehand to build accurate models. The post-hoc approaches are typically used for uncovering hidden knowledge underlying "black box" models (Murdoch *et al.,* 2019; Rudin, 2019). In our case, post-hoc approaches are more suitable because our goal is to discern and explain the complex relationship between surface morphology and the colors.

Local Interpretable Model-agnostic Explanation (LIME) is a state-of-the-art post-hoc XAI method to provide local explanations for many "black box" classification and regression models (Ribeiro *et al.,* 2016). It uses surrogate models that are inherently interpretable to derive local approximations of a "black box" model at various neighborhoods of the input predictor space. Consequently, the complicated functional relationships captured by an accurate "black box" model is dissected into a set of local explanations, such that each explanation holds for a particular neighborhood of the inputs space. One of the first applications of LIME in the manufacturing domain was to draw physical insights on material microstructures buried in the acoustic emission gathered from a nano-scratching process (Iquebal *et al.,* 2020).

Although LIME provides a local explanation for the prediction made by the "black box" model at each input, deriving a global explanation from consolidating the local explanations remains a standing issue. Such a consolidation becomes essential, albeit challenging, in the present case. This is because the surface heterogeneities, resulting from the dispersion of red microscale patterns in the blue region and vice versa, can yield conflicting local interpretations. Prior research on generating LIME-based global explanations (Ribeiro *et al.,* 2016; Chettri *et al.,* 2018; ElShawi *et al.,* 2019; Ibrahim *et al.,* 2019; Sangroya *et al.,* 2020) mostly focused on maximizing the diversity of explanations. None of the methods have considered the goodness of the local explanations, e.g., the coefficient of determination (R^2) of the linear models that are used in LIME to locally approximate the "black box" model. However, since the local explanations are directly inferred from the coefficients of the linear models, they can eventually lead to false global explanations if these models fail to capture the local behavior of the "black box" model. This usually happens when heterogeneity is present in the data, making the local decision boundary highly nonlinear (Ribeiro *et al.,* 2016).

The main methodological contribution of this article lies in the approach to reconcile the conflicting local interpretations from LIME to derive consistent global explanations, and thereby identify how the important morphological (frequency) features inform the red or blue regions. This method is based on a query-by-experts construct, and it uses recent theoretical results to guide the selection of some hyperparameters. The results from our experimental study and analysis establish a consistent set of frequency bands and their variations that inform the red and the blue regions. The method can be applied to analyze surfaces produced in processes other than MAF and be extended to explore the relationships between other morphological features and process attributes in addition to colors.

The remainder of this article is organized as follows: We present an overview of the experimental studies in Section 2. Sections 3 and 4 elucidate the technical approach. Section 5 presents the implementation results and discussions. Conclusions are given in Section 6.

2. Surface texture measurements

The samples used in this study are SS304 curved workpieces cut from a 1-mm-thick tube (with 18 mm inside diameter) and polished using the MAF process at the University of Florida (see Figure 1). As mentioned earlier, we define three types of regions on the surface based on the major colors visible to the naked eye: yellow, red, and blue. For this study, we focused on the red and the blue regions. The samples were also examined using an Olympus BX51 microscope with 10× magnification. As noted earlier, the micrographs shown in Figure 2 reveal that the color distribution is highly heterogeneous at the microscale. The defined red and blue regions are not monochromatic but contain multiple colors at the microscale, some of which even appear in both regions. This complicates the decision boundary of distinguishing the two regions.

To quantitatively analyze the surface characteristics, surface heightmaps from 18 different locations of the red and blue regions of the two samples were measured using two optical profilers with different magnifications. The purpose of using different samples and instruments is to evaluate the generalizability of our findings by studying the heightmaps from different sources separately and comparing the consistency of the results. Overall, the specifications of all the measurements are listed in Table 1. Six 1024×1024 maps each from Sample 1 and Sample 2 were obtained via Zygo Zegage at 10× magnification (at Texas A&M University). Another six 1000 × 1000 maps from Sample 1 were obtained via Zygo Zegage Pro HR at 50× magnification (at Zygo Corporation). All the heightmaps were labeled as "red" or "blue" based on the region from which they occur. The heightmaps were measured along the centerline of each curved surface to minimize errors that might have occurred in the curve-form removal process in the Zygo software.

Figure 3 exhibits representative surface geometries obtained from the red and blue regions under 50× magnification and after curve-form removal. The heightmap appears to show fewer non-uniformly distributed local spikes in the red region than in the blue region; this increases the surface height of local areas in the blue region where the oxide-film formation is prominent (Tiwari, Xu, Lakhtakia, Yamaguchi, and Bukkapatnam, 2021). Also, since the optical profilers are based on interferometry, the measurements may be affected by the semi-transparent/transparent oxide films on the surface, especially in the blue regions. We employed suitable data preprocessing and verification procedures to mitigate this issue, as outlined in the next section.

Table 1. Specifications of the heightmap measurements.

Sample	*Number of heightmaps*	*Heightmap data size*	*Heightmap actual size* (μm)	*Spatial resolution* (*nm*)	*Magnification*	*Instrument*
1	6	1024×1024	835×835	815	10×	Zegage
2	6	1024×1024	835×835	815	10×	Zegage
1	6	1000×1000	174×174	174	50×	Zegage Pro HR

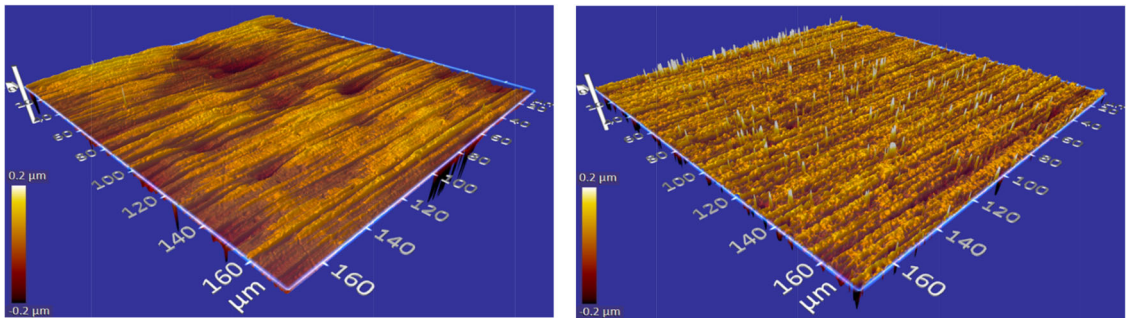

Figure 3. Examples of 1000×1000 ($174\mu m \times 174\mu m$) surface heightmaps (after curve-form removal) from (left) a red region and (right) a blue region, under 50× magnification via Zygo Zegage Pro HR, where more high-frequency content and nonuniform spikes are observed in the blue heightmap.

Figure 4. Summary of the research approach.

3. Technical approach overview

Figure 4 summarizes our approach to employ experimental data for addressing the two core questions stated in Section 1. To address Q1, the noisy heightmaps gathered from the "red" and "blue" regions are preprocessed and transformed into amplitude spectra via two-dimensional Fast Fourier Transform (2D FFT). Then, a Convolutional Neural Network (CNN) model is trained to learn to classify these labeled spectra. The classification accuracy is an indicator of the extent to which the surface morphology informs the major surface colors, and thus answers Q1. For Q2, we implemented a LIME model (Section 4) to explain predictions made by the "black box" CNN model. In this way, we can identify which frequency bands and the corresponding surface morphological patterns (extracted via inverse FFT) are influential and how they inform the surface color.

As noted from the experimental observations, each colored region consists of a heterogeneous dispersion of multiple colors at the microscale. We divided each measured heightmap into smaller tiles within which the spectral content (i.e., the underlying covariance structure) remains consistent. This can not only increase the data volume, but also reduce the color variance in each smaller tile. A nonlinear classifier such as a CNN can establish complex class boundaries to capture the effects of heterogeneity. Furthermore, the heightmap measurements are prone to significant uncertainties and often betray unusual large spikes (see Figure 3(b)). To mitigate this measurement limitation, we discarded the tiles that contain the outliers (i.e., the height values outside the interquartile range). Subsequently, a surface roughness tester (Model SJ-210 from Mitutoyo) was used to verify the consistency between *Ra* values extracted from the tiles and the surface-measurements.

Among the morphological characteristics of the surface, we focused on how the harmonic waveform patterns influence the colors. Accordingly, we treated every tile as a 2D spatial signal and employed its frequency-domain representation to capture the waveform patterns (these patterns are difficult to discern in the spatial domain). We employed a 2D FFT routine to decompose a heightmap into signal components that contain the amplitude and phase information for different frequency pairs in each dimension. One can recover a specific waveform pattern or the whole heightmap in the spatial domain by applying an Inverse FFT (IFFT) to the FFT coefficients of a specific set of frequency bands. After extracting the FFT coefficient matrix of the heightmap, we took the amplitude of every element to generate an amplitude spectrum. Additionally, because of the property of conjugate symmetry of FFT (Gonzalez and Woods, 2018), we kept only the upper two quadrants of each spectrum (including the DC (direct current) component).

Note that 2D FFT amplitudes are shift-invariant, i.e., the frequency features can be compared across the spectra extracted from the heightmaps. Comparatively, morphological features are hard to compare across heightmaps. This uniform feature format enabled us to study the global importance using LIME, which will be elaborated in the following discussions. Besides, the frequency features can be converted back into morphological features in the spatial domain.

Furthermore, the heterogeneous microscale color distribution in the red and blue regions leads to significant divergence among the amplitude spectrum values with in every class (region). The CNN uses the diverse 2D spectra as inputs to create nonlinear classification (O'Shea and Nash, 2015). The classification accuracy of the CNN model developed thus suggested the extent to which the surface

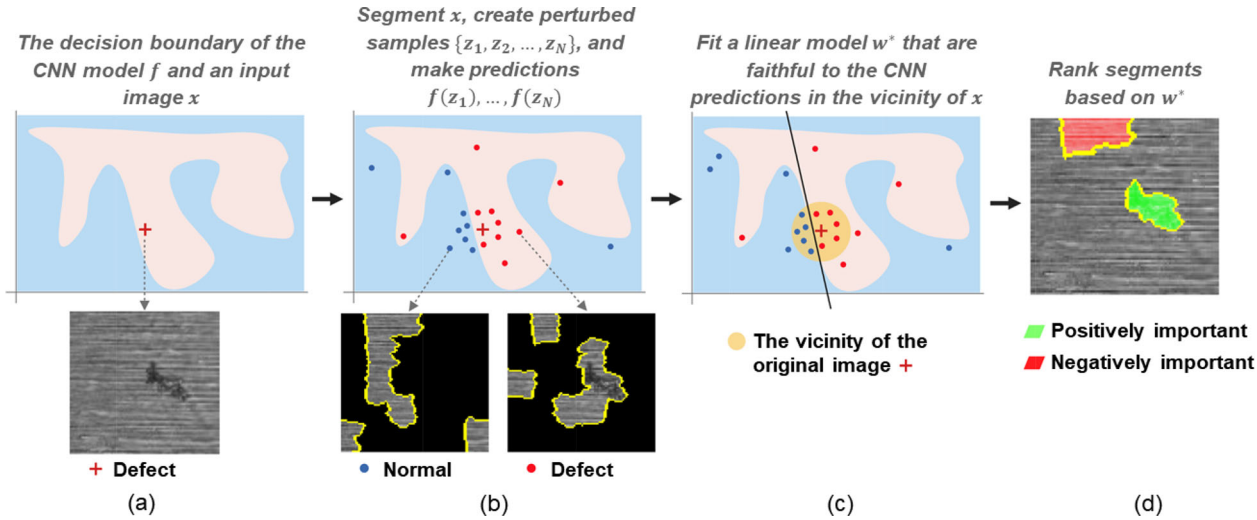

Figure 5. An illustration of LIME. LIME uses a linear interpretable model to approximate the CNN model at a local area in the feature space. The linear model coefficients indicate the importance of every image segment—a positive and larger coefficient indicates the presence of the corresponding segment would make the CNN model more likely to classify the image into the class of interest.

morphology informs the major surface colors, which answered Q1. The specific aspects of the surface morphology that inform the color (Q2) were determined by leveraging the LIME procedure, as detailed in the next section.

4. LIME and global explanations

4.1. LIME modeling to explain how the spectra of each local image region inform the colors

Towards understanding our application of LIME, let us consider a simple case of surface defect detection using micrograph images as illustrated in Figure 5. Here, the defect detection reduces to a binary classification of images into two classes, namely "normal" and "defect". In this context, an "explanation" refers to the delineation of the image segments that are important for the CNN prediction of each input image. Let f denote a CNN binary classifier model that partitions the space of the inputs constituting various "normal" and "defect" images into two regions (see Figure 5(a)). To get a local explanation about an image x in the input space, LIME first segments x into d parts. Then $x' \in \{1\}^d$ is a mask, i.e., a binary vector where each element corresponds to a segment and all the d elements are one, is used to represent this image. Then, LIME creates N different perturbation masks $z_i^{'} \in \{0,1\}^d$, $(i = 1, ..., N)$ by randomly setting some of the elements in x' to zero according to a Bernoulli distribution, i.e., Bernoulli(0.5). Each of these masks (see Figure 5(b)) corresponds to a perturbed image z_i $(i = 1, ..., N)$ in the input space.

Next, the CNN model f is used to make predictions on all these images (Figure 5(b)). Here, for the binary classification, if a softmax function was used in the CNN output layer, then each prediction is a probability vector containing probabilities towards each class; if a sigmoid function was used in the CNN output layer, each prediction is a single probability towards the class that is encoded as one (Goodfellow *et al.*, 2016).

Afterwards, for each class (if softmax was used) or the class that was encoded as one (if sigmoid was used), LIME trains a linear model $g : \mathbb{R}^d \to \mathbb{R}$, defined by its coefficient vector $w^* \in \mathbb{R}^d$ (ignoring the intercept). This linear model is trained to locally approximate the prediction outputs from f regarding the specified class, where the inputs to the linear model are the binary masks (see Figure 5(c)). Here, "locally" implies that the linear model only captures how the CNN model behaves in the vicinity of the original image x (as a simple linear model cannot mimic the nonlinear CNN model in the whole feature space). To define such locality, each perturbed image z_i is assigned with a weight $\pi_x(z_i)$ based on how similar z_i is to x. Based on this construct, a local linear model about x can be formulated as follows:

$$w^* = \underset{w}{\operatorname{argmin}} \ \mathcal{L}(f, w, \pi_x) + \Omega(w), \tag{1}$$

where $\mathcal{L}(f, w, \pi_x)$ is a loss function that measures the fidelity of the linear model to the CNN model f in the vicinity of x, and $\Omega(w)$ penalizes the complexity of w. The sign and magnitude of each coefficient in w^* indicate positive/negative importance of the presence of the corresponding segment towards the class (see Figure 5(d)).

In our case, an input 2D spectrum is segmented into d equal-sized rectangles, each representing the energy of a frequency range along the two axes. Consequently, LIME can outline which frequency bands are more dominant in determining its color classification result (see Figure 6). During the training of w^*, ridge regression is employed to regularize the model complexity. So, to sum up, LIME explains a spectrum regarding the prediction towards the qth class by solving

$$w^* = \underset{w}{\operatorname{argmin}} \ \sum_{i=0}^{N} \pi_x(z_i) \left[f_q(z_i) - \left(\left\langle w, z_i^{'} \right\rangle + \rho \right) \right]^2 + \lambda \langle w, w \rangle, \tag{2}$$

where $z_0 = x$, $z_0^{'} = x'$, $f_q(z_i)$ is the output of the CNN model regarding the qth class (with the use of softmax). ρ is

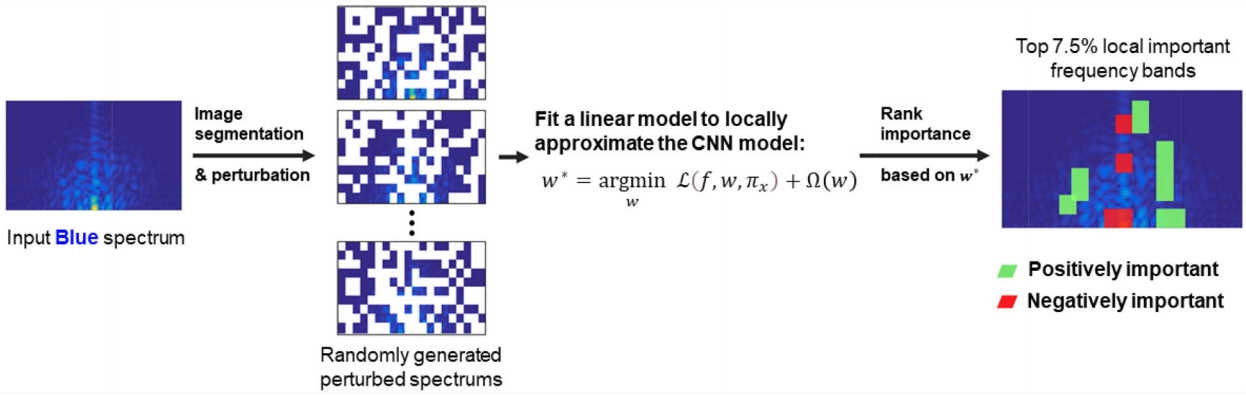

Figure 6. An illustration of the LIME model for a (2D FFT) spectrum input x. The spectrum is segmented into d rectangles (frequency bands), based on which a set of perturbed instances are generated. The local importance is ranked based on the fitted linear model w^*. The result highlights the top $\mathcal{T}$ fraction of local important bands. Here, the presence of the positively important bands would make the CNN model more likely to classify this spectrum (input x) as "blue".

the intercept and λ is the ridge regularization term. The function $\pi_x(z_i)$ is calculated using a radial basis function kernel as

$$\pi_x(z_i) = \exp\left(-\frac{D(x, z_i)^2}{\delta^2}\right), \tag{3}$$

where δ is the kernel width. The similarity $D(x, z_i)$ between x and z_i is calculated as the cosine of the angle between their vector masks x' and z_i'.

Among the hyperparameters of LIME, λ affects the sparsity and the fidelity of the linear model relative to the CNN model. A large λ can lead to underfitting, and hence, it should be chosen carefully employing guidance from van Wieringen (2015). Although a small δ would make the linear model highly sensitive to the local inaccuracies of the CNN model, the linear models constructed with a large δ may not capture the local trend accurately. The number of segments (frequency bands) d, and hence the spectral band widths, should be adjusted to ensure sufficient frequency resolution while avoiding the effects of spectral leakage (Harris, 1978). An improper choice of d can cause multicollinearity effects where different combinations of (mostly adjacent) segments spuriously emerge together as important. Additionally, the number of perturbations N should be chosen carefully towards achieving consistency of the linear model w^*. Generally, one can have a maximum of $2^d - 1$ possible local perturbations for an image x. However, considering typical pixelized image sizes, d is usually greater than 100. Therefore, much smaller N is commonly used to reduce computation costs. This, however, induces uncertainty and inconsistency in the resulting local explanations. The following result (Slack *et al.*, 2021) provides guidance to select N.

Theorem 1. (*Perturbed instance size*) (Slack *et al.*, 2021): *The number of perturbed instances N required to achieve an uncertainty interval width ξ of feature importance at a user-specified confidence level α can be estimated as*

$$N = 4\varepsilon^2 \Big/ \left\{ \overline{\pi} \times \left[\frac{\xi}{\Phi^{-1}(\alpha)} \right]^2 \right\}, \tag{4}$$

where $\overline{\pi}$ is the average weight estimated from an initial set of J perturbed instances as $\overline{\pi}_J = \sum_{i=0}^{J} \pi_x(z_i)/J$, ε^2 is the empirical sum of squared errors between a LIME model and f, weighted by $\pi_x(z_i)$ for $i = 1, \ldots, J$, and $\Phi^{-1}(\alpha)$ is the two-tailed inverse normal cumulative distribution function at confidence level α. □

It can be observed that N needs to be sufficiently large to keep the uncertainty interval ξ small. Also, the required N increases linearly with the weighted residual sum of squares ε^2 to maintain a specific ξ. Equation (2) aims to force the local response of the model g to approximate the CNN output $f_q(z_i)$, i.e., the outputs lie in the (0,1) interval. However, it does not guarantee the response to lie within, or even close to the (0,1) interval, i.e., some local models can have a large error. Therefore, it is necessary to evaluate the accuracy of a local model before using its local explanations.

4.2. Consolidating the local explanations to derive a global explanation

As noted, LIME trains a linear model w^* for each spectrum (input x) and employs the values of w^* to explain how the features (i.e., the elements of x) inform the color (i.e., the class label). However, due to surface heterogeneity, local areas may contain multiple color patterns. Consequently, the locally important frequency bands can differ drastically among the spectra taken even within the same region (i.e., the same label). In addition, as noted earlier, the local explanations themselves from a linear model can vary significantly depending on the random perturbations used to build the model. As noted earlier, the perturbations are a much smaller random sample of all possible perturbations (i.e., $N \ll 2^d - 1$) (Zhang *et al.*, 2019; Slack *et al.*, 2021).

Several efforts have been made on generating LIME-based global explanations. For example, SP-LIME (Ribeiro *et al.*, 2016; Sangroya *et al.*, 2020) aggregates the coefficients in the local explanations based on selected inputs, i.e., inputs that lead to diverse local explanations. In addition, SP-LIME also uses the absolute values of the coefficients to determine the importance of the features, and consequently, the importance values and at times the relative importance of a feature (i.e., a frequency band) can vary depending on the realization of the random perturbation sample. This can affect the consistency of the explanations. Apart from this, ILIME

(ElShawi *et al.,* 2019) and GAM (Ibrahim *et al.,* 2019) methods utilize clustering algorithms to group similar local explanations as well as their corresponding inputs, and hence, provide global explanations for different clusters of the inputs. These earlier methods focused on including the highest-possible diversity within the local explanations to derive the global explanations. However, they neglected the goodness of the local explanations, i.e., whether the linear model is faithful enough to the "black box" model locally. In our case where heterogeneity is present in the data, the decision boundary in certain local input neighborhoods can still be highly nonlinear for a linear model to capture, or outright different from those of other local models in the region, all these resulting in models with poor accuracies. Using the local explanations from such models may eventually lead to false global explanations.

Moreover, there is a challenge of getting LIME-based global explanations for image-like data such as heightmaps, mainly because the features from different heightmaps are not comparable. Fortunately, in our case, with the use of FFT spectra and the uniform segmentation across all spectra, we can easily create a measure of global importance by aggregating the local explanations from a group of representative spectra.

Algorithm 1: LIME-based global explanations for the class "red"

Input	All the n "red" spectra, CNN model f
Output	Global positive importance of all frequency bands for the class "red"
Initialization	d : number of frequency bands $I_k = 0 =$ global positive importance of the kth frequency band ($k = 1, 2, ..., d$)
Start	**Iterate for each spectrum x_r, $r \in \{1, 2, ..., n\}$:**
1	**Apply LIME towards the class "red":** optimize $w_r^* = \underset{w}{\text{argmin}}\ \mathcal{L}(f, w, \pi_{x_r}) + \Omega(w)$
2	**Aggregate local importance:** **If (R^2 of the model w_r^*) $> \theta$:** $\mathbb{P}_{\mathcal{T}}^r =$ {positively important bands among the top $\mathcal{T}$ of x_r} **Else:** $\mathbb{P}_{\mathcal{T}}^r = \emptyset$
3	**Iterate $k = 1$ to d:** $I_k = \sum_r^n 1_{\{k \in \mathbb{P}_{\mathcal{T}}^r\}}$

Algorithm 1 exhibits the details of deriving LIME-based global explanations for the class "red" (similarly for the class "blue"). The algorithm consists of three major steps: (i) *Identify the "expert" input spectra*, (ii) *Identify the local positively important frequency bands*, and (iii) *Poll the global positively important frequency bands*. The first two steps are iterated for each input "red" spectrum x_r ($r \in \{1, 2, ..., n\}$) from all the n "red" spectra to gather their local explanations. After the iterations are over, the third step calculates the global positive importance of each frequency band towards the class "red".

Identify the "expert" input spectra:

We want to only aggregate the correct local explanations as there are cases where the linear LIME models fail to locally approximate the CNN model and thus provide incorrect explanations. To do so, we first apply LIME to train a linear model w_r^* based on the given "red" spectrum x_r and its perturbed spectra. The model is trained to approximate the CNN's predictions towards the class "red", since we are only interested in the frequency bands that are important to the class "red". Then, the model identifies if the given spectrum x_r is an "expert" – one whose linear explanation model w_r^* has fairly high performance, e.g., its coefficient of determination $R^2 > \theta$. Here, R^2 measures how well the linear model captures the behavior of the CNN model locally around x_r.

Identify the local positively important frequency bands:

To acquire the frequency signature for each class, we investigate the positively important frequency bands for "red" and "blue" separately. Therefore, once the given "red" spectrum x_r is qualified as "expert", we get its local explanation towards the class "red" (towards "blue" if the given input is a "blue" spectrum), i.e., rank the d frequency bands correspondingly by the absolute magnitude of the d coefficients in w_r^* (excluding the intercept coefficient). Within the top $\mathcal{T}$ fraction of bands, only the positive bands (corresponding to positive coefficients) are recorded as a set $\mathbb{P}_{\mathcal{T}}^r$ ($\mathbb{P}_{\mathcal{T}}^r = \emptyset$ if x_r is not an "expert"). Note that the size of $\mathbb{P}_{\mathcal{T}}^r$ can vary given different spectra x_r ($r \in \{1, 2, ..., n\}$), as in some cases there are fewer positive bands in the top $\mathcal{T}$ bands. This works better than recording a fixed number of positive bands because the latter may include the positive bands that are not really important, i.e., their corresponding coefficients are in the rear rank regarding the absolute magnitude.

Poll the global positively important frequency bands:

After iterating the previous two steps for all the n "red" spectra and getting n sets $\mathbb{P}_{\mathcal{T}}^r$ ($r \in \{1, 2, ..., n\}$), for each frequency band $k \in \{1, 2, ..., d\}$ (uniformly segmented for all spectra), its global positive importance towards the class "red" I_k can be calculated by

$$I_k = \sum_r^n \mathbf{1}_{\{k \in \mathbb{P}_{\mathcal{T}}^r\}}. \tag{5}$$

The threshold θ for the R^2 in Algorithm 1 can be decided based on the distribution of the R^2 statistic. Specifically, the R^2 statistic of a linear regression model follows a $Beta(\mathcal{K}/2, (\mathcal{S} - \mathcal{K} - 1)/2))$ distribution under the null hypothesis that all coefficients are zero (Helland, 1987). Here, $\mathcal{K}$ is the degrees of freedom of the model g and $\mathcal{S}$ is the rank of the inputs (generally, $\mathcal{K} = d$ and $\mathcal{S} = N$). In other words, θ can be chosen as the critical value (to reject the null hypothesis) at a certain significance level. Conventional choices also exist for the selection of θ (Henseler *et al.,* 2009; Moore *et al.,* 2021).

The hyperparameter $\mathcal{T}$ is chosen based on the distribution of the local importance values. The local importance values $|w^*|$ follow a Pareto-type distribution, and the elbow

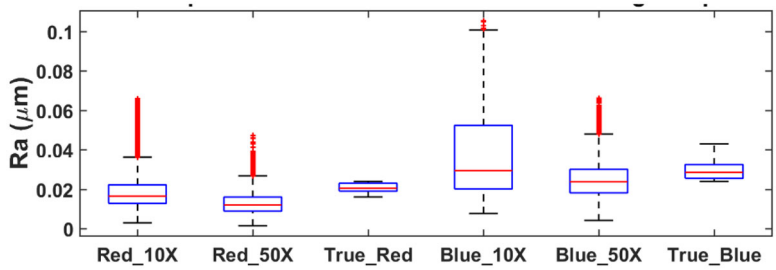

Figure 7. Boxplot of the *Ra* values of profile slices from 30 × 30 outlier-free heightmaps (under 10× and 50×) and the contact-type *Ra* values measured by the surface roughness tester (considered as ground-truth) shows a large degree of consistency.

point of the distribution guides the choice of $\mathcal{T}$ (Satopaa *et al.*, 2011). A proper choice of $\mathcal{T}$ would guarantee sufficient, yet not very redundant top local frequency bands in order to capture the consistent global bands. Additional discussion and procedural details can be found in Section F of the supplementary material. While the global importance values I_k also follow a similar highly right-skewed distribution, the important global bands stand out clearly and allow a clear identification.

In addition, for a certain, reasonably broad range of values of N, λ and δ (discussed in Section 5.5), the proposed algorithm was observed to provide consistent global explanations to identify the top global frequency bands. However, under some unusual settings, such as when the regularization term λ is excessively large, there may not be enough "expert" spectra for the polling, e.g., most of the linear models are underfitting and have $R^2 < \theta$. Also, as noted earlier, a judicious choice of d is desirable to reduce the effect of the spectral leakage induced by 2D FFT.

5. Implementation and results

5.1. Data preprocessing and 2D FFT

As noted in Section 3, to ensure that each instance (e.g., training sample for machine learning) is reasonably homogeneous and enough number of instances are available for later study, the original heightmaps were divided into 30 × 30 heightmap tiles. The maps containing outliers (as detected in the original heightmaps) were discarded. The boxplot in Figure 7 suggests that the *Ra* values calculated from the 30x30 heightmaps after outlier removal are consistent with the *Ra* measured by the Mitutoyo surface roughness tester (regarded as the ground-truth).

After data preprocessing, the 30 × 30 heightmaps were transformed into 2D FFT spectra. Note that the dimension of the output spectrum is the same as the input heightmap. To get a finer and consistent frequency bin resolution, each heightmap was zero-padded to 100×100 (for the maps at 10× magnification) or 500 × 500 (for the maps at 50× magnification) before 2D FFT. After 2D FFT, we only kept the upper two quadrants of each spectrum (including the DC component) considering the conjugate symmetry property. Overall, 2000 30x30 heightmaps (600 for Sample 1 under 10× and 50×, 800 for Sample 2 at 10×) were transformed into 1400 51 × 100 and 600 251 × 500 amplitude spectra. An example is given in Figure 8. Here, a negative frequency indicates an opposite spatial direction to the same frequency with a positive sign.

5.2. CNN results

While understanding to what extent the data can be discriminated, we also tried to develop the most predictive models. Two CNN models were built, respectively, based on the 600 51 × 100 spectra and 600 251 × 500 spectra from Sample 1. The details of the model architecture are summarized in Figure B1 in the supplementary material. The CNN model based on 600 51 × 100 spectra was constructed with two hidden layers, including a convolutional layer and a max pooling layer, to extract the features. Then, a fully connected layer was used to flatten the feature maps (of dimension 4 × 6×8) into a vector and classify it into the two classes via a softmax function. This model was trained for 90 epochs.

The CNN model based on 600 251 × 500 spectra was constructed with three hidden layers and a fully connected layer, and this model was trained for 60 epochs. Rectified Linear Unit (ReLU) (Glorot *et al.* 2011) was employed as the nonlinear activation function in every convolutional layer.

Moreover, the data set was split into 90% for training and 10% for testing. Both CNN models were trained under the same learning rate of 0.0002 with an Adam optimizer (Kingma and Ba, 2014). Note that the architecture and learning rate are determined through hyperparameter-tuning trials using a validation dataset within the training data (Yamashita *et al.*, 2018).

According to Figure 9, the CNN first model has training and testing accuracies of 99.6% and 98.3%, respectively. The second CNN model also has high accuracies (99.1% for training and 98.3% for testing). Hence, the results suggest that both CNN models have learned the pattern in the spectra for discriminating the surface colors. This answer to Q1 therefore is that surface morphology can inform the major surface colors to a large extent.

5.3. Global important frequency bands

As introduced earlier, LIME can identify which segments in a spectrum, i.e., frequency bands, as each segment is rectangular, contribute significantly to the color-classification result of the spectrum. Nevertheless, the heterogeneous color

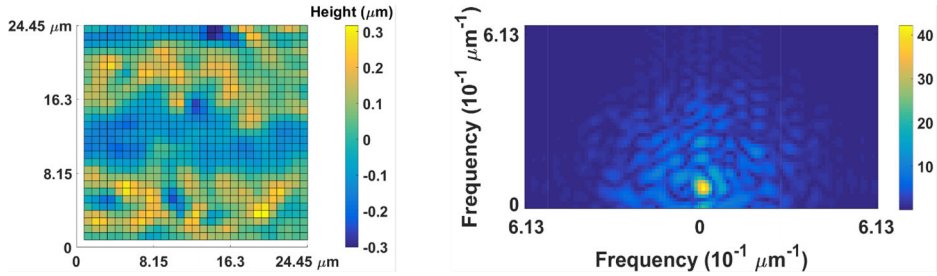

Figure. 8. A 30×30 heightmap from a blue region (left) and its 51×100 FFT amplitude spectrum (right).

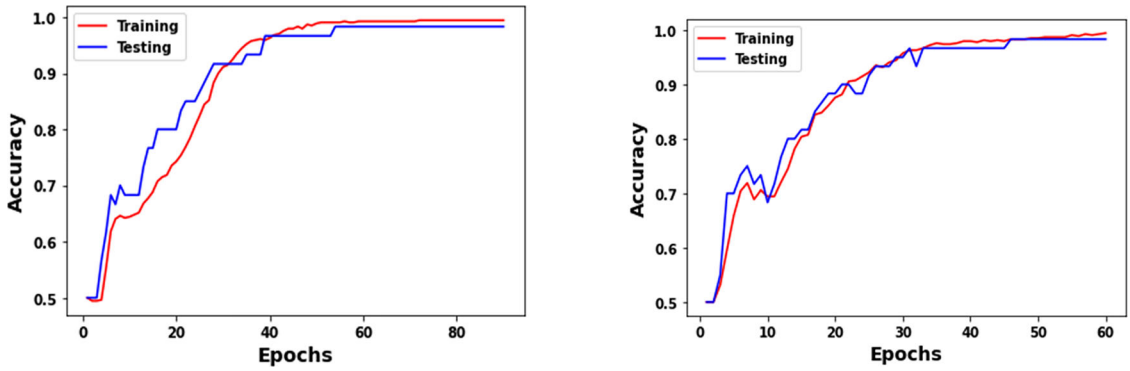

Figure 9. Training and testing accuracies of the CNN models based on (left) 51×100 and (right) 251×500 spectra. The model with the best testing accuracy is considered as the final model.

distribution complicates the decision boundary in the feature space, and thus, the neighborhoods defined by very diverse spectra can lead to inconsistent local explanations. Therefore, to find the general patterns corresponding to the two major colors, consistent global explanations obtained by aggregating the local explanations would be a better option.

During the implementation, each 51×100 spectrum was segmented into 200 frequency bands of size 5×5 (bands at the bottom of the spectrum are of size 6×5), based on which 300 perturbed spectra were generated for each spectrum, whereas each 251×500 spectrum was segmented into 5000 bands of size 5×5 (bands at the bottom of the spectrum are of size 6×5) and the number of perturbed spectra was set to 300. In both cases, the kernel width δ was set as 0.25 and the ridge penalty term λ was set as one based on the guidance in Section 4.2. Moreover, Algorithm 1 was used to identify the global positively important frequency bands for "red" and "blue", respectively. The algorithm considers only the spectra that were locally well explained (their interpretable linear models have high R^2). This avoids the consolidation of the local explanations that are highly divergent with the CNN classifier. As mentioned in Section 4.2, the R^2 threshold θ was set as 0.75, and $\mathcal{T}$ was set to be 7.5%. Here, for d and N set as 200 and 300, respectively and at a significance level of 0.01, $Beta(d/2, (N-d-1)/2) \approx 0.75$. Here, to be more conservative, d and N were set as 200 and 300, respectively, based on the case of 51×100 spectra.

Figure 10 visualizes the results of the global positive importance of all the frequency bands towards the two colors for 51×100 spectra and 251×500 spectra, in the form of heat maps. Each rectangular segment represents a frequency range along the vertical and horizontal directions.

The results for 51×100 spectra (corresponding to heightmaps at $10\times$ magnification) indicate that the global positive important bands (used with "global bands" interchangeably in the rest of this article) for "red" tend to be along the centerline of a spectrum, discontinuously covering the spatial frequency range from -0.72×10^{-4} to 0.48×10^{-4} nm^{-1} on the horizontal direction and from 0 to 4.93×10^{-4} nm^{-1} vertically. In contrast, global positive important bands for "blue" are mostly off-centerline, where the top ones are mainly at two different horizontal ranges: -2.52×10^{-4} to -1.92×10^{-4} nm^{-1} and 1.68×10^{-4} to 2.28×10^{-4} nm^{-1}, and vertically range from 0 to 2.52×10^{-4} nm^{-1}. This pattern can also be observed in the results of the 251×500 spectra (corresponding to $50\times$ heightmaps), even though the global bands may not cover the same range. Specifically, for 251×500 spectra, the top "red" global bands approximately span from -1.83×10^{-4} to 1.6×10^{-4} nm^{-1} horizontally and 0 to 3.54×10^{-4} nm^{-1} vertically, whereas the top "blue" global bands can approximately reach as high as -6.97×10^{-4} and 5.03×10^{-4} nm^{-1} horizontally and 12.69×10^{-4} nm^{-1} vertically.

Additionally, the results from the 251×500 spectra further suggest that the presence of a lower-frequency pattern in a "red" spectrum increases the probability of this spectrum being classified as "red", whereas the presence of a higher-frequency pattern in a "blue" spectrum increases the probability that the spectrum is classified as "blue". Under the two different magnifications, there is a considerable overlap between the results of "red" global bands, especially those near the DC component. Nevertheless, the "blue" global bands of 251×500 spectra cover much higher frequency than those of 51×100 spectra do, but it is hard to conduct further comparison due to different Nyquist frequencies under these two magnifications.

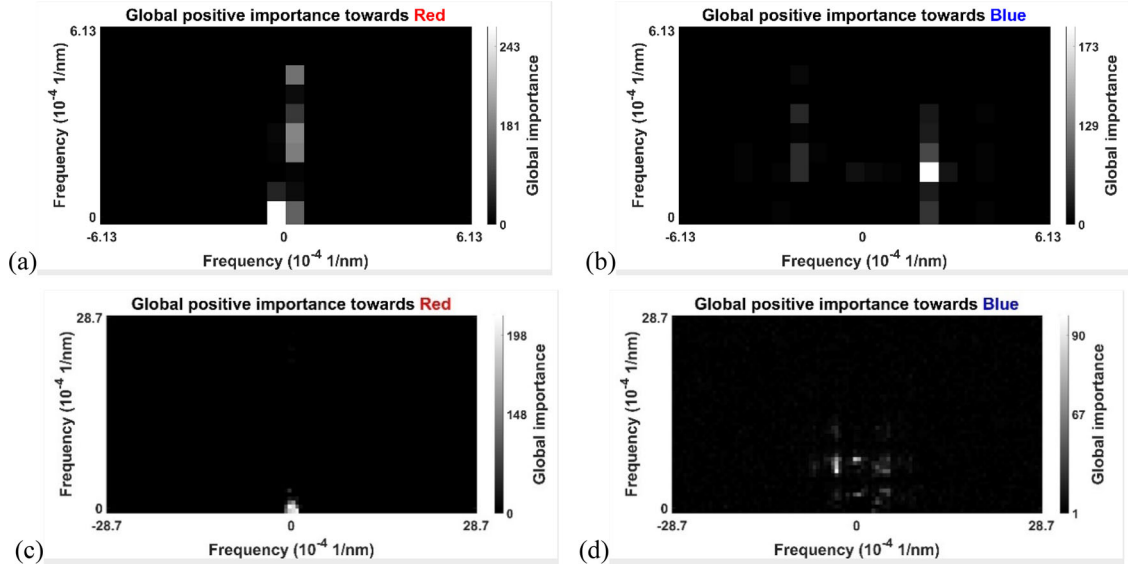

Figure 10. Global positive importance of all the frequency bands towards the class (a) "red" and (b) "blue" from the 51 × 100 spectra. Global positive importance of all the frequency bands towards the class (c) "red" and (d) "blue" from the 251 × 500 spectra.

Next, we evaluate the generated explanations in terms of their descriptive accuracy, relevancy and consistency (Horel, 2020). For descriptive accuracy, we verify the discriminative power of the identified global frequency bands (see Section 5.4). Next, we study the consistency of Algorithm 1 regarding two aspects: different randomization settings and different hyperparameter values (see Section 5.5). For relevancy, we assess the extent to which the surface waveform patterns recovered from applying IFFT to the global bands can make connections to prior physical understandings (see Section 5.6).

5.4. Descriptive accuracy verification

5.4.1. Validity of the global bands

Two separate attempts were made to verify whether the frequency bands with high global importance are truly dominant in discriminating "red" and "blue". In the first attempt, the same data was classified using only the top global bands, in the sense that the presence of only the global bands should have enough predictive power to classify most of the spectra. In the second attempt, the actual values, i.e., FFT magnitudes, of the top global bands were compared for the "red" and "blue" spectra. The latter is a further exploration beyond the LIME-based explanation since LIME only delineates the effect of the presence of a band instead of the specific values of the band that were actually used by the CNN model.

In the first attempt, we identified the top 2.5% global bands for each class based on the results in Section 5.3. Next, we masked all the other bands, (i.e., set the values of them to zero) in all the previously correctly classified spectra and input the masked spectra to the same CNN models. As a result, the classification accuracy was 88.6% and 88% for 51 × 100 spectra and 251 × 500 spectra, respectively. This implies that the global importance calculated according to Algorithm 1 can indeed identify the truly discriminative bands.

In the second attempt, since each frequency band is of size 5 × 5 or 6 × 5 containing multiple FFT magnitudes, we took the average of all the values. Figures 11(a) and (b) show the average FFT magnitudes of the top 1 "red" global band (-0.72×10^{-4} to -0.12×10^{-4} nm^{-1} horizontally, 0 to 0.72×10^{-4} nm^{-1} vertically) and the top 1 "blue" global band (1.68×10^{-4} to 2.28×10^{-4} nm^{-1} horizontally, 1.32×10^{-4} to 1.92×10^{-4} nm^{-1} vertically), respectively, for all the "red" and the "blue" 51 × 100 spectra that were qualified as "experts". Similarly, Figures 11(c) and (d) show the average FFT magnitudes of the top 1 "red" global band (0.5×10^{-4} to 1×10^{-4} nm^{-1} horizontally, 0.1×10^{-4} to 0.7×10^{-4} nm^{-1} vertically) and the top 1 "blue" global band (-3.5×10^{-4} to -3×10^{-4} nm^{-1} horizontally, 5.8×10^{-4} to 6.4×10^{-4} nm^{-1} vertically), respectively, for all the "red" and the "blue" 251 × 500 "expert" spectra. Clearly, the average FFT magnitudes for the "blue" spectra in all plots cover almost the same value range as that of the "red" spectra do, but with a slightly higher median. Furthermore, based on the *p*-values from a paired-sample *t*-test (shown in Figure. 11 below each boxplot), the difference between the means of the average FFT magnitudes for the "red" and the "blue" spectra is statistically significant at the 0.05 level in all four cases. Overall, it implies that the selected global bands are indeed globally discriminative regarding the two classes. However, it does not imply that the bands without statistical difference between the means are not globally important. This is because the classification is not simply a "black and white" problem, i.e., it involves a highly nonlinear decision boundary and possible interaction effects of multiple frequency bands, especially when heterogeneity is present.

5.4.2. Generalization on a new sample

In addition to evaluating the generalizability of the results at different magnifications, we also evaluated the generalizability

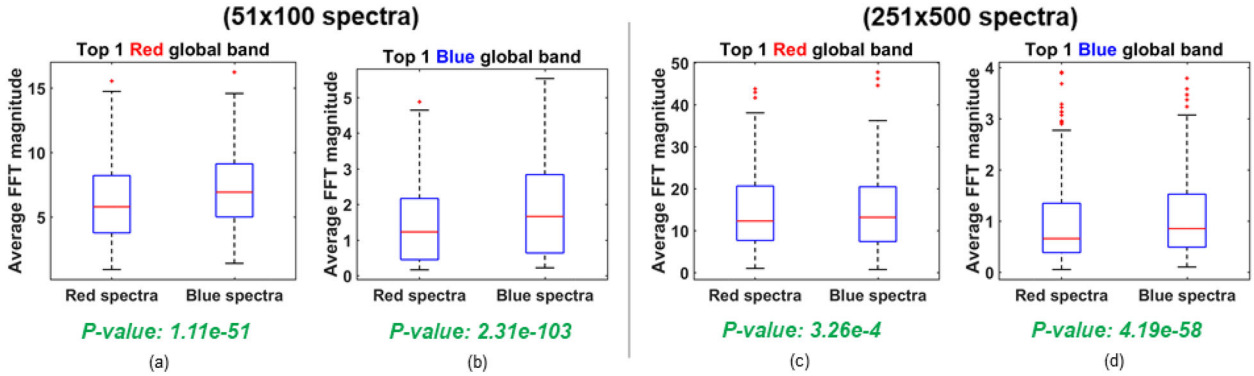

Figure 11. Comparison between "red" and "blue" spectra regarding the average FFT magnitudes of the four top 1 global frequency bands. (a) and (b) correspond to the result of 51×100 spectra. (c) and (d) correspond to the result of 251×500 spectra. In all cases, the value ranges of the two classes are similar, but "blue" spectra always have a slightly higher median in these bands. The *p*-values from a paired-sample *t*-test also suggest that the means of the two groups are statistically different. These reinforce the result that the bands are globally discriminative regarding the two classes.

of our findings on another sample, with the use of the 800 51×100 spectra (400 "red" ones and 400 "blue" ones) from Sample 2. Sample 2 was cut off from the same polished tube as Sample 1. The idea was to test if the trained CNN model and the global frequency bands identified using Sample 1 could be applied to classify a different surface (Sample 2) which was polished through the same process and measured by the same instrument. Therefore, in these 800 spectra, all the frequency bands were masked except the top 2.5% globally important ones for "red" and "blue". As a result, the classification accuracy of the previously trained CNN model on Sample 2 was found to be 83.88%, which implies that our findings can be generalized to different samples from the same process.

5.5. Consistency verification

For consistency, we conduct the implementation stated in Section 5.3, albeit with different randomization settings and different hyperparameter values. This is to check the extent to which randomization and hyperparameter values affect the global bands and their importance. In particular, the values for λ include 0, 0.1, 1 (default), 10 and 50. The values of δ include 0.25 (default) and 10, and the values of N include 200, 300 (default), 600 and 1000. It turned out that the results were highly consistent, and in fact, the global bands did not vary, under different randomizations (for examples, see Section C in the supplementary material). The top global bands identified using Algorithm 1 were also fairly robust to changes in the hyperparameter values over the ranges (despite the inevitable variations in importance values).

We compared the performance of Algorithm 1 relative to an extant approach, SP-LIME (see Section E in the supplementary material). SP-LIME aggregates the local explanations of inputs regardless of the accuracy of their LIME (local) models to derive global explanations. Besides, it uses the square root of the absolute value of w^* as opposed to the use of the ranks in Algorithm 1 to derive the global importance of the bands. As a net result of using the values of the local model coefficients w^* themselves, and not being selective with local models, 14 times more bands can emerge as globally important with SP-LIME compared to those with Algorithm 1. The same factors can also cause SP-LIME to identify different global bands and assign different importance values under different realizations of the perturbations. In contrast, Algorithm 1 uses the ranking of coefficients without the exact values and considers the positive and negative values separately. Therefore, as noted earlier, the identified global bands remain fairly consistent and their importance has a smaller variance, relative to the actual realizations of the perturbations.

5.6. Relevance verification

To completely answer Q2, the global frequency bands were transformed back into the spatial domain via inverse 2D FFT while masking other uninterested frequency bands. Specifically, first, one "red" and one "blue" spectrum at each level of magnification were selected, where the LIME models for these spectra had the highest R^2. Then, based on each of these four spectra, the results of inverse 2D FFT of the top 2.5% "red" and "blue" global frequency bands, and the original heightmap were plotted, respectively, in both 2D (the upper row) and 3D views (the lower row) (see Figure 12). For simplicity, we will call the waveform patterns corresponding to the top global frequency bands "red" or "blue" waveform patterns. Here, to distinctly show the waveform patterns, the color scales within each case are set differently.

Based on Figure 12, we made the following four observations:

1. Within each of the four instances, the "blue" waveform patterns tend to have smaller height values than the waveform patterns corresponding to "red" top global bands.
2. At both 10× and 50× magnification, the "red" waveform patterns have a clearer directionality than the "blue" ones, but both align with the lay pattern generated by the abrasive action during the MAF process.
3. The globally important morphological features are consistent across all the magnifications and instruments considered in this study. The "red" features capture the general form of the mountain ridges, i.e., hills and valleys that result from polishing. The "blue" features

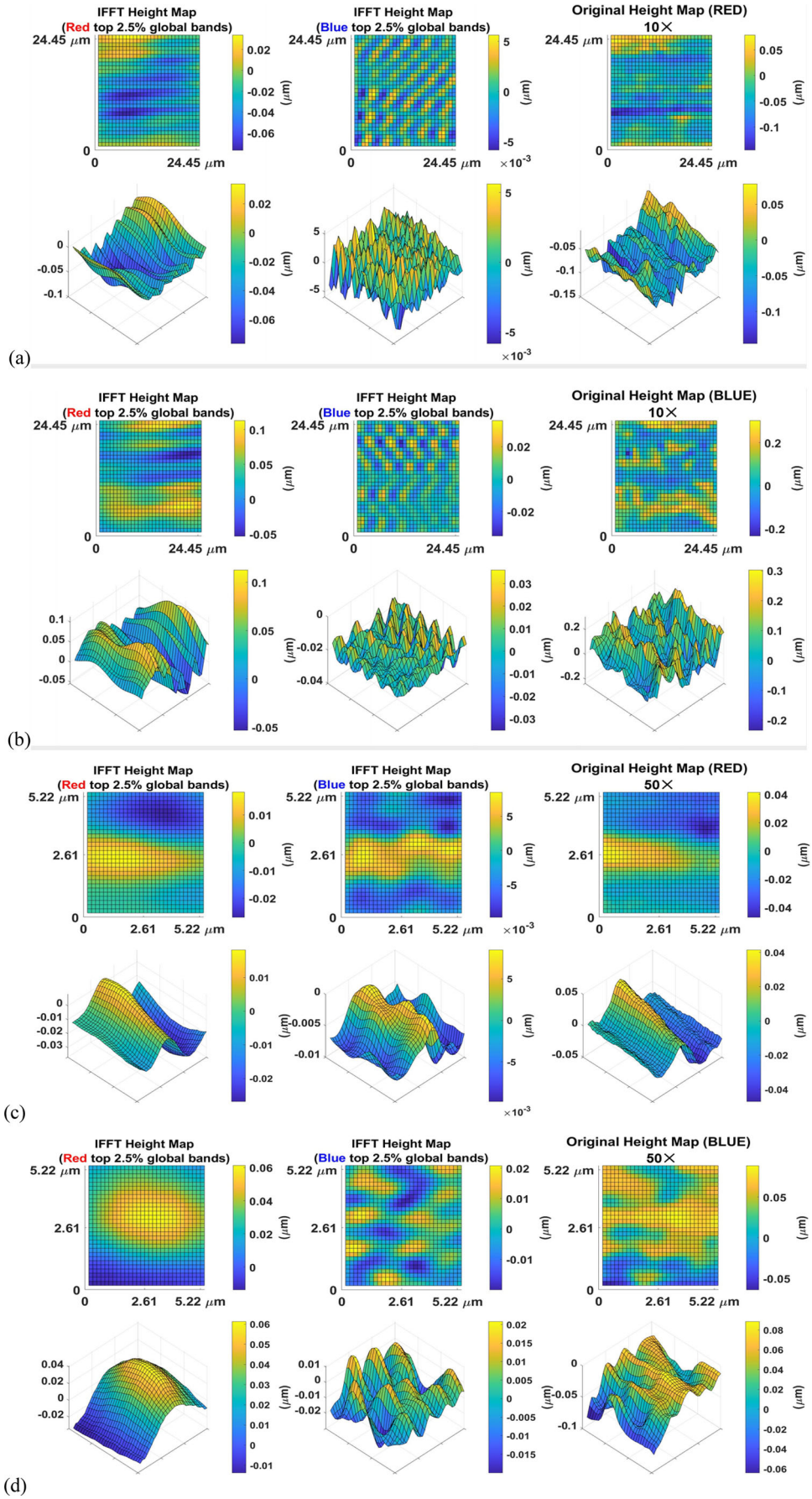

Figure 12. Waveform patterns corresponding to the top 2.5% "red" global bands (left) and the top 2.5% "blue" global bands (middle) (obtained via inverse 2D FFT) and the original heightmap (right), based on four representative spectra (with high-R^2 LIME models): (a) a "red" 10× spectrum, (b) a "blue" 10× spectrum, (c) a "red" 50× spectrum, (d) a "blue" 50× spectrum. In each case, the upper row shows the waveform patterns in a 2D view, while the lower row provides the corresponding 3D view. Plots of the "red" and the "blue" bands employ different color scales to facilitate an adequate visualization of the waveform patterns.

capture the peaky high-frequency hills that seemingly formed above the "red" ones.

4. Both the "red" and the "blue" waveform patterns can be observed in each of the four original heightmaps, regardless of the region from which the heightmap was sampled. This indicates that the waveform patterns also have a heterogeneous distribution. However, the "red" waveform patterns seem to be more dominant in the heightmaps sampled from the "red" region, and vice versa. This can be used as a guidance to visually classify the heightmaps.

These observations also offer some insight into the physical sources responsible for the way these morphological features influence the colors. Generally, the colors on metal surfaces can be caused due to interference, pigments or diffraction (Liu *et al.*, 2019). Pertinently, the frequency of the red light (700 nm) was among the top 2.5% most important global frequency bands (along the vertical direction) of the surface morphology, as identified in this study. This suggests that at least a part the observed red color is informed by the waveforms on the surface. However, the wavelength of the most important global band (the brightest band in Figure 10(c)) lies above 2.825 μm, and it is within the spatial wavelengths of the horizontal lay patterns which are between 1.5 and 15μm (measured using ZEISS EVO MA-10 Scanning Electron Microscope). This indicates that the "red" colors are strongly influenced by the lay patterns.

An earlier study (Tiwari, Xu, Lakhtakia, Yamaguchi and Bukkapatnam, 2021) suggested that different colors arising over the polished surface is due to the combined effects of the phenomenon of interference of transparent oxide films and other pigmented chemical species. The resulting colors are affected by the polishing intensity during the MAF process, i.e., the down force exerted by the magnetic abrasive material onto the surface during polishing. Specifically, they argue that the occurrence of flash temperatures – proportional to the polishing intensity – induces chemical reactions leading to growth of transparent oxide films and heterogeneously distributed pigmented chemical species. The flash temperatures occurring during polishing is a local phenomenon. This causes a heterogenous distribution of the oxide film thickness and pigmented oxides over the polished surface. Thinner oxide film growth results in the red color when the surface undergoes milder polishing intensity. The thicker oxide films resulting in blue color are observed in regions with heavier polishing intensity. The pigmented colors arise due to iron oxides (visually red) and chromium oxides (visually blue). Additionally, they also reported that the colors appear along the circumferential lay pattern which is aligned along the direction of the abrasive action of the MAF process. These conclusions support our observations: in the red region the global morphological features appear to be mostly the underlying surface structure, e.g., the lay patterns, or the thin layer of oxide films formed along the lay patterns. In contrast, in the blue region, the global morphological features are representative of the heterogeneously distributed thicker oxide films over the surface, and the heterogenous distribution of chromium oxides. The heterogeneity captured by the morphological features agrees with the observation 4 and the micrographs in Figure 2.

6. Conclusion

In this study, we identified the influence of morphology on colors on polished stainless-steel surfaces using frequency domain analysis and XAI methods. First, we confirmed that the surface morphology can be used to accurately classify between the major surface colors. Next, by using a query-by-experts algorithm, we identified the globally important frequency bands and their corresponding morphological features, i.e., waveform patterns that are more influential on determining the major colors.

The study also indicates that the patterns of the identified global morphological features from the heightmaps measured under different magnifications are consistent. But, the globally important frequency bands at 50× magnification cover a broader and higher frequency range, possibly because higher magnification brings higher spatial resolution and thus more high-frequency information. Additionally, the findings (at the same magnifications) are consistent for different samples from the same process.

Subsequently, we explored the connections of the global explanations to the underlying physics by comparing our results with the findings in a prior study (Tiwari, Xu, Lakhtakia, Yamaguchi and Bukkapatnam, 2021). Both investigations suggest that the polishing intensity affects the surface colors. Over the intensely polished regions, the non-uniform and high-frequency global waveform patterns are consistent indicators of an oxide film formation and a consequent blue coloration. The global waveform patterns in the mildly polished red region appear to be merely related to the lay pattern. Understanding such a relationship between the surface morphology and colors through XAI can open the possibility to gainfully employ surface colors for product authentication and encryption, thereby saving expensive and time-consuming inspections for authentication and security.

Nonetheless, this study also has a few limitations. Since each of the waveform patterns corresponds to multiple frequency bands. The causal relationship between a specific frequency band and the colors needs a further study and confirmations.

From a methodological standpoint, this study has tackled various challenges such as the color heterogeneity shown in Figure 2, the measurement accuracy, as well as addressed the consistency and relevancy of global explanations in the context of the present application. This is also one of the few efforts aimed to provide global explanations for image inputs. Additional theoretical studies are needed to gather a deeper understanding of the effects of the various hyperparameters and randomizations in LIME (Zhang *et al.*, 2019).

In essence, our work can be deemed as an inchoate effort towards the application of MAF to impart controlled large-area structural colors on manufactured surfaces, thus offering a product authentication method that can be robust to

falsification. More importantly, this work can spur future applications of XAI to address various issues related to knowledge discovery, model diagnostics, and localization and causal inferences in the manufacturing domain. For instance, in knowledge discovery, the explanations generated by XAI can be regarded as hypotheses. The plausibility of such hypotheses being true can then be evaluated using the domain knowledge or subsequent confirmatory experiments. In model diagnostics, the goal is to identify the potential bias and errors in the model to ensure that the prediction mechanism aligns well with the expectation. In terms of localization and causal inferences, the surface defect detection illustrated in Figure 5 can serve as an example, where XAI is used for an automated identification of the image segments that contain a "defect" or such a product feature (Karthikeyan *et al.*, 2022).

Funding

U.S. National Science Foundation under grants 1849085 and 1953694; Texas A&M University Xgrants Program.

ORCID

Yuhao Zhong http://orcid.org/0000-0002-2560-3911
Satish T.S. Bukkapatnam http://orcid.org/0000-0003-3312-8222

Notes on contributors

Yuhao Zhong received his BEng degree in safety engineering from China University of Petroleum (Beijing) in 2018 and MS degree in industrial and systems engineering from Texas A&M University in 2020. He is currently a PhD student in industrial and systems engineering at Texas A&M University. His research interests include explainable machine learning and data analytics for manufacturing and healthcare. He is a student member of INFORMS and IEEE.

Akash Tiwari received a BTech degree in industrial and systems engineering from the Indian Institute of Technology, Kharagpur, India, in 2019. He is currently working towards a PhD degree with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA. His research interests include cybersecurity and data analytics for manufacturing. He was a Summer Intern with the Royal Enfield Motors Factory, Chennai, India, in 2017. In 2018, he was a Summer Research Intern with the Durham University Business School, Durham, U.K. He is a student member of the Institute for Operations Research and the Management Sciences (INFORMS).

Hitomi Yamaguchi is a professor in the Department of Mechanical and Aerospace Engineering at the University of Florida. Her research interests include magnetic field-assisted finishing, surface functionalization, and medical-device development. She has received many awards, including the 2021 Numata Memorial Paper Award from the Japan Society for Precision Engineering (JSPE). She is currently the Chair of the Scientific Technical Committee for Abrasive Processes (STC-G) of CIRP (International Academy for Production Engineering). She has been elected a fellow of both the American Society of Mechanical Engineers (ASME) and the Society of Manufacturing Engineers (SME).

Akhlesh Lakhtakia is Evan Pugh university professor and Charles Godfrey Binder professor of engineering science and mechanics at The Pennsylvania State University. He is currently interested in surface waves, solar cells, sculptured thin films, biologically inspired design, mimumes, and forensic science. Elected a fellow of eight learned societies, he received the 2010 SPIE Technical Achievement Award, the 2016 Walston Chubb Award for Innovation, the 2022 SPIE Smart Materials and Structures Lifetime Award, the 2022 IEEE Antennas and Propagation Distinguished Achievement Award, and a 2022-23 Jefferson Science Fellowship.

Satish T.S. Bukkapatnam is the Rockwell International Professor of Industrial & Systems Engineering at Texas A&M University, and the Director of Texas A&M Engineering Experiment Station Institute of Manufacturing Systems. He received his PhD degree in industrial and manufacturing engineering from Pennsylvania State University (1997). His research interests are broadly in smart manufacturing systems, and ultraprecision manufacturing. Dr. Bukkapatnam is a Fellow of IISE and SME, and was a Fulbright-Tocqueville Distinguished Chair.

References

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, **58**, 82–115.

Baxter, J., Lesina, A.C., Guay, J.-M., Weck, A., Berini, P. and Ramunno, L. (2019) Plasmonic colours predicted by deep learning. *Scientific Reports* 9:8074 (1), 1–9.

Chettri, B., Mishra, S., Sturm, B.L. and Benetos, E. (2018) Analysing the predictions of a CNN-based replay spoofing detection system.Presented at the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18-21 December.

ElShawi, R., Sherif, Y., Al-Mallah, M. and Sakr, S. (2019) ILIME: Local and global interpretable model-agnostic explainer of black-box decision, in *23rd European Conference on Advances in Databases and Information Systems*, Bled, Slovenia, Springer-Verlag Berlin, Heidelberg, pp. 53–68.

Ganguly, V., Schmitz, T., Graziano, A. and Yamaguchi, H. (2013) Force measurement and analysis for magnetic field-assisted finishing. *ASME Journal of Manufacturing Science and Engineering*, **135**(4), 041016.

Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep sparse rectifier neural networks, in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Volume 15, JMLR Workshop and Conference Proceedings, pp. 315–323.

Gonzalez, R.C. and Woods, R.E. (2018) *Digital Image Processing*, 4th edn. Pearson, New York, NY.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*, MIT Press, Cambridge, MA.

Gunning, D. (2017) Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)IJCAI-16DLAIWS.pdf (accessed May 2021).

Harris, F.J. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, **66**(1), 51–83.

He, X., Gu, Y., Yu, B., Liu, Z., Zhu, K., Wu, N., Zhao, X., Wei, Y., Zhou, J. and Song, Y. (2019) Multi-mode structural-color anti-counterfeiting labels based on physically unclonable amorphous photonic structures with convenient artificial intelligence authentication. *Journal of Materials Chemistry C*, **7**(45), 14069–14074.

Helland, I.S. (1987) On the interpretation and use of R2 in regression analysis. *Biometrics*, **43**(1), 61–69.

Henseler, J., Ringle, C.M. and Sinkovics, R.R. (2009) The use of partial least squares path modeling in international marketing, in *New Challenges to International Marketing*, Emerald Group Publishing Limited, Bingley, UK, pp. 277–319.

Horel, E. (2020) *Towards Explainable AI: Feature Significance and Importance for Machine Learning Models*, Stanford University, Stanford, CA.

Ibrahim, M., Louie, M., Modarres, C. and Paisley, J. (2019) Global explanations of neural networks: Mapping the landscape of

predictions, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, pp. 279–287.

Iquebal, A.S., Pandagare, S. and Bukkapatnam, S.T.S. (2020) Learning acoustic emission signatures from a nanoindentation-based lithography process: Towards rapid microstructure characterization. *Tribology International*, **143**, 106074.

Karthikeyan, A., Tiwari, A., Zhong, Y. and Bukkapatnam, S.T. (2022) Explainable AI-infused ultrasonic inspection for internal defect detection. *CIRP Annals* **71**(1), 449–452.

Kingma, D.P. and Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint *arXiv:1412.6980*.

Letham, B., Rudin, C., McCormick, T.H. and Madigan, D. (2015)Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, **9**(3), 1350–1371.

Li, O., Liu, H., Chen, C. and Rudin, C. (2018) Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI Press, pp. 3530–3537.

Liu, H., Lin, W. and Hong, M. (2019) Surface coloring by laser irradiation of solid substrates. *APL Photonics*, **4**(5), 051101.

Lundberg, S. and Lee, S.-I. (2017) A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, pp. 4768–4777.

Mahesh, P., Tiwari, A., Jin, C., Kumar, P.R., Reddy, A.N., Bukkapatanam, S.T., Gupta, N. and Karri, R. (2020) A survey of cybersecurity of digital manufacturing. *Proceedings of the IEEE*, **109** (4), 495–516.

Manzano, C.V., Ramos, D., Pethö, L., Bürki, G., Michler, J. and Philippe, L. (2018) Controlling the color and effective refractive index of metal-anodic aluminum oxide (AAO)–Al nanostructures: Morphology of AAO. *The Journal of Physical Chemistry C*, **122**(1), 957–963.

Moore, D.S., Notz, W. and Fligner, M.A. (2021) *The Basic Practice of Statistics*. 9th edn. W.H. Freeman, New York, NY.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019) Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, **116**(44), 22071–22080.

O'Shea, K. and Nash, R. (2015) An introduction to convolutional neural networks. arXiv preprint *arXiv:1511.08458*.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 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*, San Francisco, CA. Association for Computing Machinery New York, NY, pp. 1135–1144.

Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1**(5), 206–215.

Sangroya, A., Rastogi, M., Anantaram, C. and Vig, L. (2020) Guided-LIME: Structured sampling based hybrid approach towards explaining blackbox machine learning models.Presented at the CIKM 2020 Workshops, Galway, Ireland.

Satopaa, V., Albrecht, J., Irwin, D. and Raghavan, B. (2011) Finding a "kneedle" in a haystack: Detecting knee points in system behavior, in *Presented at the 31st International Conference on Distributed Computing Systems Workshops*, Minneapolis, MN.

Scott, A.C., Clancey, W.J., Davis, R. and Shortliffe, E.H. (1977) *Explanation capabilities of production-based consultation systems*. Department of Computer Science, Stanford University, Stanford, CA.

Slack, D., Hilgard, A., Singh, S. and Lakkaraju, H. (2021) Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, **34**, 9391–9404.

Srinivasan, S. and Ranganathan, S. (2004) *India's Legendary Wootz Steel: An Advanced Material of the Ancient World*, National Institute of Advanced Studies, Bengaluru, Karnataka.

Tiwari, A., Villasenor, E.J., Gupta, N., Reddy, N., Karri, R. and Bukkapatnam, S.T. (2021) Protection against counterfeiting attacks in 3D printing by streaming signature-embedded manufacturing process instructions, in *Proceedings of the 2021 Workshop on Additive Manufacturing (3D Printing) Security*, Association for Computing Machinery New York, NY, pp. 11–21.

Tiwari, A., Xu, F., Lakhtakia, A. Yamaguchi, H. and Bukkapatnam, S.T. (2021) On colors of stainless-steel surfaces polished with magnetic abrasives. *Applied Optics*, **60**(9), 2549–2559.

Van Wieringen, W.N. (2015) Lecture notes on ridge regression. arXiv preprint *arXiv:1509.09169*.

Veiko, V., Karlagina, Y., Moskvin, M., Mikhailovskii, V., Odintsova, G., Olshin, P., Pankin, D., Romanov, V. and Yatsuk, R. (2017) Metal surface coloration by oxide periodic structures formed with nanosecond laser pulses. *Optics and Lasers in Engineering*, **96**, 63–67.

Wang, J., Wang, Y., Yang, Y., Yang, R., Liao, W.-H. and Guo, P. (2020) Fabrication of structurally colored basso-relievo with modulated elliptical vibration texturing. *Precision Engineering*, **64**, 113–121.

Yamaguchi, H., Shinmura, T. and Ikeda, R. (2007) Study of internal finishing of austenitic stainless steel capillary tubes by magnetic abrasive finishing. *ASME Journal of Manufacturing Science and Engineering*, **129**, 885–892.

Yamashita, R., Nishio, M., Do, R.K.G. and Togashi, K. (2018) Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, **9**(4), 611–629.

Zhang, Q., Cao, R., Shi, F., Wu, Y.N. and Zhu, S.-C. (2018) Interpreting CNN knowledge via an explanatory graph, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI Press, New Orleans, LA, pp. 4454–4463.

Zhang, Y., Song, K., Sun, Y., Tan, S. and Udell, M. (2019) Why should you trust my explanation? Understanding uncertainty in LIME explanations, in *Proceedings of the International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, CA. arXiv preprint *arXiv:1904.12991*.

Zheng, H., Lim, G., Wang, X., Tan, J. and Hilfiker, J. (2002) Process study for laser-induced surface coloration. *Journal of Laser Applications*, **14**(4), 215–220.