Acta Materialia 295 (2025) 121132

Contents lists available at ScienceDirect

Acta Materialia

journal homepage: www.elsevier.com/locate/actamat

Full length article

Machine learning pipeline for Structure–Property modeling in Mg-alloys using microstructure and texture descriptors

Mahish K. Guru [a,*], Jan Bohlen [a], Roland C. Aydin [a,c], Noomane Ben Khalifa [a,b]

[a] Institute of Material and Process Design, Helmholtz-Zentrum Hereon, Geesthacht, Germany
[b] Institute for Production Technology and Systems, Leuphana Universität Lüneburg, Lüneburg, Germany
[c] Institute of Continuum and Materials Mechanics, Hamburg University of Technology (TUHH), Hamburg, Germany

ARTICLE INFO

Keywords:
Structure–Property
Magnesium alloys
Microstructure
Texture
Descriptor
Grain boundary
Orientation distribution function
Property prediction
Machine learning
Deep learning

ABSTRACT

Identifying the relationships between material structure and mechanical properties has been crucial for accelerating the exploration of the material design space for advanced alloys. However, traditional approaches for magnesium (Mg) alloys often fall short in providing quantitative and broadly applicable structure–property linkages. To address this challenge, a comprehensive machine learning pipeline is presented for structure–property modeling in extruded Mg-alloys, leveraging both microstructure and texture descriptors derived from experimental data. The pipeline encompasses a robust workflow for data extraction from optical microscopy and X-ray diffraction, advanced image processing and deep learning techniques for microstructure binarization and grain statistics, and the computation of statistical descriptors including n-point spatial correlations, gram matrices for microstructure, and generalized spherical harmonics (GSH) for texture. Dimensionality reduction techniques such as principal component analysis (PCA), isomap, and autoencoders are employed to manage the high-dimensionality of the descriptor space. Subsequently, non-linear regression models—Gaussian Process, XGBoost, and Multi-Layer Perceptron regressors—are evaluated to predict mechanical properties, specifically strain hardening exponent (**n**) and yield stress (σ_y). Our results demonstrate that XGBoost consistently outperforms other regressors, achieving a notably low mean absolute percentage error (MAPE) of 6.67% for strain hardening exponent and 7.01% for yield stress, using a combination of PCA-reduced 3-point spatial correlations and isomap-reduced gram matrices as microstructure descriptors, and isomap-reduced GSH coefficients as texture descriptors at a 150 μm length scale. Shapley Additive exPlanations (SHAP) analysis further reveals that texture descriptors and aspect ratio distribution are the most influential features in predicting mechanical properties. This established ML framework for structure–property modeling in Mg-alloys, surpasses state-of-the-art benchmarks and provides a valuable template for materials design and discovery.

1. Introduction

The integration of machine learning (ML) into material science has revolutionized the field, particularly in the domains of material and process design [1]. Machine learning algorithms, with their ability to analyze vast datasets and identify complex process–structure–property relationships, are showing critical importance in synthesizing new materials [2]. Simulation-driven material design has long played a crucial role in establishing structure–property relationships. Now, tremendous advancements in Integrated Computational Materials Engineering (ICME) are leveraging experimental data alongside material science principles and simulation to accelerate material design exploration [3]. The advent of data-driven and hybrid techniques has enabled researchers and engineers to tailor material properties to meet specific performance requirements [4]. Due to the applications of polycrystalline metallic materials in aerospace, automotive, bio-materials, renewable energy, and electronics, machine learning has been widely used to uncover structure–property relationships in these materials. Machine learning-embedded workflows have been used to develop structure–property relationships in steels [5], aluminum alloys [6], titanium alloys [7], copper alloys [8], and more.

Magnesium (Mg) alloys are another highly popular polycrystalline metallic material, used in aerospace, transportation, and biomedical applications due to their properties such as low corrosion resistance, high strength, and low density. The mechanical properties of wrought magnesium alloy (produced through mechanical processes such as

* Corresponding author.
E-mail address: mahish.guru@hereon.de (M.K. Guru).

https://doi.org/10.1016/j.actamat.2025.121132
Received 7 February 2025; Accepted 9 May 2025
Available online 28 May 2025

rolling, forging, and extrusion) can be tailored by setting the processing parameters appropriately, such as temperatures and processing times [9]. For these reasons, Mg-alloys have been extensively studied as functional materials in recent years as lightweight material for replacement of steel and aluminum in the aerospace and automobile industry. However, the widespread industrial adoption of Mg-alloys has been hindered by their low formability, as well as ability to adjust strength, resulting in the low yield strength (less than 300 MPa) of most commercial Mg-alloys [10]. Consequently, significant efforts have been undertaken to enhance the strength of magnesium alloys by thoroughly understanding the effects of various fabrication processes. Extrusion has been employed as a forming technique for many polycrystalline metallic materials to produce fine-grained microstructures. For industrial Mg-alloys, it has been found that by adjusting the temperature and speed of extrusion, the tensile yield strength (TYS) can be significantly enhanced [11].

Numerous experimental studies have been conducted to understand the effects of alloying and to correlate microstructure and textures with mechanical and formability properties of extruded Mg-alloys. The impact of rare earth elements, such as cerium (Ce) and neodymium (Nd), on the microstructure and texture evolution of magnesium–manganese (Mg–Mn) alloys during extrusion, and their subsequent influence on mechanical properties, has also been investigated. The study reveals that these elements weaken the typically strong recrystallization textures seen in the base alloy. Among them, neodymium proves to be the most effective in modifying the texture by reducing the formation of grains with orientations unfavorable for deformation. This modification leads to enhanced ductility and a reduction in asymmetric yield behavior, ultimately resulting in lower yield strength and ultimate tensile stress [12]. In a study conducted by [13], direct extrusion of flat magnesium alloy bands was used to investigate the effects of alloying and processing on microstructure and texture development. The study focused on how extrusion speed influenced dynamic recrystallization, grain growth, and texture modification in alloys like AZ31, ZX10, and ZN10. By adjusting alloying elements such as Nd and Ca, the study achieved weaker basal textures, characterized by a less pronounced alignment of basal planes. This texture weakening significantly enhances mechanical properties like ductility and formability, particularly in biaxial tests, by allowing easier activation of slip systems. Another study investigated the effect of silver (Ag) alloying in magnesium (Mg) on microstructure and mechanical properties during indirect extrusion. It found that increasing Ag content enhanced tensile yield strength (TYS) and ultimate tensile strength (UTS) by promoting non-basal slip mechanisms, improving elongation. The development of a weak basal texture, where basal planes align with the extrusion direction, facilitated these mechanical improvements by enhancing ductility and reducing twinning during compression [14].

While these experimental studies provide deep insights into the correlations between microstructure, textures, and mechanical properties of extruded Mg-alloys, their primary focus is to offer knowledge-based insights into material behavior by revealing dominant metallurgical mechanisms. However, these approaches do not provide a quantitative overview for a broader range of alloying or processing conditions. Additionally, these studies are often expensive and time-consuming, leading to prolonged material development timelines. Despite these challenges, experimental studies remain a crucial source of primary data. However, there is a pressing need for a more quantified approach to studying these correlations using machine learning based methodologies, enabling the development of more precise and efficient ways to analyze structure–property relationships [4].

A comprehensive Gaussian process (GP)-based workflow was developed to predict the process–structure–property (PSP) linkages in metal additive manufacturing, specifically focusing on laser powder bed fusion processes using 316L stainless steel. The methodology involved creating functional GPs trained on a limited set of experimental and simulation data to predict melt pool characteristics, microstructure features, and mechanical properties. The study found that the GP-based framework could achieve approximately 95% accuracy in predicting grain stress and strain for previously unseen process parameter combinations [15]. Different machine learning methods have also been previously compared by [16] in establishing structure–property linkages for high-contrast 3D elastic composites. They used finite element simulations to generate a dataset of voxelized microstructures and applied principal component analysis for dimensionality reduction. The study compared local/global and parametric/non-parametric approaches, with a focus on Gaussian Process (GP) models to predict the FE estimated effective stiffness. The findings revealed that the locally approximated Gaussian Process model achieved the highest accuracy, with a mean absolute error (MAE) of 1.87 GPa and a mean squared prediction error (MSPE) of 2.68 GPa. A reduced order structure–property (S–P) linkage framework for polycrystalline microstructures of α-titanium was developed by [17] using 2-point spatial correlations, generalized spherical harmonics (GSH), and principal component analysis (PCA) for compact representation of microstructure. They tested this framework on data generated using crystal plasticity finite element (CPFE) simulation. The findings showed that the framework could predict elastic stiffness with an average error of about 0.2% and yield strength with an average error of about 1.5%. While all the aforementioned approaches successfully link microstructure features (and in one case, texture features [17]) to mechanical properties with high accuracy, these accuracies are primarily derived from testing on simulation datasets or a combination of simulation and experimental datasets. This raises the question of whether the simulation datasets are sufficiently grounded in real-world experimental data, which is crucial for accelerating material design exploration. This is particularly important for leveraging the experimental data for industrial Mg-alloys.

Machine learning has indeed also been applied to predict properties in Mg-alloys. An in-depth study was conducted by [18] to predict the mechanical properties using experimental data collected from published literature. This data was used to establish a comprehensive database of descriptors, including processing parameters (such as rolling temperature, extrusion temperature, and extrusion ratio), alloying element properties, thermodynamic parameters. Finally, several ML models were employed to predict ultimate tensile strength (UTS), elongation (EL), yield strength (YS), and hardness (HV). Gradient Boosting Regression (GBR) achieved high accuracy for UTS, YS, and HV achieving an R^2 value of more than 0.8 on the test set, although the prediction for EL was less satisfactory. Despite the adequate accuracies achieved, the descriptor space was primarily composed of thermal processing parameters and elemental properties, with no information about the microstructure or texture of the material. This now highlights the need for a structure–property (S–P) linkage approach that is independent of thermal processing parameters or alloying elemental properties. Such an approach would ensure a more robust and generalizable model, capable of accurately predicting mechanical properties across a wider range of conditions and compositions. Another ensemble-based approach utilizing shallow artificial neural networks (ANNs) has been applied by [19] for extruded magnesium–gadolinium (Mg–Gd) alloys to predict grain size, tensile yield stress, compressive yield stress, ultimate tensile strength, and hardness. The models were trained on a dataset where descriptors included various process parameters, such as extrusion velocity, temperature, and gadolinium content. However, the models exhibited limited generalizability, with their accuracy declining when predicting properties beyond the measured process parameter space. This limitation can be attributed to the insufficiency for processing parameters and alloying content as a complete indicator of these properties. Additionally, the models only utilized grain size as a descriptor of microstructure and the texture is only simply considered as 3 pole related intensities rather than the representative orientation distribution.

Therefore, in this paper,

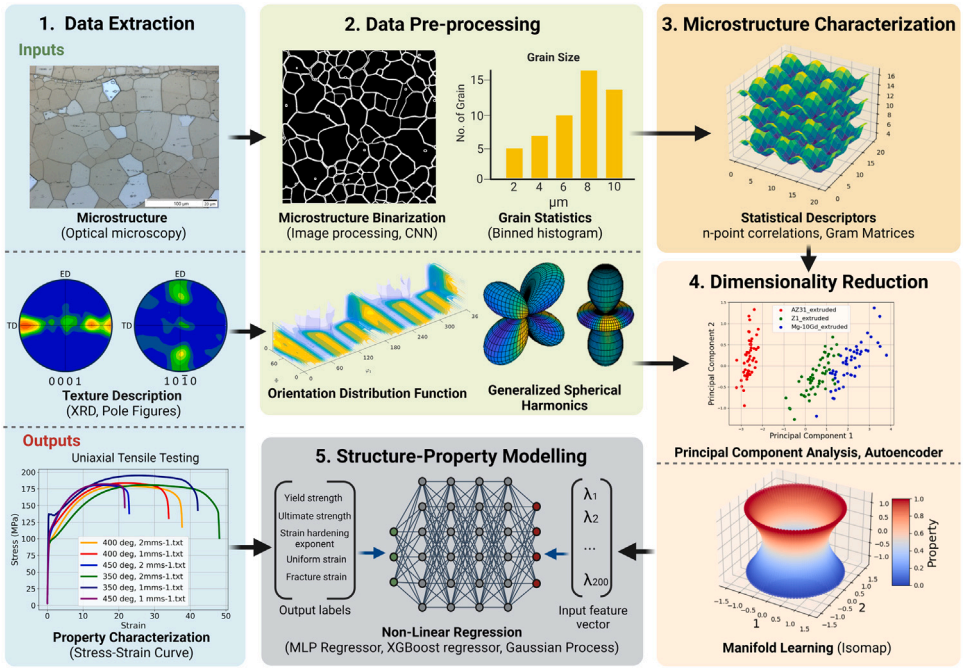

Fig. 1. Pipeline for Structure–property modeling of Mg-alloys with 5 key modules: 1. Data extraction, input–output database creation, 2. Pre-processing the data for augmentation, information enrichment and feature engineering, 3. Microstructure characterization using statistical descriptors, 4. Dimensionality reduction of the calculated descriptors and 5. Structure–Property modeling to predict mechanical properties from the input lower order descriptors.

- We develop a database and a generalized approach for extruded Mg-alloys that bases the prediction of mechanical properties entirely on descriptors of microstructure and texture, as illustrated in the pipeline in Fig. 1.
- We investigate various statistical descriptors for the most accurate and efficient characterization of microstructure, the orientation distribution of the microstructure grains and grain statistics.
- We explore dimensionality reduction techniques to create a reduced-order structure–property (S–P) linkage.
- We examine various non-linear regression models to achieve the most accurate and efficient prediction of mechanical properties.

2. Structure–property linkage framework

To develop a quantified and reduced-order approach for the structure–property relationship, efficient characterization, dimensionality reduction, and model selection are essential components of the pipeline. This necessitates a thorough investigation of these components to tailor the approach to specific use cases while maintaining generalizability across different use cases or material types. For instance, our pipeline utilizes experimental data from various extruded Mg-alloys, following a specific protocol for data on microstructure and texture features. However, efforts are made to achieve a more generalized structure–property approach for characterization, dimensionality reduction, and model selection, applicable not only to one specific data protocol but also to other widely available data protocols. To this end, we made a conscious decision to accommodate optical microscopy (OM) images for micrographs (with minimal effort required to extend the approach to scanning electron microscopy (SEM) images) and X-ray diffraction (XRD) data for texture (which can also be extended to electron backscatter diffraction (EBSD) data). These data types are more accessible and economical to acquire, yet they contain sufficient information indicative of the mechanical properties. By leveraging OM and XRD data, we ensure that the pipeline remains cost-effective and widely applicable across diverse datasets. This adaptation enhances the robustness and applicability of the method across various material systems and alloying elements, making it independent of specific compositions [20].

Microstructure information is extracted as the first part of the data extraction block (Fig. 1) from the experimental data using specific routines to obtain the micrographs. These micrographs, which are images obtained through metallography of the sample and observed under OM, reveal grain boundaries and grain structure. They often include a scale (generally in μm) for spatial and scale perspective [21]. More information about this data is given in Section 3. In the next stage of the pipeline, these micrographs undergo preprocessing to maintain uniform scale by cropping and tiling, ensuring data uniformity and coherence in representation, which is essential for making automated predictions using machine learning.

For a quantified structure–property (S–P) linkage, raw microstructure images cannot effectively characterize the microstructure due to the presence of a lot of unimportant information in direct pixel-to-pixel comparisons. To address this, the images are binarized to create a clearer and more computer-readable format. This process is particularly advantageous for machine learning applications as it reduces noise, and enriches the information about grain boundary structure and topology, making it easier for algorithms analyze the information [22]. Workflow used for microstructure binarization for our this pipeline is discussed in Section 3.2. In this pipeline, we focus on single-phase Mg-alloys, treating any other phase as an inclusion with a pixel value of 1. Additionally, grains are segmented using the binarized images of the microstructure to obtain statistics for grain size and grain aspect ratios. These statistics are then represented in the form of binned histograms. Conventionally, only these statistics have been employed to characterize microstructure features [23,24]. However, they cannot fully capture the key features of grain morphology, which are crucial for property prediction applications. To address this, extensive research has been conducted to develop algorithms that calculate statistical descriptors for microstructure characterization [25–27]. An n-point spatial correlation function (or n-point statistics) is a statistical descriptor used to capture the spatial arrangement of grains and grain boundaries in polycrystalline materials. It can represent complex patterns and arrangements, including long-range patterns and connectivity features [28]. However,

this method requires higher-dimensional representation and is more susceptible to noise in the data. In contrast, a Convolutional Neural Network (CNN) based descriptor, gram matrices are generally used for a more compact representation of complex spatial relationships of grains and grain boundaries in a polycrystalline microstructure. They provide direct geometric insights into the microstructure features, offering a description of the overall geometry of the grain boundaries in the microstructure image. This method also helps capture any anisotropy present in the material [29,30]. Spatial correlations capture local spatial arrangements, while gram matrices provide a global view of grain boundary features. The implementation of both descriptors is discussed in Section 4. Considering the advantages and disadvantages of grain statistics, n-point spatial correlations, and gram matrices, this research will evaluate their individual and combined performance in predicting mechanical properties during the extrusion of Mg-alloys.

Another part of the first phase of the S–P pipeline shown in Fig. 1 is the extraction of texture description from the experimental data. This texture description is available in the form of metadata of pole figures, which are obtained from X-ray diffraction (XRD) experiments using a Malvern-Panalytical file structure (.xrdml). X-ray diffraction is a crucial tool for studying the atomic structure of polycrystalline materials. XRD is based on the interference of X-ray waves elastically scattered by a series of atoms oriented along particular lattice plains in a crystal [31]. XRD is used for obtaining pole figures, which are essential for determining if a material exhibits a texture, i.e., the orientation of the grains in the material. Pole figures display the accumulation of poles around specific directions relevant to the material's processing, such as the normal direction (ND), extrusion direction (ED), and transverse direction (TD). For instance, in rolled metals, certain crystallographic directions may align preferentially due to the deformation processes involved during manufacturing [32]. These crystallographic orientations of grains are very indicative of the mechanical performance of the material. While pole figures are useful for visualizing these orientations and making quick assessments, they lack the depth of information needed for a data-driven approach to property prediction. Therefore, pole figures are further processed to calculate the Orientation Distribution Function (ODF), which captures the number of grains oriented in specific directions, expressed in terms of Euler angles [33]. The ODF is more suitable for characterizing complex orientations and enables more detailed orientation feature extraction using machine learning. When represented in Bunge-Euler space $\{g = (\varphi_1, \Phi, \varphi_2) \mid 0 \leq \varphi_1 < 2\pi,\ 0 \leq \Phi \leq \pi/2,\ 0 \leq \varphi_2 < \pi/6\}$ with 5-degree increments, the ODF is generally high-dimensional data. For compactness, ODFs are most commonly represented in terms of expansion coefficients of GSH. More specifics about the implementation for converting pole figures to ODFs and then to GSH coefficients in our pipeline will be explained in Section 5. Further details about the texture descriptors and the calculation of GSH coefficients can be found in Section 5. It is also important to note that setting the upper limit for truncating the series expansion of GSH, L = 22, is sufficient to smoothly capture the orientation density [34]. However, this still results in more than 2000 coefficients, which is a high number for our intended black-box modeling [33].

Descriptors used in the pipeline for microstructure and texture are high-dimensional data. With high-dimensional data comes the curse of dimensionality, which makes it difficult for machine learning algorithms to learn meaningful patterns in these descriptors, as they can overfit to unimportant patterns [35]. Therefore, dimensionality reduction techniques such as PCA and manifold learning are employed to identify principal dimensions for n-point statistics and gram matrices used as microstructure descriptors. Additionally, Autoencoders and manifold learning are applied to reduce the dimensions of the direct ODF descriptions and the compact representations with GSH coefficients. In Section 6, the implementation for these dimensionality reduction techniques are discussed. Furthermore, to determine the most effective dimensionality reduction method for our processed data types, these techniques are compared against each other based on their performance in predicting mechanical properties.

In this study, data extracted from the experimental sets for mechanical properties is in the form of stress–strain curves, which is characterized into scalar values that best represent the mechanical performance of the alloy using specific routines. These scalar values include Yield Strength ($\boldsymbol{\sigma_y}$), Ultimate Strength ($\boldsymbol{\sigma_{max}}$), Strain Hardening Exponent ($\mathbf{n}$), Uniform Strain ($\boldsymbol{\varepsilon_{pu}}$), and Fracture Strain ($\boldsymbol{\varepsilon_{pf}}$). These values are crucial for the material design workflow in both practical and industrial stages of material development and are discussed in detail in Section 3.3. They are used as labels for the black box modeling of phase 5. Structure–property models (see Fig. 1 of the pipeline. We investigate the performance of non-linear regression techniques, including Multi-Layer Perceptron (MLP), XGBoost (based on random forest regression), and Gaussian Process (GP) Regressor, to model the complex relationships between features (microstructure and texture descriptors) and the labels (mechanical properties). These models are evaluated for their ability to efficiently and accurately predict mechanical properties, facilitating efficient industrial material design process for extruded Mg-alloys.

3. Dataset and materials

For every supervised machine learning-based pipeline, a coherent and well-labeled dataset with uniform structure is the backbone. When modeling complex non-linear relationships between data and labels, the accuracy of the model depends significantly more on the quality and coherence of the input data than on the preprocessing and selection techniques [36]. In this section, we will detail the materials used in this study, specifically the Mg-alloys, along with their sources and the specific data collected for each alloy. Furthermore, we will elaborate on the preprocessing techniques performed on the data, such as microstructure binarization, grain state analysis, ODF characterization, and GSH coefficient calculation for texture data. The methods of implementation for these preprocessing steps will be thoroughly discussed. Additionally, we will address the data labels used in this pipeline, specifically the stress–strain data along with the scalar values and their significance in the context of material design and property prediction.

3.1. Material data

Data were extracted for various Mg-alloys from previous experimental studies aimed at understanding the effect of the extrusion process on microstructure, texture, and consequently, mechanical properties. The selected studies were published and met specific criteria to ensure uniformity in process variables, namely extrusion velocity ($\mathbf{v_{ext}}$) and extrusion temperature ($\mathbf{T_{ext}}$). Further criteria for the availability of microstructure information as images from OM (optionally, also from EBSD) and availability texture descriptions in the form of raw data from XRD (optionally, from EBSD) for pole figures. Finally, stress–strain data (.csv files) for each specimen, for which the aforementioned data were available, were included to create a coherent, well-structured, and labeled dataset, as shown in Table 1.

The data for a wrought magnesium alloy AZ31, which contains approximately 2.5% to 3.5% aluminum, 0.7% to 1.3% zinc, and a minimum of 0.2% manganese, were extracted from a doctoral thesis. In this thesis, flat bands of AZ31 were processed at extrusion speeds of 0.6 and 2.4 mm/s and extrusion temperatures of 200, 250, 300, 350, 400, 450 and 500 °C, with and without heat treatment (HT) post extrusion, and subsequently analyzed for their microstructure, texture, and mechanical properties. The mechanical properties were evaluated via uni-axial tensile testing [37]. Furthermore, similar data were collected for zinc-based magnesium alloys with additions of calcium (Ca) and neodymium (Nd) from a study conducted by [38]. These alloys were extruded to obtain round bars under extrusion temperatures of 250, 300, 350 and 400 °C and extrusion velocities of 2, 5 and 7.5 mm/s.

Table 1
Material data with Mg-alloy names and corresponding extrusion process parameters, type of microstructure and texture data available, the range of mechanical properties [13,37–41].

Alloy Name	v_{ext} (mm/s)	T_{ext} (°C)	Microstructure ($IMG = [a_{i,j}]_{i,j \in M}$)	Texture ($ODF = f(\varphi_1, \Phi, \varphi_2)$)	σ_y (MPa)	σ_{max} (MPa)	n	ε_{pu}	ε_{pf}
AZ31	0.6, 2.4	200 – 500	OM, EBSD	XRD, EBSD	143 – 184	240 – 278	0.13 – 0.23	0.10 – 0.18	0.13 – 0.25
AZ31(HT10min450)	0.6, 2.4	200 – 500	OM, EBSD	XRD, EBSD	140 – 180	245 – 265	0.14 – 0.23	0.13 – 0.17	0.17 – 0.24
ME21	0.75 – 7.5	300 – 450	OM	XRD	154 – 237	228 – 270	0.09 – 0.19	0.08 – 0.15	0.14 – 0.22
Mg-2Gd	0.5 – 2	350 – 450	OM	XRD	87 – 208	179 – 226	0.10 – 0.31	0.07 – 0.26	0.09 – 0.47
Mg-2Gd-0.5Mn	0.5 – 2	350 – 450	OM	XRD	104 – 184	180 – 216	0.13 – 0.23	0.07 – 0.19	0.09 – 0.31
Mg-2Gd-1Mn	0.5 – 2	350 – 450	OM	XRD	105 – 183	188 – 221	0.13 – 0.23	0.13 – 0.17	0.17 – 0.30
Mg-5Gd	0.5 – 2	350 – 450	OM	XRD	91 – 193	240 – 278	0.12 – 0.28	0.12 – 0.26	0.14 – 0.35
Mg-5Gd-0.5Mn	0.5 – 2	350 – 450	OM	XRD	109 – 173	198 – 225	0.16 – 0.23	0.13 – 0.19	0.15 – 0.26
Mg-5Gd-1Mn	0.5 – 2	350 – 450	OM	XRD	114 – 187	206 – 237	0.15 – 0.22	0.15 – 0.18	0.21 – 0.27
Mg-10Gd	0.5 – 2	350 – 450	OM	XRD	132 – 232	242 – 296	0.16 – 0.22	0.18 – 0.24	0.22 – 0.32
Mg-10Gd-0.5Mn	0.5 – 2	350 – 450	OM	XRD	138 – 212	248 – 286	0.17 – 0.20	0.18 – 0.20	0.22 – 0.26
Mg-10Gd-1Mn	0.5 – 2	350 – 450	OM	XRD	154 – 232	262 – 296	0.16 – 0.19	0.18 – 0.20	0.24 – 0.26
Z1	2 – 7.5	250 – 400	OM, EBSD	XRD, EBSD	125 – 149	216 – 235	0.19 – 0.23	0.07 – 0.09	0.16 – 0.24
ZNd10	2 – 7.5	250 – 400	OM, EBSD	XRD, EBSD	79 – 282	203 – 287	0.08 – 0.33	0.09 – 0.23	0.22 – 0.48
ZNd10(HT10min450)	2.4	350	OM, EBSD	XRD, EBSD	72	197	0.32	0.22	0.31
ZX10	2 – 7.5	250 – 400	OM, EBSD	XRD, EBSD	77 – 132	194 – 225	0.23 – 0.34	0.13 – 0.25	0.25 – 0.44
ZX10(HT10min450)	0.6, 2.4	300	OM, EBSD	XRD, EBSD	62 – 69	184 – 189	0.36 – 0.37	0.25 – 0.27	0.35 – 0.39

The three alloys are designated as follows: the Mg-Zn alloy is named Z1, the Mg-Zn-Ca alloy is named ZX10, and the Mg-Zn-Nd alloy is named ZNd10. Additional data for AZ31, Z1, ZX10, and ZNd10, as well as heat-treated ZX10 and ZNd10, were collected from a study that investigated the effect of varying processing parameters on the mechanical properties of flat bands of Mg-alloys at extrusion speeds of 0.6 and 2.4 mm/s [13]. Data for ME21 alloys, a specialized aluminum-free magnesium alloy (Mg–2Mn–0.6Ce–0.3) was extracted from a study that investigated the effect of varying extrusion parameters on tailoring mechanical properties with respect to microstructure and texture. The extrusion temperature was varied at 300, 350, 400, and 450 °C, and the extrusion speed was varied at 0.75, 1.4, 2.8, 5.5, and 7.5 mm/s [39]. The minimum and maximum tensile yield strengths for ME21 were found to be 228 MPa and 270 MPa, respectively (see Table 1). Another specialized class of Mg-alloys, which is very popular for lightweight and biomedical applications, are Mg–Gd alloys due to their corrosion resistance and high strength. Data for Mg-2Gd, Mg-5Gd, and Mg-10Gd were taken from a study by [40], which investigated the effect of extrusion rates (0.5, 1, and 2 mm/s) and extrusion temperatures (350, 400, and 450 °C) on mechanical properties. This study achieved a range of tensile yield strengths ($\boldsymbol{\sigma_y}$) from 180 to 280 MPa and tensile maximum strengths ($\boldsymbol{\sigma_{max}}$) from 300 to 450 MPa. The final addition to the dataset includes data for Mg–Gd alloys with manganese (Mn). This data was taken from a study that investigated Mg-2Gd-0.5Mn, Mg-5Gd-0.5Mn, Mg-10Gd-0.5Mn, Mg-2Gd-1Mn, Mg-5Gd-1Mn, and Mg-10Gd-1Mn. The study examined the effects of varying extrusion rates (0.5, 1, and 2 mm/s) and extrusion temperatures (350 and 450 °C) on microstructure, texture properties mechanical performance of these alloys [41].

3.2. Microstructure binarization and statistics

The OM data presented in Table 1 includes the following conditions: AZ31 with an extrusion speed ($\mathbf{v_{ext}}$) of 0.6 mm/s and an extrusion temperature ($\mathbf{T_{ext}}$) of 500 °C; ME21 with $\mathbf{v_{ext}}$ = 5.5 mm/s and $\mathbf{T_{ext}}$ = 450 °C; and Mg-5Gd-1Mn with $\mathbf{v_{ext}}$ = 1 mm/s and $\mathbf{T_{ext}}$ = 450 °C. These data are illustrated at scales of 200, 100, and 100 μm, respectively, in Fig. 2. Upon initial inspection, it is evident that the images contain a high density of information, including a significant amount of extraneous data. To address this, it is crucial to first maintain a uniform scale to achieve a coherent image dataset for microstructure analysis, ensuring a consistent micrometer-to-pixel ratio ($\mu m/px$). Currently, for example, Figs. 5(a) and 5(b) have different $\mu m/px$ ratios. Therefore, we need to crop tiles from images according to these ratios to create uniformly scaled microstructure images. For this purpose, tiles will be cropped at sizes of 90 × 90, 100 × 100, 120 × 120, and 150 × 150 μm and resized to 256 × 256 pixels. These four OM microstructure image datasets will also be compared against each other to understand the effect of scale selection on the mechanical property prediction for the extruded Mg-alloy dataset.

To further reduce the information in these uniformly scaled images and enhance consistency, it is necessary to extract grain boundaries and inclusions. We will adapt and employ previously published image processing and machine learning techniques that have been successfully applied to OM-based microstructure images [42,43]. Both of these approaches utilize pixel value-based image thresholding to enhance contrast, edge detection for grain boundary segmentation, CNN-based grain segmentation and various other image processing operations. These are employed in a specific order and combination to generate binary images, where pixel values of 1 represent grain boundaries and 0 represent all other areas. These methods also incorporate statistical analysis of the segmented grains to generate grain statistics, such as aspect ratio and equivalent diameter for each pixel. The microstructure binarization approach used in this research is illustrated in Fig. 3. This approach involves two separate workflows: one for segmenting and binarizing the grain boundaries, including calculating grain size and aspect ratio statistics, and another for segmenting out the inclusions (black blobs in Fig. 2, which may be non-metallic particles or other alloy phases). The final binary microstructure image is obtained by combining these two segmentations.

In the first workflow, the uniformly scaled microstructure image is initially denoised to eliminate any background jittering of the pixels. This is achieved using an image processing technique called non-local means (NLM) denoising. NLM denoising replaces each pixel's value with a weighted average of similar pixels from across the entire image, not just nearby ones. This technique was introduced by [44] in their 2005 paper "A Non-Local Algorithm for Image Denoising". The denoised value at pixel p,

$$u(p) = \frac{1}{C(p)} \int_{\Omega} v(q) f(p,q)\, \mathrm{d}q, \tag{1}$$

where $v(q)$ is the original pixel value, $f(p,q)$ is a similarity weight function, and $C(p)$ is a normalizing factor.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is subsequently applied to enhance the contrast of the images while preventing noise amplification in homogeneous areas. This step is crucial for the subsequent grain segmentation process, as it helps to clearly distinguish between grains and grain boundaries. Unlike standard histogram equalization, CLAHE restricts the amplification of noise by clipping the histogram at a predefined threshold at a clip limit. The CLAHE thresholding function at pixel p,

$$T(I(p)) = \mathrm{CDF}(I(p)) \cdot \text{clip limit}, \tag{2}$$

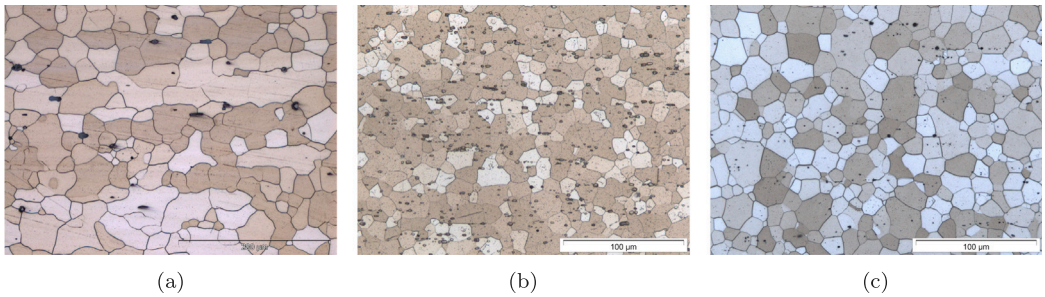

Fig. 2. Sample microstructure images from the dataset in Table 1, generated via Light Optical Microscopy. (a) AZ31, $\mathbf{v}_{\mathbf{ext}} = 0.6\,\text{mm/s}$, $\mathbf{T}_{\mathbf{ext}} = 500\,^{\circ}\text{C}$, scale $= 200\,\mu\text{m}$, $\mu\text{m}/px = 0.178$. (b) ME21, $\mathbf{v}_{\mathbf{ext}} = 5,5\,\text{mm/s}$, $\mathbf{T}_{\mathbf{ext}} = 450\,^{\circ}\text{C}$, scale $= 100\,\mu\text{m}$, $\mu\text{m}/px = 0.107$. (c) Mg-5Gd-1Mn, $\mathbf{v}_{\mathbf{ext}} = 1\,\text{mm/s}$, $\mathbf{T}_{\mathbf{ext}} = 450\,^{\circ}\text{C}$, scale $= 100\,\mu\text{m}$, $\mu\text{m}/px = 0.107$.

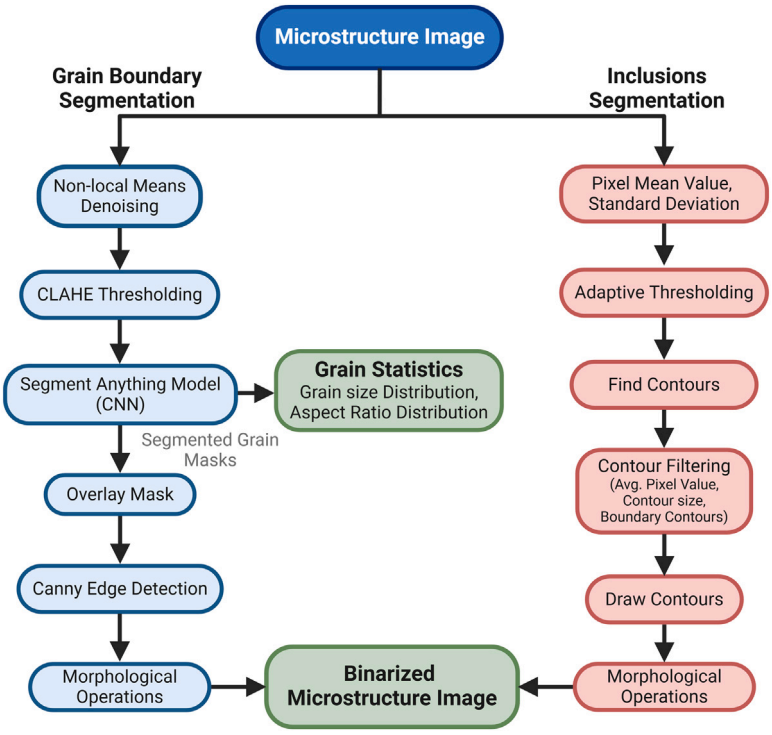

Fig. 3. Microstructure Binarization workflow with calculation of grain statistics via grain boundary segmentation (blue) and inclusion segmentation (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where CDF($I(p)$) is the cumulative distribution function calculated from the histogram of the local tile and clip limit is typically the maximum pixel value (e.g., 255 for an 8-bit image) [45]. This method is particularly effective for OM imaging due to varying lighting conditions, as it improves visibility without over-amplifying noise.

The zero-shot capability of the Segment Anything Model (SAM) from Meta is utilized to segment grain masks from the high-contrast microstructure images. SAM automatically adapts to different imaging conditions without requiring any labeled training data [46]. This makes SAM particularly suitable for microstructure grain segmentation in our dataset, as it effectively handles images with varied grain topology and lighting conditions. An implementation of the Segment Anything Model (SAM) is used to generate grain masks from the thresholded images, as illustrated by the three examples in Fig. 4. SAM creates a mask for each grain, and these masks are then overlaid to produce an image with all segmented grains, each colored differently. The area and bounding box of each mask are used to calculate statistics for each micrograph, specifically the grain size distribution, which includes the equivalent diameter and aspect ratio distribution of the grains. These statistics are presented in the form of binned histograms and are used as input for the final structure–property (S–P) model. The values for three different specimen micrograph at scale $150 \times 150\,\mu\text{m}$, are compared in binned histograms, as depicted in Fig. 5. A template of 30 bins ranging from $5 - 200\,\mu\text{m}$ for grain boundary distribution and 3 bins ranging from 0–5 for aspect ratio distribution is used to maintain data coherence. This approach facilitates the S–P model's ability to learn and recognize patterns effectively.

Next, to generate the binarized grain boundary image from the overlaid mask image, a multi-stage algorithm called Canny edge detection is used to effectively identify the edges of the grain masks. The algorithm computes the intensity gradients of the smoothed image using derivative filters (such as Sobel filters). This results in two gradient components, (G_x) and (G_y), which represent changes in intensity in the horizontal and vertical directions, respectively. The magnitude of the gradient is given by:

$$G = \sqrt{G_x^2 + G_y^2}, \tag{3}$$

which signifies the strength of the edge, while the direction is given by:

$$\theta = \tan^{-1}\left(\frac{G_y}{G_x}\right). \tag{4}$$

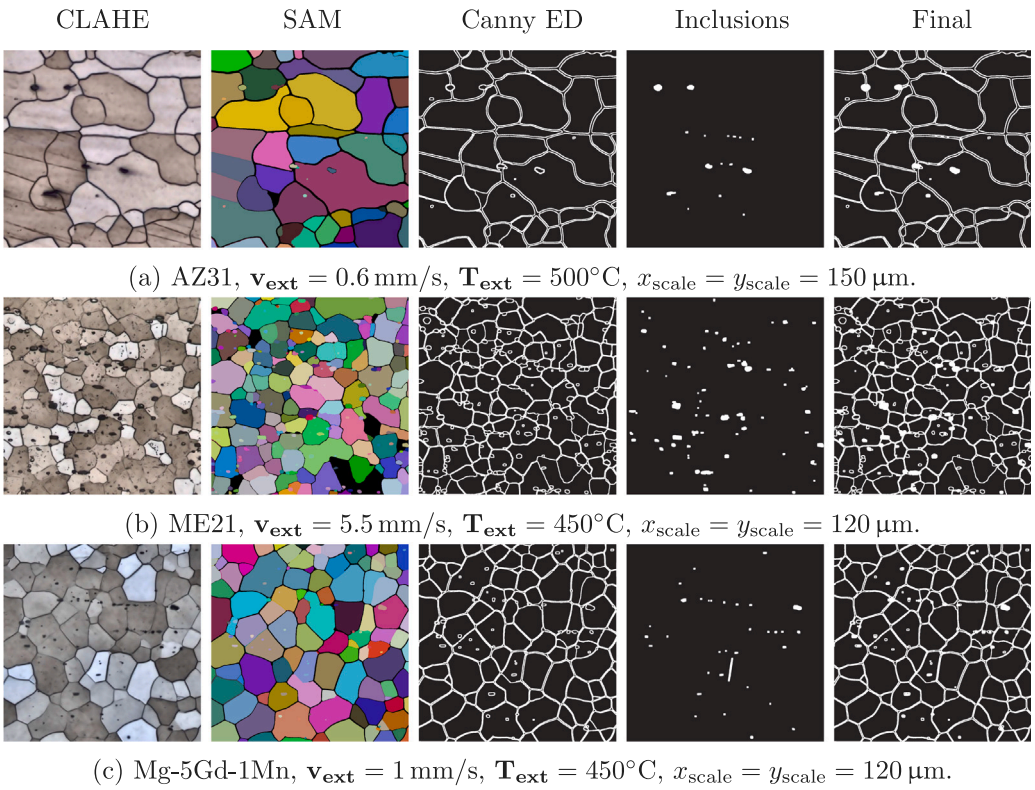

Fig. 4. Three examples of Mg-alloy sample images at different stages of the microstructure binarization workflow.

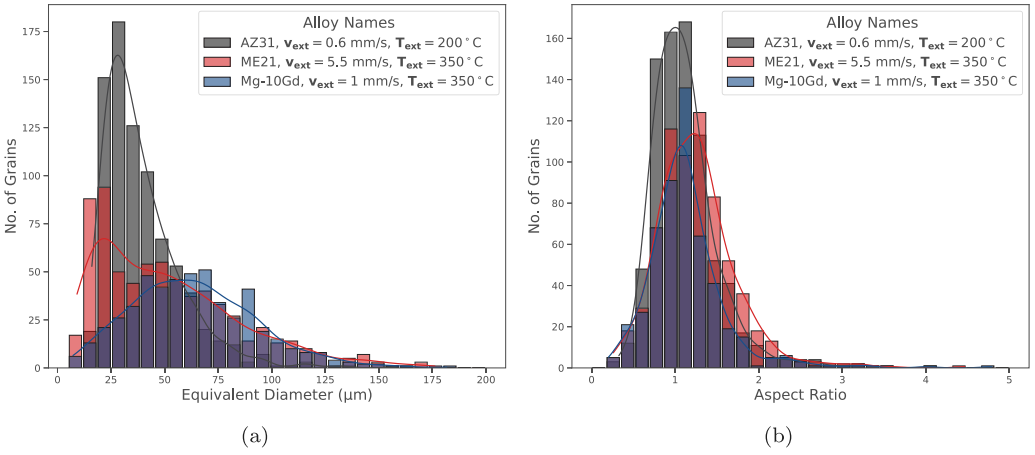

Fig. 5. Microstructure statistics for scale = 150 μm. (a) Grain size distribution and (b) Aspect ratio distribution for AZ31 ($\mathbf{v_{ext}} = 0.6\,\mathrm{mm/s}$, $\mathbf{T_{ext}}$ = 500 °C), ME21 ($\mathbf{v_{ext}} = 5,5\,\mathrm{mm/s}$, $\mathbf{T_{ext}}$ = 350 °C) and Mg-10Gd ($\mathbf{v_{ext}} = 1\,\mathrm{mm/s}$, $\mathbf{T_{ext}}$ = 450 °C).

The next step is Non-Maximum Suppression, which thins out the edges by retaining only local maxima in the gradient direction, effectively suppressing non-edge pixels. Finally, two thresholds are applied to determine which edges are significant. Pixels with gradient magnitudes above a high threshold are marked as edges, while those below a low threshold are discarded [47].

Finally, morphological operations such as erosion and dilation of pixel boundaries are applied to the detected edges at the end of the binarization process. These operations help to form a clean, binarized grain boundary image (as seen in Canny ED images from Fig. 4), which can then be used to derive statistical descriptors.

In the second workflow (red workflow from Fig. 3) for inclusion segmentation, the mean pixel value $\bar{I}_{\mathrm{pixel}}$ and the standard deviation of the pixel value $\sigma_{I_{\mathrm{pixel}}}$ are calculated for the cropped, uniformly scaled micrograph. These values are then used for adaptive thresholding to create binary image of inclusion. Adaptive thresholding is a technique in image processing that segments an image by dynamically determining the threshold value for each pixel based on the local characteristics of its neighborhood. The algorithm examines a small region around each pixel to compute a local threshold. The local threshold T can be calculated using:

$$T = \mathrm{mean}(I_L) - C, \tag{5}$$

where $\mathrm{mean}(I_L)$ is the mean intensity of the local neighborhood around pixel p and C is a constant, which is determined differently for each micrograph using $\bar{I}_{\mathrm{pixel}}$ and $\sigma_{I_{\mathrm{pixel}}}$.

To identify inclusions (which appear as black blobs) in the thresholded image, we use the 'findContours' function from the OpenCV library. This powerful tool detects contours of pixel islands in images. Contours are simply curves that join all continuous points along a

boundary with the same intensity. These contours are stored in a hierarchy based on their detection and retrieval order. Once detected, the contours and their hierarchy are filtered based on the average pixel value (intensity) of each contour being greater than 200 and the area of the contour being greater than 5. Additionally, any contours touching the rectangular boundary of the image are excluded from the final selection. All the contours are then drawn on a black background and filled with pixels. To obtain clean binary inclusions, further morphological operations (such as erosion and dilation) are performed, as shown in Fig. 4 under "Inclusions". The two binary images (one with grain boundaries and the other with inclusions) are then overlaid to create the finalized binary microstructure image. This image will be used to derive statistical descriptors of grain boundary features and inclusions, which will, in turn, be utilized by the structure–property (S–P) model to make property predictions.

3.3. Uni-axial tensile property analysis

Uni-axial tensile test evaluates material's ability to resist deformation under tension, providing crucial data such as yield strength ($\boldsymbol{\sigma_y}$), ultimate tensile strength ($\boldsymbol{\sigma_{max}}$), strain hardening exponent (**n**) [48], uniform strain ($\boldsymbol{\varepsilon_{pu}}$) and fracture strain ($\boldsymbol{\varepsilon_{pf}}$). Typical values for these properties for various Mg-alloys are presented in Table 1. These values are derived from the aforementioned studies where tensile tests were conducted on cylindrical rod samples with a diameter of 5 mm and a gauge length of 30 mm for the Z1, ZNd10, ZX10, Mg–Gd, and Mg–Gd-Mn alloys. For the AZ31, AZ31(HT10min450), ME21, ZNd10(HT10min450), and ZX10(HT10min450) alloys, flat specimens with a width of 40 mm and a thickness of 2 mm were used. All tests were performed at room temperature along the extrusion direction using a 50 kN testing machine (Zwick Z050) with a constant strain rate of $10^{-3}\ s^{-1}$.

The data obtained from these studies are presented in the form of engineering stress vs. strain tables, with yield point values assigned to each table. These yield point values are used to calculate the $\boldsymbol{\sigma_y}$. The maximum stress value in the tabular data is considered as $\boldsymbol{\sigma_{max}}$, and the corresponding plastic strain value is taken as the uniform strain. The plastic strain experienced by the material at the point of fracture is called fracture strain. It is determined for each table by identifying the point where either the data ends or the stress value starts to drop rapidly for each strain step. To characterize the plastic behavior of a material in the strain hardening region, we use the relation between true stress ($\boldsymbol{\sigma_t}$) and true strain ($\boldsymbol{\varepsilon_t}$),

$$\boldsymbol{\sigma_t} = \mathbf{K}\boldsymbol{\varepsilon_t^n}, \tag{6}$$

where K is the strength coefficient, which represents the stress level at a true strain of $\boldsymbol{\varepsilon_t} = 1$. **n** is the strain hardening exponent, which indicates the rate of hardening of the material; it reflects how quickly the material strengthens with increasing strain in the plastic region, it begins at $\boldsymbol{\sigma_y}$ and extends to $\boldsymbol{\sigma_{max}}$. The corresponding indices for these points are identified from the stress–strain data. Subsequently, engineering stress and strain values within the plastic region are converted into true stress ($\boldsymbol{\sigma_t}$) and true strain ($\boldsymbol{\varepsilon_t}$) using,

$$\boldsymbol{\sigma_t} = \boldsymbol{\sigma_e}(1 + \boldsymbol{\varepsilon_e}), \tag{7}$$

$$\boldsymbol{\varepsilon_t} = \ln(1 + \boldsymbol{\varepsilon_e}), \tag{8}$$

where $\boldsymbol{\sigma_e}$ and $\boldsymbol{\varepsilon_e}$ are the engineering stress and strain, respectively. To facilitate the calculation of **n** and **K**, these true stress and true strain values are then transformed into their logarithmic forms, yielding the expression for Eq. (6) as,

$$\ln(\boldsymbol{\sigma_t}) = \mathbf{n}\ln(\varepsilon_t) + \ln(\mathbf{K}). \tag{9}$$

A linear regression analysis is then conducted on the log-transformed data to determine the slope and intercept of the resulting line. The slope of this regression line corresponds to **n**, while the intercept provides the natural logarithm of **K**. This approach allows for the characterization of material hardening behavior in the plastic deformation region of the stress–strain curve.

4. Microstructure characterization

In this section, we will explain the statistical descriptors used to characterize the binarized microstructure images, along with the libraries and implementations employed to calculate them. We will discuss the n-point spatial correlation function, which captures the spatial correlation within the microstructure images, and the gram matrices, which extract feature information such as grain boundary texture and grain boundary patterns from the 256 × 256 pixel images across four scales.

4.1. N-point spatial correlation function

The n-point spatial correlation function is a statistical tool used to describe the spatial arrangement and relationships within spatially distributed set of points of a material's microstructure. This function quantifies how grain boundary features in a microstructure is positioned relative to each other at different distances. The term "n-point" refers to the number of points considered simultaneously, with higher values of n capturing more complex spatial relationships. The most common form for heterogeneous media is the 2-point function, which considers pairs of points and provides a measure of how the probability of finding a specific feature at one point is related to finding the same or another feature at a given distance and direction from it. The 2-point spatial correlation of a microstructure image $m[s, l]$ with s pixels and l phases (0 or 1 in our case due to the binary microstructure) can be represented as the conditional probability of finding local states,

$$f_2 = f\left[r \mid l, l'\right] = \frac{1}{S}\sum_s m[s, l]m\left[s + r, l'\right], \tag{10}$$

where S is the total number of pixels, l and l' are the local states at two ends of a distance and orientation defined by vector r [49]. The implementation of 2-point statistics by [49] in the Python library PyMKS (Materials Knowledge Systems in Python), an open-source framework designed for materials data science, is utilized for our purpose. The final dimension of the calculated 2-point spatial descriptor for each image is (256 × 256 × 2).

By examining triplets (3 points), one can capture more complex spatial patterns. This third order correlations are more sensitive to the exact geometrical arrangements in the material and can reveal intricate details about the grain boundary structure. The 3-point spatial correlation function extends the concept of the 2-point function by considering triplets of points. This function provides a measure of how the probability of finding a specific feature at one point is related to finding the same or another feature at two other points, defined by vectors r_1 and r_2,

$$f_3 = f\left[r_1, r_2 \mid l, l', l''\right] = \frac{1}{S}\sum_s m[s, l]m\left[s + r_1, l'\right]m\left[s + r_2, l''\right], \tag{11}$$

where S is the total number of pixels, and l, l', and l'' are the local states at the distances and orientations defined by r_1 and r_2. This function captures more complex spatial relationships and provides detailed information about the geometrical arrangements and interactions within the microstructure. For the purpose of this study, the implementation of 2-point statistics from PyMKS is extended to 3-point statistics, as represented by Eq. (11). The final dimension of the calculated 3-point spatial descriptor for each image is (256 × 256 × 3).

4.2. Gram matrices representations

Gram matrices are a more recent, compact and powerful statistical descriptor used for microstructure images. They provide a compact representation of spatial relationships among points of grain boundary features, making them particularly useful for analyzing grain boundary structures and texture. [50] developed a robust algorithm for the statistical description of input image, where internal activations in

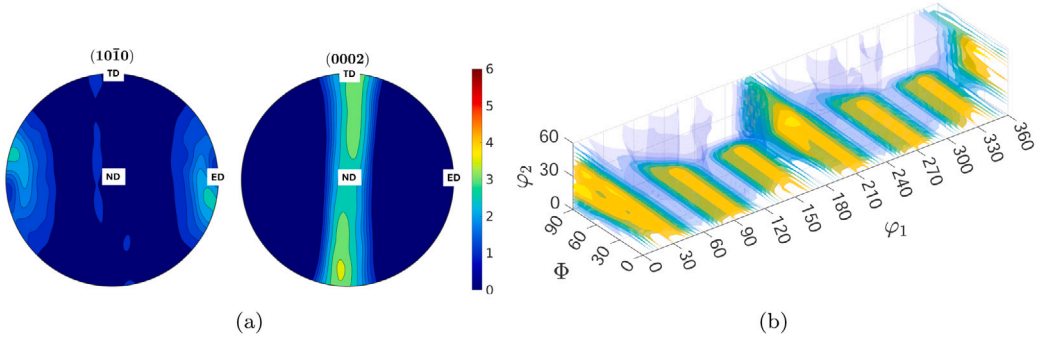

Fig. 6. Material texture characterization for Z1 ($\mathbf{v_{ext}} = 2$ mm/s, $\mathbf{T_{ext}} = 300$ °C). (a) Pole figure example from the dataset for texture on 1010 (left) and 0002 (right) planes with extrusion direction (ED), normal direction (ND) and transverse direction (TD) annotated. (b) 3D representation of Orientation distribution function (ODF) for the corresponding pole figure.

convolutional layers of a pre-trained CNN is denoted as F_{ij}^{l}, where l is the layer index, i is the feature index, and j is the pixel index. At each layer, the gram matrices can be captured between the feature index i and feature index k:

$$G_{ik}^{l} = \sum_{j} F_{ij}^{l} F_{kj}^{l}, \tag{12}$$

where the summation represents the transpose of the two feature vectors, which encodes invariance to translations in the image, up to its boundaries [29]. Compared to just the feature vector (activation filters from CNN), gram matrices provide a much richer statistical description. In this work, the implementation of gram matrices by [51] is utilized for the binarized microstructure. This approach uses a pre-trained, normalized version of a 16-layer VGG network, based on the ResNet architecture. Activation filters from layers l = 2, 4, 8, 12 and 16 are extracted, and gram matrices are calculated for each layer. To form the final texture vector, all the gram matrices are normalized and concatenated as: $\hat{G}^{l} = \left(w_l / A_l\right) G^{l}$, to get a scaled texture vector, $\hat{G} = \left(\hat{G}^{2}, \hat{G}^{4}, \ldots\right)$. For a binary image of size (256 × 256), the calculated gram matrices dimensions are $(64, 64)$, $(128, 128)$, $(512, 512)$ and $(512, 512)$.

5. Texture characterization

The texture data extracted from the experiments presented in Table 1 is in the form of pole figures generated by XRD experiments, as shown in Fig. 6(a). The material data comprises extruded flat bands and extruded cylindrical bars. For flat bands, the preparation and measurement of the specimen are done on the normal plane, whereas for round bars, the specimen is prepared and measured in XRD on the cross-sectional plane. This creates a difference in the texture description in the data. Therefore, rotations in the Bunge Euler angles are performed using an open-source MATLAB toolbox for crystallographic texture analysis, MTEX [52]. The rotation of the pole figure is done using the built-in functions within MTEX to maintain a uniform description of texture in the pole figure, as shown in Fig. 6(a). Sample symmetry is explained in this figure, the extrusion direction (ED) is oriented to the right (x-axis), the normal or second transverse direction for round bars is coming out of the plane (z-axis), and the other transverse direction is upwards (y-axis). Further, the uniform description of pole figures is used to calculate the ODF (see Fig. 6(b)), which describes the probability density of finding grains with crystal orientations represented by Bunge Euler angles: $\left\{g = \left(\varphi_1, \boldsymbol{\Phi}, \varphi_2\right) \mid 0 \le \varphi_1 < 2\pi,\ 0 \le \boldsymbol{\Phi} \le \pi/2,\ 0 \le \varphi_2 < \pi/6\right\}$ for Mg-alloys.

Because of the discrete description of the ODF in 5-degree increments, the ODF becomes high-dimensional data, making it challenging to use directly for Structure–Property (S–P) modeling. Therefore, the ODF can be analytically reduced into the expansion coefficients of the Generalized Spherical Harmonics (GSH) series. The harmonic series can be further reduced by leveraging the hexagonally close-packed (HCP) crystal symmetry of the Mg-alloy and the sample symmetry (orthotropic). This reduced representation of ODF, utilized in our study and also described by [33],

$$f(g) = \sum_{l=0}^{\infty} \sum_{m=1}^{M(l)} \sum_{n=1}^{N(l)} F_{l}^{mn} \dddot{T}_{l}^{\cdot\, mn}(g), \tag{13}$$

where F_{l}^{mn} are the expansion coefficients and $\dddot{T}_{l}^{\cdot\, mn}(g)$ is the GSH basis function, reduced for orthotropic sample symmetry and HCP crystal symmetry. The combination of numbers l, m and n defines the number of dimensions that the series will be expanded for. Each combination corresponds to one expansion coefficient F_{l}^{mn}, and the symmetry constraints up to the chosen number for the expansion determine the dimensionality of the space [33]. For our purpose, and most purposes, l is chosen to be 16. This value is the default for orientation imaging microscopy analysis software, TexSEM Laboratories (TSL OIM).

Expansion coefficient pertaining to each orientation $\left(\varphi_1, \boldsymbol{\Phi}, \varphi_2\right)$ in the ODF denoted by g_k, can be represented as

$${}^{k}F_{l}^{mn} = (2l+1)\dddot{T}^{\cdot} * {}_{l}^{mn}\left(g_k\right), \tag{14}$$

where $\dddot{T}^{\cdot} * {}_{l}^{mn}$ is the complex conjugated reduced basis function. To calculate the expansion coefficients for the complete ODF, represented by a total of $N_{\mathbf{crys}}$ orientations, the volume average of each individual expansion coefficient is computed,

$$\bar{F}_{l}^{mn} = \sum_{k=1}^{N_{\mathbf{crys}}} {}^{k}\alpha\, {}^{k}F_{l}^{mn},\ \sum_{k=1}^{N_{\mathbf{crys}}} {}^{k}\alpha = 1, 0 < {}^{k}\alpha < 1, \tag{15}$$

where ${}^{k}\alpha$ represents the weighting factor for each orientation k [53]. Implementing the above set of equations for the ODF in our dataset yields 551 coefficients, each with a real and imaginary value. These coefficients are vertically stacked to form a 1D vector of dimension 1102.

6. Dimensionality reduction techniques

In this section, we will discuss the machine learning-based techniques used for reducing the dimensionality of the microstructure descriptors (discussed in Section 4) and texture descriptors (discussed in Section 5). The techniques that will be discussed include Principal Component Analysis (PCA), Isomap manifold learning, and Autoencoder-based vector embeddings. The discussion will also encompass which technique is used for which descriptor and the suitability of each technique for the respective descriptors.

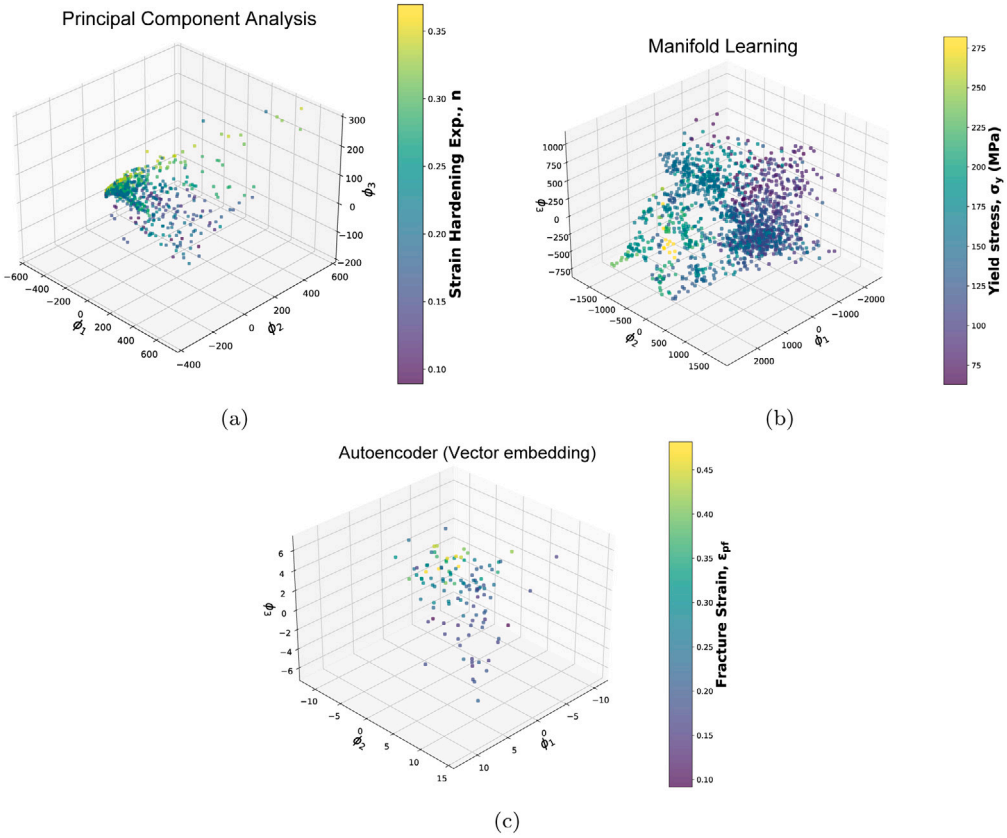

Fig. 7. First three components (ϕ_1, ϕ_2 and ϕ_3) for microstructure and texture descriptors with reduced dimensions using techniques like (a) Principal Component Analysis (PCA) for 3-point spatial correlations (scale = $100 \times 100\,\mu m$), colored for strain hardening exponent (**n**). (b) Isomap manifold learning for gram matrices (at scale = $100 \times 100\,\mu m$), colored for tensile yield stress (σ_y). (c) Autoencoder based vector embeddings for orientation distribution function (ODF), colored for fracture strain (ε_{pf}).

6.1. Principal component analysis (PCA)

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique. It centers the feature matrix of the descriptors around the mean by subtracting the average of all data points and then transforms it into a set of orthogonal components, capturing the significant variance in the data. The PCA representation for any descriptor replaces as

$$f_r = \sum_{k=1}^{R} \alpha_k \phi_{kr} + \bar{f}_r, \tag{16}$$

where R is the total number of principal components, $\bar{f}_r$ is the ensemble mean for the descriptor, ϕ_{kr} are the orthogonal basis, and α_k are the weights computed by PCA for each corresponding component [54].

Although PCA can handle some forms of non-linearity in the data, it is most effective when applied to data with linearity due to the linear nature of PCA. For this reason, PCA is commonly chosen to reduce the dimensions of 2-point and 3-point spatial correlation functions used for statistically describing the binarized microstructure. This is because of the inherent linearity in the description of these correlation functions, as shown in Eqs. (10) and (11) [17]. Scikit-learn, a widely used open-source machine learning library by [55], is utilized for PCA to reduce the dimensions for $f\left[r \mid l, l'\right]$ and $f\left[r_1, r_2 \mid l, l', l''\right]$ to 100 and 120 respectively. Fig. 7(a) shows the 3D representation of the first three principal components (ϕ_1, ϕ_2 and ϕ_3) for $f\left[r_1, r_2 \mid l, l', l''\right]$ of binary microstructure images from the dataset with scale = $100 \times 100\,\mu m$. Each data point in the figure is colored according to the value of the plastic property, strain hardening exponent (**n**). The color gradient in the figure represents variations in property values captured by the reduced dimensions. The figure reveals clusters of points with similar property values and a discernible distribution trend. However, it provides an incomplete representation, as it is based on a single microstructure descriptor and only the first three principal components of that descriptor. Consequently, it is beneficial to incorporate multiple microstructure descriptors, including texture descriptors, and utilize additional principal components for the reduced descriptors as input features to enhance property prediction accuracy

6.2. Manifold learning (Isomap)

Isomap, a type of manifold learning technique, is a non-linear dimensionality reduction method that extends the concept of Multidimensional Scaling (MDS) by incorporating geodesic distances between data points along the manifold of reduced dimensional space. For each data point in the feature matrix, its k-nearest neighbors are identified and represented as a graph where nodes are the data points and edges represent connections between the neighboring points. Shortest distances are further calculated between each pair of points in the graph, usually by algorithms like Dijkstra's, resulting in a distance matrix (D). Finally, classical MDS is used to embed these geodesic distances into lower-dimensional space by finding a configuration (Y) that minimizes the stress function:

$$\Omega(y) = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \left(D_{ij} - \left| y_i - y_j \right| \right)^2 \right)^{-\frac{1}{2}}, \tag{17}$$

where D_{ij} is the matrix of geodesic distances, $\left| y_i - y_j \right|$ are the Euclidean distances, and N is the total number of components [56].

Isomap is very effective for non-linear data that lies on a curved manifold, as it preserves the local structure much better than linear techniques like PCA. By preserving geodesic distances, Isomap maintains the global geometric relationships in the data, leading to a more

faithful representation of the intrinsic data structure in the lower-dimensional embedding. Multidimensional scaling has been previously applied to features based on gram matrices due to the non-linearity in the descriptors [57]. Further graph-based approaches to implement MDS, like Isomap, are also used to achieve low-dimensional embeddings that leverage the properties of gram matrices [58]. The implementation in Scikit-learn is again utilized to compute 200 low-dimensional embeddings of gram matrices-based texture vector ($\hat{G}^l$), for all the images in the dataset. Fig. 7(b) shows the first three dimensions (ϕ_1, ϕ_2, and ϕ_3) for the binary microstructure images from the dataset with a scale of $100 \times 100\,\mu m$. Each data point in the figure is colored according to the value of the tensile yield stress (σ_y). Here, the clustering of the colors in one portion of the 3D space and the gradient of the cluster hints at some relationship, although not fully complete, but at least indicative, between the mechanical property and the gram matrices-based texture vector. Isomap is also used for the ODF due to the non-linearity in the probability density function used to represent the orientation of grains using Bunge-Euler angles (see Fig. 6(b)). This approach helps to observe the effect of this dimensionality reduction on texture descriptors and their impact on the performance of property prediction.

6.3. Vector embedding (Autoencoder)

The ODF is defined in a multidimensional space with Euler angles as each dimension. The relationship between these orientations and the probability is inherently non-linear due to the periodic nature of angular measurements. For instance, the integration over angular parameters to obtain volume fractions involves non-linear transformations, which complicate the direct interpretation. Additionally, the reconstruction of the ODF from experimental pole figure data often involves approximations that introduce more non-linearity, which does not necessarily lie on a curved manifold, even in higher dimensions [59]. Understanding this non-linearity is essential for accurately predicting mechanical properties, especially anisotropic properties in polycrystalline metallic alloys. Therefore, we explore a contemporary dimensionality reduction technique that is particularly effective for processing volumetric data with complex non-linearity. This technique utilizes a 3D CNN Autoencoder to efficiently extract features from the ODF.

The 3D Convolutional Autoencoder comprises two main components: an encoder and a decoder. The encoder compresses the data into a lower-dimensional latent space by applying 3D convolutional layers, batch normalization layers, and max pooling layers. These layers capture the spatial hierarchies and feature patterns inherent in the three-dimensional input data, such as the ODF. Conversely, the decoder reconstructs the original input from the compressed representation using 3D transposed convolutional layers. This process involves 3D upsampling of the latent representation back to its original dimensions, aiming to minimize the reconstruction error. Table B.6 presents the Autoencoder architecture, implemented using an open-source deep learning library with a Python interface, Keras, and employed for the compact learning of the ODF. The first column in Table B.6 represents the operator and the activation layer applied after the operator. The input to the encoder has a shape of $72 \times 19 \times 12$, corresponding to the three Bunge-Euler angles and one channel, which is the probability density function. The autoencoder consists of five convolutional layers and a final fully connected linear layer, which encodes the data to learn a 100-dimensional latent feature representation of the ODF. A dropout ratio of 0.2 is used to prevent overfitting and encourage more generalized feature learning. The decoder architecture starts with a fully connected linear layer, followed by five 3D transposed convolutional layers with sigmoid activation functions. This setup allows for learning complex patterns in the data with smooth gradients, ultimately reconstructing the ODF. The mean squared error between the input and the output of the Autoencoder is minimized using the Adam optimizer, with an exponential decay learning rate scheduler starting at 0.0001. The model is trained for a maximum of 1000 epochs with a batch size of 8. An early stopping criterion is applied to halt training if the metrics do not improve for 5 consecutive epochs. Fig. 7(c) illustrates the first three dimensions of the 100-dimensional features learned by the autoencoder for the 119 ODFs in our dataset. Each data point in the figure is colored according to the value of the fracture strain ($\varepsilon_{\mathbf{pf}}$). The color gradient indicates a relationship/trend between the lower-order representation of texture descriptors and mechanical property, $\varepsilon_{\mathbf{pf}}$. (See Table 2)

7. Lower-order modeling of structure–property linkages

The reduced-order microstructure descriptors, including 2-point correlations, 3-point correlations, gram matrices, binned histogram representations of grain size distribution, and aspect ratio distribution, as well as reduced-order texture descriptors, are all shaped into a 1D vector. This serves as the input to the non-linear regression models discussed in this section for predicting mechanical properties. The descriptors will be investigated in various combinations to evaluate their performance in property prediction. For this purpose, non-linear, non-parametric regressors are chosen. Non-linear regressors are selected due to the complex patterns in the data that need to be linked to the corresponding mechanical properties. Non-parametric regressors are preferred because there are no predefined functional forms specified for the relationships between the structure and properties. The shape and complexity of the model should adapt to these linkages, making non-parametric regression models more suitable for this problem. Gaussian Process (GP) regression is a powerful non-parametric method used for modeling complex, non-linear relationships without assuming a specific functional form. The performance of the GP regressor is compared to that of a random forest and a Multi-Layer Perceptron (MLP) regressor. XGBoost, a non-parametric regressor, which is based on random forest and gradient boosting, is also evaluated. Although the MLP regressor is parametric due to the predetermined number of layers and neurons per layer, it remains a good choice for this application. This is because these numbers are considered hyper parameters and can be fine-tuned to accurately model the S–P relationship. For predicting each mechanical property a different regressor is trained with same inputs but different mechanical property as label.

7.1. Random forest regression (XGBoost)

XGBoost, or eXtreme Gradient Boosting, is a powerful machine learning algorithm that enhances the gradient boosting framework through piecewise constant approximations in decision trees. It is most effective when interactions between features and threshold values are important. The core of XGBoost's optimization lies in its objective function,

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \Omega(f), \tag{18}$$

where $l(y_i, \hat{y}_i)$ represents the loss function, which measures how well the predicted values $\hat{y}_i$ match the actual values y_i. XGBoost typically uses a squared loss function for regression tasks,

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \tag{19}$$

which quantifies the difference between actual and predicted values. $\Omega(f)$ is a regularization term that penalizes model complexity, helping to prevent overfitting by controlling the size of the trees in the ensemble. The regularization term can be defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{20}$$

where T is the number of leaves in the tree, w_j are the weights of each leaf, γ controls the minimum loss reduction required to make a

Table 2
3D Convolutional Autoencoder architecture for encoding Orientation Distribution Function (ODF).

Encoder Architecture				
Layer (Activation Function)	Kernel Size	Channels	Output Shape	# of Param
Input Layer	–	1	(72, 19, 12, 1)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	16	(72, 19, 12, 16)	448
Batch Normalization	–	–	(72, 19, 12, 16)	64
Convolutional Layer (LeakyReLU)	3 × 3 × 3	32	(72, 19, 12, 32)	13,856
Batch Normalization	–	–	(72, 19, 12, 32)	128
Max-Pooling Layer	2 × 2 × 2	–	(36, 9, 6, 32)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	64	(36, 9, 6, 64)	55,360
Batch Normalization	–	–	(36, 9, 6, 64)	256
Convolutional Layer (LeakyReLU)	3 × 3 × 3	64	(36, 9, 6, 64)	110,656
Batch Normalization	–	–	(36, 9, 6, 64)	256
Max-Pooling Layer	2 × 2 × 2	–	(18, 4, 3, 64)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	128	(18, 4, 3, 128)	221,312
Batch Normalization	–	–	(18, 4, 3, 128)	512
Max-Pooling Layer	2 × 2 × 2	–	(9, 2, 1, 128)	0
Flatten Layer	–	–	(2304)	0
Dropout Layer (0.2)	–	–	(2304)	0
Dense Layer (Linear)	–	100	(100)	230,500
Decoder Architecture				
Layer (Activation Function)	Kernel Size	Channels	Output Shape	# of Param
Dense Layer (ReLU)	–	128	(4608)	465,408
Reshape Layer	–	–	(9, 4, 6, 128)	0
Upsampling Layer	2 × 2 × 2	–	(18, 8, 12, 128)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	128	(18, 8, 12, 128)	442,496
Upsampling Layer	2 × 3 × 1	–	(36, 24, 12, 128)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	64	(36, 24, 12, 64)	221,248
Convolutional Layer (LeakyReLU)	3 × 3 × 3	64	(36, 24, 12, 64)	110,656
Upsampling Layer	2 × 1 × 1	–	(72, 24, 12, 64)	0
Convolutional Layer (LeakyReLU)	3 × 3 × 3	32	(72, 24, 12, 32)	55,360
Convolutional Layer (LeakyReLU)	1 × 3 × 1	16	(72, 22, 12, 16)	1,552
Convolutional Layer (Sigmoid)	1 × 3 × 1	1	(72, 20, 12, 1)	49
Cropping Layer	–	–	(72, 19, 12, 1)	0

further partition on a leaf node and λ is a regularization parameter that penalizes large weights. XGBoost employs a gradient descent approach to minimize the objective function. At each iteration, it fits a new model (often a decision tree) to the negative gradient of the loss function [60]. Hyperparameter selection is done using a 5-fold cross validation strategy over the training set. An early stopping criterion is applied for regularization, if the evaluation metric (mean squared error) does not improve for 10 consecutive iterations.

7.2. Multi layer perceptron (MLP)

An MLP regressor is a type of fully connected artificial neural network used for predicting continuous values. It consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. Each neuron in the hidden layers computes a weighted sum of its inputs and applies an activation function $h_i^{(\ell)}$, for the ℓ-th layer, the output of the i-th neuron

$$h_i^{(\ell)} = \Psi^{(\ell)}\left(\sum_j w_{ij}^{(\ell)} h_j^{(\ell-1)} + b_i^{(\ell)}\right), \tag{21}$$

where $w_{ij}^{(\ell)}$ is the weight connecting neuron j from layer $\ell-1$ to neuron i in layer ℓ, $b_i^{(\ell)}$ is the bias term for neuron i and $\Psi^{(\ell)}$ is the activation function for layer ℓ. The activation function chosen for hidden layer of MLP regressor is Rectified Linear Unit (ReLU) which introduces non-linearity while being computationally efficient.

The final output of the MLP, which represents the predicted value,

$$y = h^{(L)} = \Psi^{(L)}\left(\sum_j w_j^{(L)} h_j^{(L-1)} + b^{(L)}\right), \tag{22}$$

where L denotes the output layer, which uses linear activation function for alloying continuous output values (important for regression tasks) for real numbers [61]. MLP regressor is typically trained using back propagation along with Adam optimizer with the goal to minimize mean squared error loss function. The number of hidden layers, the number of neurons per layer, the learning rate, and the regularization parameter (alpha) are considered for hyperparameter optimization to enhance prediction accuracy. The training set is further divided into five folds, and a k-fold cross-validation strategy is employed for hyperparameter optimization.

7.3. Gaussian process regression (GP)

GP Regression utilizes the properties of Gaussian processes to make predictions about unknown functions. In this context, a Gaussian process can be understood as a collection of random variables, any finite number of which have a joint Gaussian distribution. This characteristic allows GP regression to provide a flexible framework for modeling the underlying relation between co-variates and response. GP not only predicts the mean of the output variable but also provides a measure of uncertainty associated with these predictions [62]. This is achieved through the posterior distribution derived from the prior distribution and observed data.

The GaussianProcessRegressor class from Scikit-learn is utilized because of its efficient implementation with support for various kernel functions and automatic hyperparameter optimization. In GP regression, the underlying function $f(\mathbf{x})$ is modeled as a realization of a Gaussian process,

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{23}$$

where $m(\mathbf{x})$ is the mean function (often assumed to be zero), and $k(\mathbf{x}, \mathbf{x}')$ is the covariance kernel function that encodes assumptions about the smoothness or periodicity of $f(\mathbf{x})$. Radial Basis Function (RBF) kernel,

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right), \tag{24}$$

where l is the length scale parameter, is chosen as $k(\mathbf{x}, \mathbf{x}')$. The observed data y_i at $\mathbf{x}_i$ are assumed to be noisy realizations,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2). \tag{25}$$

Given n training samples $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, the predictive distribution at a new input $\mathbf{x}_*$ is Gaussian with mean $\mu(\mathbf{x}_*)$ and variance $\sigma^2(\mathbf{x}_*)$,

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y}, \tag{26}$$

$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_*, \tag{27}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the covariance matrix with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_* \in \mathbb{R}^n$ is the covariance vector between $\mathbf{x}_*$ and the training points, σ_n^2 is the noise variance, and $\mathbf{y} \in \mathbb{R}^n$ is the vector of observed targets. Further, GaussianProcessRegressor optimizes the hyperparameter to train the model on the data by maximizing log-marginal likelihood,

$$\log\left(p(\mathbf{y} \mid \mathbf{X}, \theta)\right) = -\frac{1}{2}\mathbf{y}^\top \left(\mathbf{K}_\theta + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} - \frac{1}{2}\log\left(\det\left(\mathbf{K}_\theta + \sigma_n^2 \mathbf{I}\right)\right) - \frac{n}{2}\log 2\pi. \tag{28}$$

Finally, predictions at test points are made by computing the joint distribution of observed outputs and outputs at new inputs. From this, the conditional distribution of $f(X_*)$ is computed, which ultimately yields mean predictions and uncertainty quantification [63]. Since GP regression models are numerically similar to kernel regression, they benefit from an additional well-defined statistical framework. This framework enables the optimization of the likelihood function for estimating hyperparameters without the need for cross-validation.

7.4. Shapley additive explanation value (SHAP value)

The non-linear regression models discussed above are often regarded as "black boxes" due to their ability to take in input arguments and return target output values without revealing the internal decision-making process. The complexity of these models increases with the intricacy of input features and their interdependencies during output value prediction, which in turn hampers their interpretability. Interpretability refers to the ability of a human to comprehend the impact of each input feature on the model's output. This becomes particularly challenging in this research due to the diverse range of input features used to predict mechanical properties. To address such complexities, the quantification of feature dependence or importance is crucial. In the field of machine learning, feature additive contribution models are employed to enhance the explainability of these complex models [64]. SHAP (SHapley Additive exPlanations) leverages the concept of game theory known as Shapley values, which provide a method for fairly distributing gains and losses among multiple contributors in a collaboration. In SHAP value analysis, the estimation of Shapley values is performed using a weighted linear regression with a kernel estimator [65]. It is utilized in this research to understand the feature importance of microstructure statistics, microstructure, and texture descriptors in predicting mechanical properties.

The goal in SHAP value analysis for machine learning models is to fairly attribute the difference between the model's prediction for a given data point and the mean prediction to each feature. The value of feature (i) is calculated as,

$$\text{SHAP}(x_i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \Big[f(S \cup \{i\}) - f(S) \Big], \tag{29}$$

where, N is the set of all features, S is a subset of features excluding i, $S \subseteq N \setminus \{i\}$. $f(S)$ is the model's prediction when only the features in subset S are present. $|S|$ is the cardinality of subset S and $|N|$ is the total number of features.

The marginal contribution of each feature i is represented by the term $f(S \cup \{i\}) - f(S)$ ie. difference in the model's output when the feature is included in the subset S. The term before this in Eq. (29), ensures that the contributions are averaged across all possible subsets S and summation at the end is done to attribute equal considerations to all feature combinations. Computation of SHAP values involves considering all 2^n subsets of features, making it computationally expensive for large n. Therefore, approximations such as Kernel and Tree explainer are used according to the nature of the regression model. For the purpose of this study, Tree explainer from the implementation done in the SHAP library by [66], is used to understand the feature importance for GP, MLP and XGBoost regressors.

8. Results and discussions

In this section, we compare the prediction performance of the models discussed in Section 7 using the input features outlined in Sections Section 4, 5, and 6. This comparison is conducted using parity plots, which compare true and predicted values, as well as metrics such as mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). Among the mechanical properties discussed in Section 3.3, we focus the discussion on the prediction results for the strain hardening exponent, $\mathbf{n}$, and yield stress, $\boldsymbol{\sigma}_\mathbf{y}$. The best performance for prediction of ultimate tensile stress, $\boldsymbol{\sigma}_\mathbf{max}$, uniform strain, $\boldsymbol{\varepsilon}_\mathbf{pu}$, and fracture strain $\boldsymbol{\varepsilon}_\mathbf{pf}$ are further shown in Appendix A. In this section, we will discuss four different investigations conducted in this study to identify the best configuration for the modules of the pipeline mentioned in Fig. 1 for accurate mechanical property prediction. Table 3 lists the input feature vectors and sets of these input feature vectors that will be used for these investigations. It also includes the symbol assigned to them, dimensionality reduction technique applied to them, their final reduced size, and whether they describe microstructure, microstructure statistics, or texture.

- In the first investigation, the suitability of the three regression models—GP, XGBoost, and MLP—to the dataset is evaluated. This is crucial to determine which model performs better than the others for the same input set, $\boldsymbol{\Omega}$.
- In the second investigation, the three reduced microstructure descriptors (${}^{\mathbf{PCA}}f_2$, ${}^{\mathbf{PCA}}f_3$, and ${}^{\mathbf{ISO}}\hat{G}$) and their combinations are examined for their performance in predicting $\mathbf{n}$ and $\boldsymbol{\sigma}_\mathbf{y}$ using the input set $\boldsymbol{\Psi}$.
- In the third investigation, three different reduced-order texture descriptors (${}^{\mathbf{ISO}}\bar{F}_l^{mn}$, ${}^{\mathbf{VEC}}g(\varphi_1, \boldsymbol{\Phi}, \varphi_2)$, and ${}^{\mathbf{ISO}}g_{(\varphi_1,\boldsymbol{\Phi},\varphi_2)}$) are assessed for their performance in property prediction using the input set $\boldsymbol{\Phi}$.
- Finally, the fourth investigation is conducted using the input set $\boldsymbol{\Omega}$ to understand the impact of length scale (90, 100, 120 and 150 μm) in reduced-order microstructure descriptors on property prediction accuracy.

These investigations are also further employed to understand the dependence of the target mechanical properties on the input microstructure and texture descriptors. This interpretation is further quantified using SHAP values, as presented in Section 8.5.

8.1. Investigating non-linear regressors

The best-performing configurations of hyperparameters for the GP, XGBoost, and MLP regressors are used to predict $\mathbf{n}$ and $\boldsymbol{\sigma}_\mathbf{y}$ for the input set $\boldsymbol{\Omega}$ at a scale of 100 μm. The fit of the model to the training and testing data for predicting $\mathbf{n}$ is illustrated using parity plots (true vs. predicted values) in Fig. 9. It is observed that in all the parity plots, there are groups of points with the same true value but different predicted values. This occurs because tiling and multiple images taken from the same specimen might have different microstructure descriptors and microstructure statistics, yet share the same mechanical performance label. The grouping and alignment of these points also indicate that the models are able to learn some patterns in linking slightly different microstructure descriptors to the same output property label.

A similar comparison is observed between the models for predicting $\boldsymbol{\sigma}_\mathbf{y}$ for the input set $\boldsymbol{\Omega}$ at a scale of 100, μm, as shown in Fig. 10. The MLP regressor exhibits higher train MAPE values of 3.98%, with train

Table 3
Input feature vectors and the sets of their various combinations which are used for investigating the performance of non-linear regression models in predicting mechanical properties, $\mathbf{\Omega}$ is used for regressor performance and length scale, $\mathbf{\Psi}$ is used for microstructure descriptors and $\mathbf{\Phi}$ is used for texture descriptors.

	Input feature vectors	Dimensionality reduction	Size	Symbol	Input sets for investigation		
					$\mathbf{\Omega}$	$\mathbf{\Psi}$	$\mathbf{\Phi}$
Microstructure	2-point spatial correlation function	PCA	100	${}^{\mathbf{PCA}}f_2$			
	3-point spatial correlation function	PCA	120	${}^{\mathbf{PCA}}f_3$	✓		✓
	Gram Matrices based texture vector	Isomap	200	${}^{\mathbf{ISO}}\hat{G}$	✓		✓
Texture	GSH coefficient vector	Isomap	60	${}^{\mathbf{ISO}}\hat{F}_l^{mn}$	✓	✓	
	Orientation Distribution Function (ODF)	Isomap	60	${}^{\mathbf{ISO}}g_{(\varphi_1,\Phi,\varphi_2)}$			
	Vector embeddings of ODF	Autoencoder	100	${}^{\mathbf{VEC}}g_{(\varphi_1,\Phi,\varphi_2)}$			
Stats.	Aspect Ratio distribution	None	30	$\mathbf{AR}$	✓	✓	✓
	Equivalent diameter distribution	None	30	$d_{\mathbf{eq}}$	✓	✓	✓

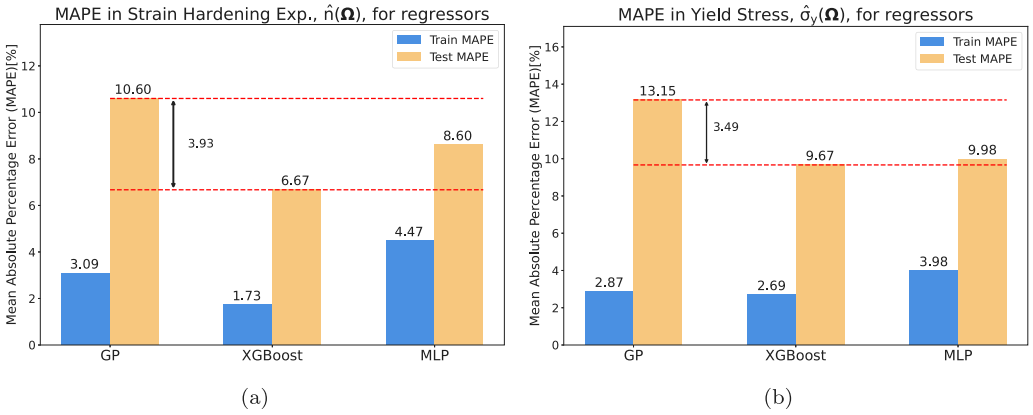

Fig. 8. Overall performance comparison of Gaussian Process, Multi Layer Perceptron and XGBoost regressors using mean absolute percentage error (MAPE) bar graph for predicting (a) Strain Hardening Exponent, $\mathbf{n}$ and (b) Yield Stress, $\boldsymbol{\sigma}_\mathbf{y}$, using the input set $\mathbf{\Omega}$, at scale $= 100\,\mu$m.

data points scattered further from the perfect agreement line in Fig. 10(c), indicating underfitting and an inability to effectively learn the patterns in the data. This is further supported by high variance in the test set and a low R^2 value of 0.42. The GP regressor again shows overfitting compared to the XGBoost regressor, as evidenced by the difference in its train and test MAPE. Consequently, the model does not generalize well to the unseen test data, yielding a test MAPE of 13.15%, which is the highest among the three regressors (see Fig. 8(b)). XGBoost once again performs best in terms of predicting the unseen data, with a test MAPE of 9.67%. However, a few data points from a single material specimen perform considerably worse than others due to under representation in the train set (see Fig. 10(b)). Overall, the performance of all the regressors is not as good for predicting $\boldsymbol{\sigma}_\mathbf{y}$ compared to $\mathbf{n}$ for the input set $\mathbf{\Omega}$ at a scale of $100\,\mu$m. Additionally, Appendix Appendix B summarizes all the performance metrics for different configurations of input features/sets across all three regressors.

All GP, XGBoost, and MLP models fit well on the training data, with MAE of 0.007 and MAPE of 3.09% for GP, train MAE of 0.004 and MAPE of 1.73% for XGBoost, and train MAE of 0.01 and MAPE of 4.47% for MLP. Nevertheless, the R^2 value of 0.95 and slightly worse training metrics indicate underfitting of the training data for the best-performing MLP model. This is also evident by examining the training data (blue) in Fig. 9(c), which is not the case with GP and XGBoost. Although the R^2 value of 0.97 is better for GP, it does not necessarily mean that it learns the inherent patterns of the material data better than MLP. For predicting the test data, GP performs worse than MLP, with MAE and MAPE of 0.026 and 10.60% for GP, compared to MAE and MAPE of 0.018 and 8.6% for MLP. The GP regressor comparatively overfits slightly more to the training data when compared to the XGBoost regressor, as indicated by the greater difference in Train and Test MAPE for GP.

Fig. 8(a) compares the test and train predictions of all the models and shows that XGBoost performs superior to the other two models on the unseen test data, with a MAPE of 6.67% in estimating the value of $\mathbf{n}$. This is also clear from Fig. 9(b), where the test data (orange points) are mostly very close to the perfect agreement line, apart from data from one or two material specimens, which reduced the test R^2 value to 0.72. This variance can be attributed to the fact that this material type from the test set does not have adequate representation in the train set, which can be improved by increasing the number of data points to enhance this representation.

8.2. Investigating microstructure descriptors

In Section 8.1, we found that XGBoost model learns and generalizes well over the data, as evidenced by its test lowest test and train MAPE in predicting both $\mathbf{n}$ and $\boldsymbol{\sigma}_\mathbf{y}$. Therefore, XGBoost models will be used for the next three investigations. Fig. 11(a) shows the comparison of the XGBoost regressor in predicting $\mathbf{n}$ using ${}^{\mathbf{PCA}}f_2$, ${}^{\mathbf{PCA}}f_3$, ${}^{\mathbf{ISO}}\hat{G}$, and the combination of $({}^{\mathbf{PCA}}f_3,\ {}^{\mathbf{ISO}}\hat{G})$ along with the input set $\mathbf{\Psi}$ at a scale of $100,\mu$m. The PCA-reduced 3-point spatial correlation function performs the worst with a test MAPE of 9.5%, while the isomap-reduced gram matrices, ${}^{\mathbf{ISO}}\hat{G}$, perform much better with a test MAPE of 6.97%. This indicates that the style and texture of grain boundaries are more indicative of $\mathbf{n}$, making gram matrices a better descriptor of the microstructure features when directly compared to 2-point or 3-point spatial correlations for our curated dataset of extruded Mg-alloys. Although spatial correlation functions contains information of a spatial distribution of pixels with grain boundaries, which is absent in gram matrices, it is observed that the combination of $({}^{\mathbf{PCA}}f_3,\ {}^{\mathbf{ISO}}\hat{G})$ performs slightly better than just ${}^{\mathbf{ISO}}\hat{G}$, with test MAPE of 6.67%. The combination of $({}^{\mathbf{PCA}}f_2,\ {}^{\mathbf{ISO}}\hat{G})$ was also tested but not included in the showcased results, as the improvement of the combination $({}^{\mathbf{PCA}}f_3,\ {}^{\mathbf{ISO}}\hat{G})$ was more significant.

There is no significant difference in the prediction performance of $\boldsymbol{\sigma}_\mathbf{y}$ using the microstructure descriptors—${}^{\mathbf{PCA}}f_2$, ${}^{\mathbf{PCA}}f_3$, ${}^{\mathbf{ISO}}\hat{G}$, and the

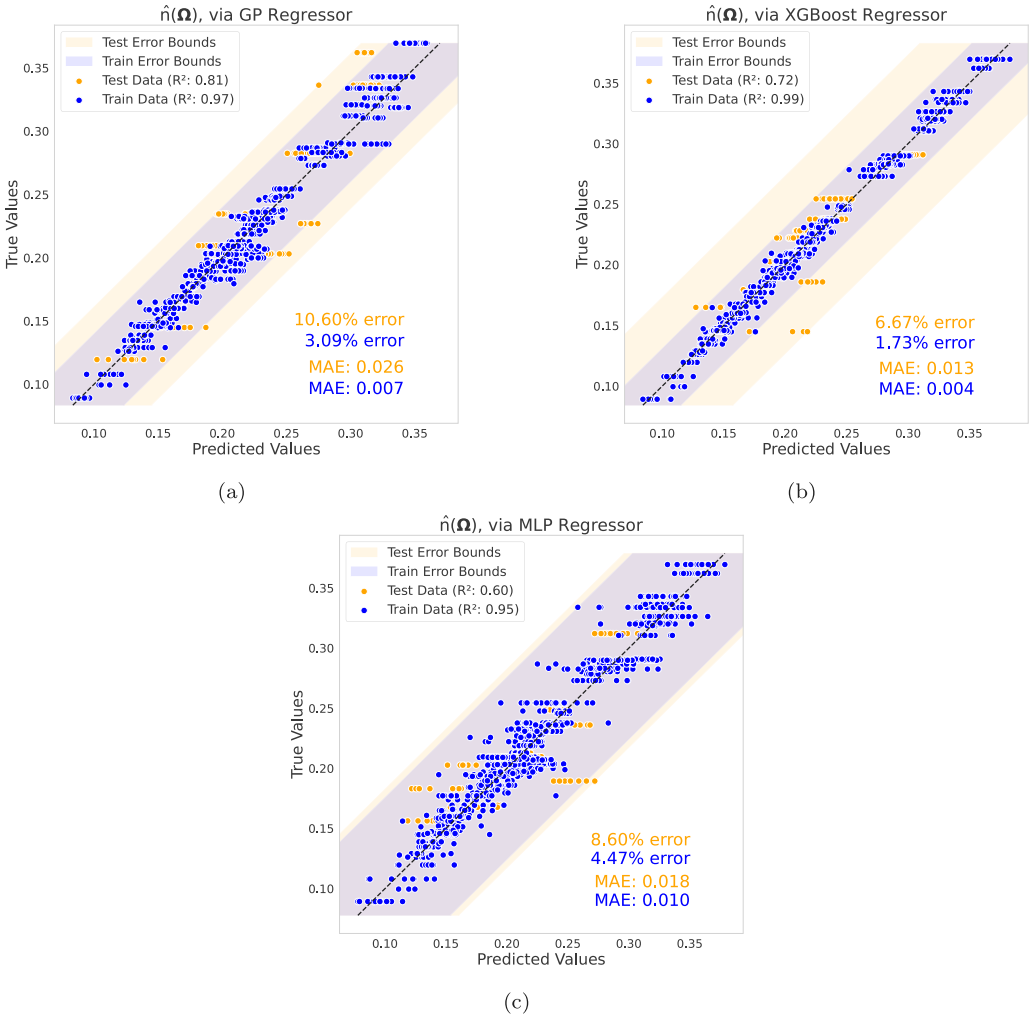

Fig. 9. True and predicted values of Strain Hardening Exponent, **n**, for (a) Gaussian Process, (b) XGBoost and (c) Multi Layer Perceptron regressors using input feature set **Ω**, at scale = 100 μm.

combination of (${}^{\mathbf{PCA}}f_3$, ${}^{\mathbf{ISO}}\hat{G}$)—along with the input set **Ψ** at a scale of 100, μm (see Fig. 11(b)). The difference in test MAPE between the best-performing ${}^{\mathbf{PCA}}f_2$ and the worst-performing ${}^{\mathbf{PCA}}f_3$, ${}^{\mathbf{ISO}}\hat{G}$ is just 0.4%. This indicates that contribution of microstructure descriptor in prediction of $\boldsymbol{\sigma}_\mathbf{y}$ is limited. Additionally, all microstructure descriptors perform comparatively worse (test MAPE greater than 9.2%), suggesting that the length scale of 100 μm may not be suitable and may not contain enough information to learn the patterns or indicate $\boldsymbol{\sigma}_\mathbf{y}$. The influence of length scales is further investigated in Section 8.4. The comparison of true vs. predicted values of **n** and $\boldsymbol{\sigma}_\mathbf{y}$ for the best-performing microstructure descriptors identified in this investigation is shown in Fig. 12. The figure again demonstrates that the model is able to accurately learn the patterns for all material specimens in the train and test sets, except for one or two materials in the test set due to their under representation in the train set. For checking all the parity plots for the performance comparison of microstructure descriptors discussed in Fig. 11(a), refer to Appendix Appendix C.

8.3. Investigating texture descriptors

An overall comparison of the performance of three different reduced texture descriptors in predicting **n** and $\boldsymbol{\sigma}_\mathbf{y}$ is shown in Fig. 13 using bar graph with MAPE values. For predicting **n** using the input set **Φ**, ${}^{\mathbf{ISO}}\bar{F}_l^{mn}$ performs the best with a test MAPE of 6.67%, which is 2.41% lower than the worst-performing ${}^{\mathbf{ISO}}g(\varphi_1, \boldsymbol{\Phi}, \varphi_2)$. Similarly, for predicting $\boldsymbol{\sigma}_\mathbf{y}$ using the input set **Φ**, ${}^{\mathbf{ISO}}\bar{F}_l^{mn}$ works better with a test MAPE of 9.67%, although the difference between this and the worst-performing ${}^{\mathbf{ISO}}g(\varphi_1, \boldsymbol{\Phi}, \varphi_2)$ is only 0.76%. This indicates that isomap dimensionality reduction on the GSH coefficient vector works better than directly implementing it on the ODF for preserving the patterns in material texture that are indicative of **n** and $\boldsymbol{\sigma}_\mathbf{y}$. This is due to the analytical dimensionality reduction inherent in the mathematical structure of GSH. Autoencoder-based vector embedding performed comparatively better than isomap for directly implementing dimensionality reduction on the ODF, with a test MAPE of 8.17% for **n** and 10.28% for $\boldsymbol{\sigma}_\mathbf{y}$. However, ${}^{\mathbf{ISO}}\bar{F}l^{mn}$ is overall a better choice for a reduced texture descriptor in the prediction of both **n** and $\boldsymbol{\sigma}_\mathbf{y}$. Fig. 14 shows the true vs. predicted values of **n** and $\boldsymbol{\sigma}_\mathbf{y}$ with the input set **Φ** and vector embeddings of ODF, ${}^{\mathbf{VEC}}g_{(\varphi_1,\boldsymbol{\Phi},\varphi_2)}$. There is a good agreement of the training data seen in Fig. 14(a) for **n**, but a comparatively worse fit to the training data in Fig. 14(b) for $\boldsymbol{\sigma}_\mathbf{y}$. This can be attributed to the insufficiency in the dataset required to effectively learn the links between all the descriptors and $\boldsymbol{\sigma}_\mathbf{y}$. However, overall, the XGBoost model is able to learn and effectively predict **n** and $\boldsymbol{\sigma}_\mathbf{y}$ with ${}^{\mathbf{VEC}}g_{(\varphi_1,\boldsymbol{\Phi},\varphi_2)}$ as the texture descriptor, within a certain amount of accuracy, except for slight underfitting in Fig. 14(b). Refer to Appendix Appendix C for all the parity plots comparing the performance for rest of the texture descriptors.

8.4. Investigating length scales

Fig. 15(b) compares the performance of the combination of microstructure descriptors, ${}^{\mathbf{PCA}}f_3$ and ${}^{\mathbf{ISO}}\hat{G}$, calculated at length scales of

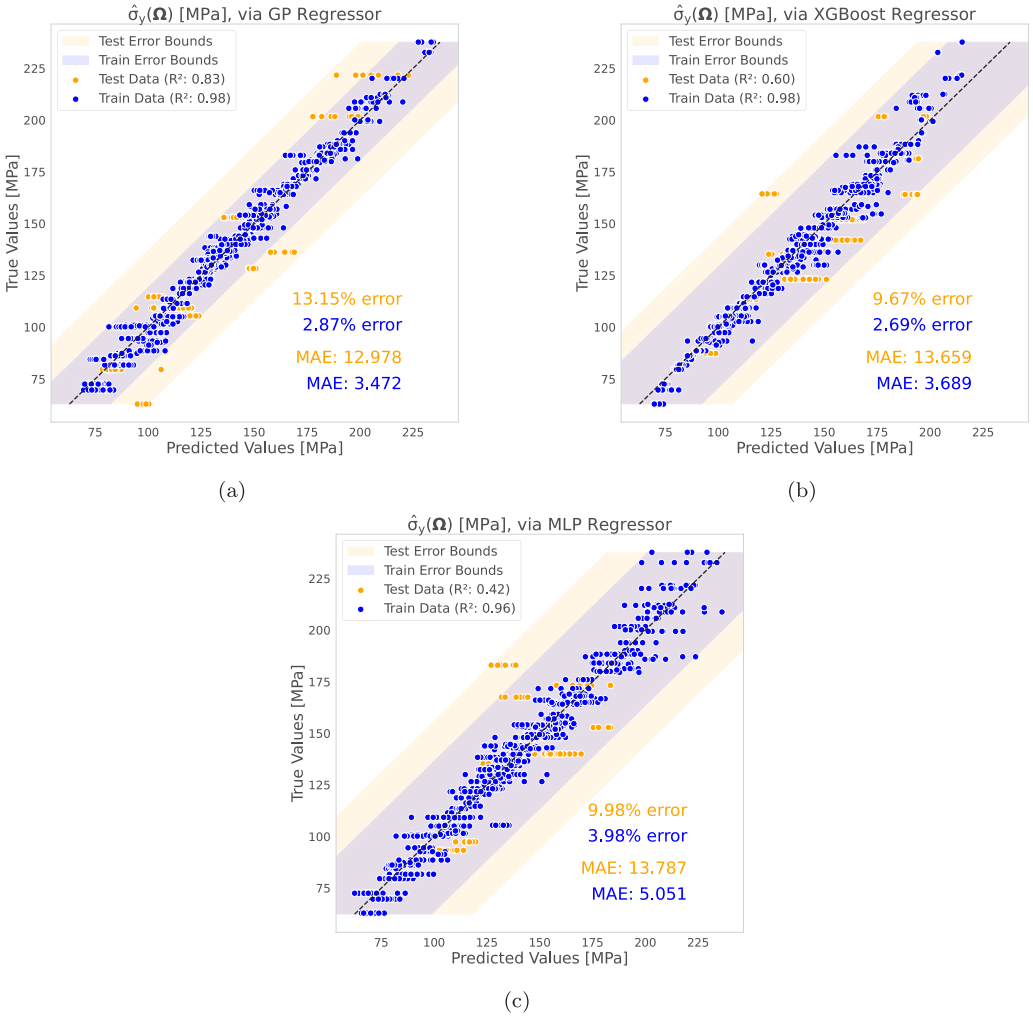

Fig. 10. True and predicted values of Yield Stress, σ_y, for (a) Gaussian Process, (b) XGBoost and (c) Multi Layer Perceptron regressors using input feature set $\mathbf{\Omega}$, at scale = 100 μm.

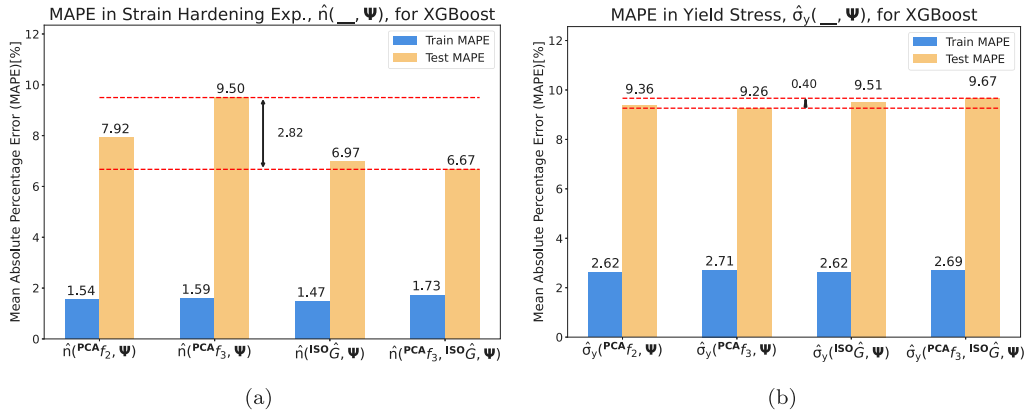

Fig. 11. Overall performance comparison of different microstructure descriptor based input features using mean absolute percentage error (MAPE) bar graph for predicting (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, using the input set $\mathbf{\Psi}$, at scale = 100 μm using XGBoost regressor.

90, 100, 120, and 150 μm to predict σ_y. Unlike previous investigations, this investigation shows that changing the length scale significantly impacts the test MAPE, with performance improving considerably at larger scales (120 and 150 μm). The test MAPE at 150 μm is 7.01%, which is 3.71% better than at the scale of 90 μm. At larger length scales, relatively more grains and more grain boundary features are included in the images which used to calculate the reduced-order microstructure descriptors. The increased accuracy in σ_y prediction can be attributed to this additional information captured in the descriptors, which was missing at smaller length scales. This indicates that more grain boundary structure information is required to learn patterns for predicting σ_y. For the prediction of **n**, microstructure descriptors at a scale of 90 μm perform the worst with a test MAPE of 8.59%, while those at a scale of 150 μm perform the best with a test MAPE of 5.35%

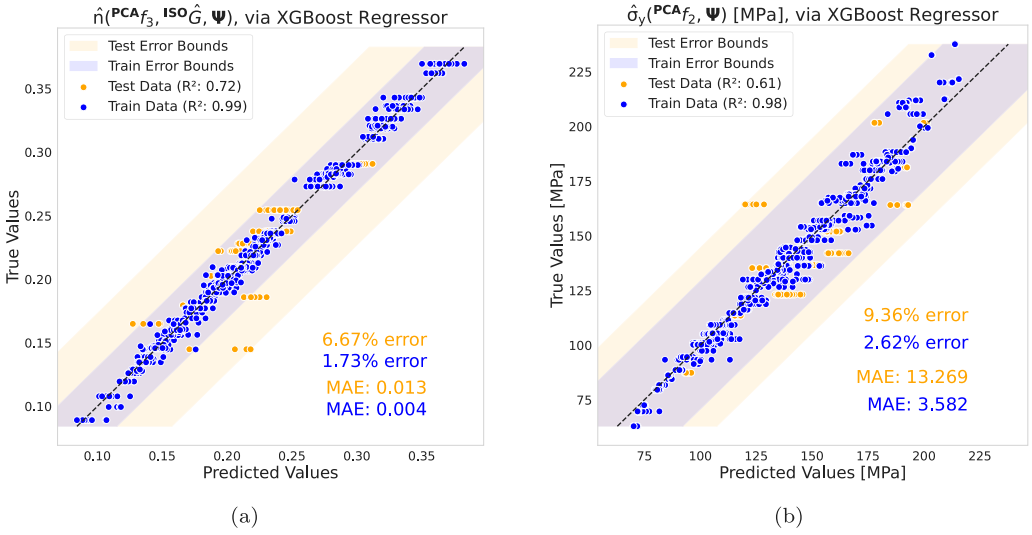

Fig. 12. True and predicted values of (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, for XGBoost regressor using input features $^{PCA}f_3$, $^{ISO}\hat{G}$ and input set Ψ, at scale = 100 μm.

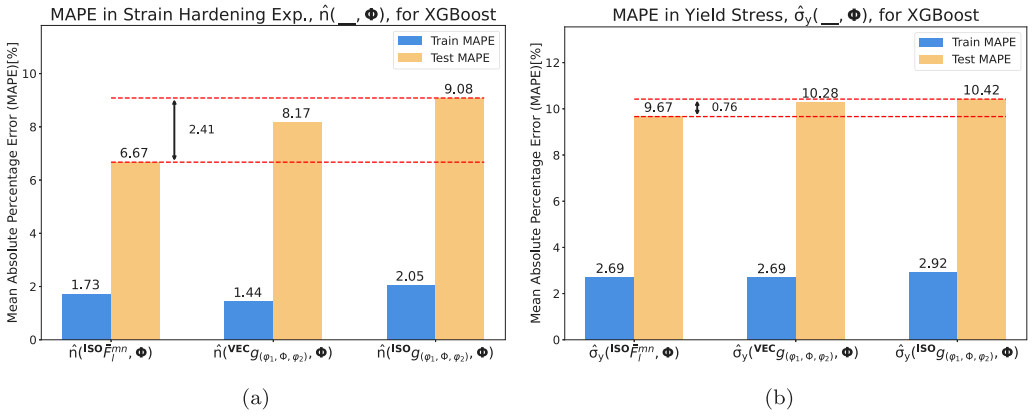

Fig. 13. Overall performance comparison of different texture descriptor based input features using mean absolute percentage error (MAPE) bar graph for predicting (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, using the input set Φ, at scale = 100 μm using XGBoost regressor.

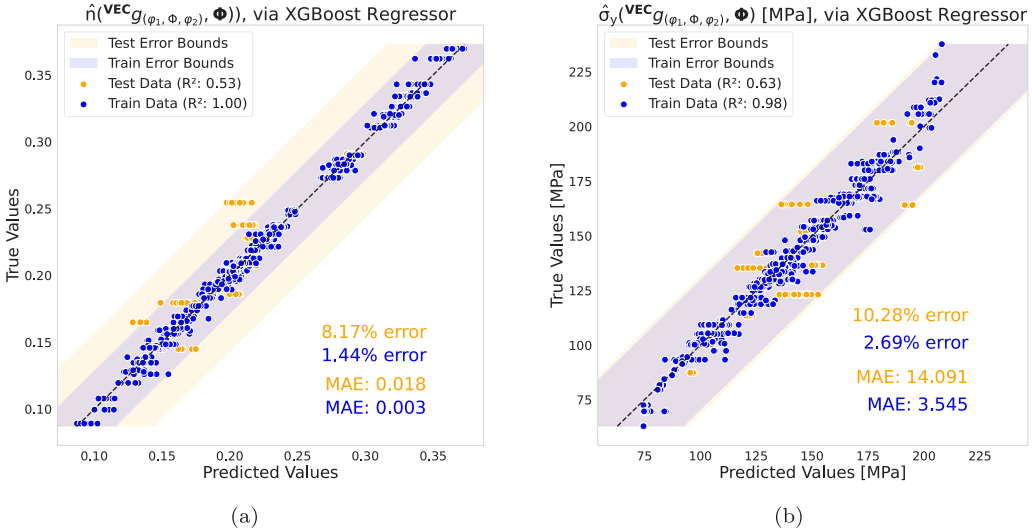

Fig. 14. True and predicted values of (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, for XGBoost regressor using input feature $^{VEC}g_{(\varphi_1,\Phi,\varphi_2)}$ and input set Φ, at scale = 100 μm.

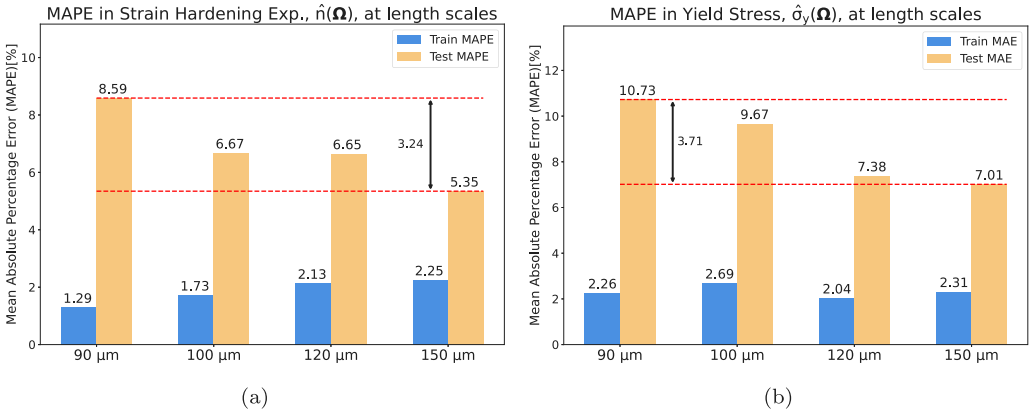

Fig. 15. Overall performance comparison at different length scales using mean absolute percentage error (MAPE) bar graph for predicting (a) Strain Hardening Exponent, $\mathbf{n}$ and (b) Yield Stress, σ_y, using the input set $\mathbf{\Omega}$ using XGBoost regressor.

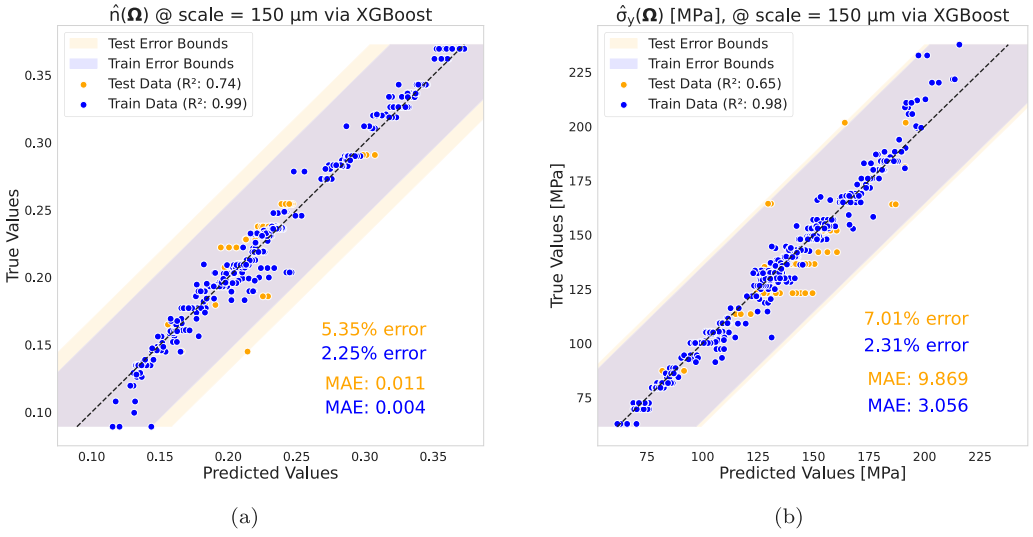

Fig. 16. True and predicted values of (a) Strain Hardening Exponent, $\mathbf{n}$ and (b) Yield Stress, σ_y, for XGBoost regressor using input set $\mathbf{\Omega}$ at scale = 150 µm.

(see Fig. 15(a)). Although the difference in test MAPE between scales 100, 120, and 150 µm is only approximately 1.5%, this indicates that sufficient information is captured between scales 100 and 120 µm for learning patterns that link to accurate $\hat{\mathbf{n}}$. This information is further enriched at 150 µm with test MAPE of 5.35%.

At a scale of 150 µm, there are fewer data points per material specimen due to the reduced number of images and tiles that can be cropped from those images. In Fig. 16(a), the best-performing hyperparameters are employed to train and test the prediction for $\mathbf{n}$. Although the model's test prediction is good, some underfitting is observed due to insufficient training data. Therefore, the robustness and stability of the XGBoost model in predicting $\mathbf{n}$ can be improved by increasing the number of material specimens and data points per specimen at a scale of 150 µm. A similar observation can be made for the prediction of σ_y at a scale of 150 µm. In this case, the training data fits relatively well along the best prediction line, but the same cannot be said for the test data due to insufficient representation of these test data points in the training data (see Fig. 16(b)). Appendix C contains the rest of parity plot comparisons at length scale 90, 100 and 120 µm. For predicting σ_y, the issue of insufficient data is mitigated by the amount of grain boundary information contained in each data point.

8.5. Feature importance using SHAP value analysis

The SHAP value analysis as discussed in Section 7.4 is visualized here using beeswarm plots of the best performing input and model configurations of XGBoost, to understand the importance of microstructure descriptors, statistics, and texture descriptors in predicting $\mathbf{n}$ and σ_y. These SHAP beeswarm plots illustrate feature importance in an XGBoost regressor. Each dot represents a data point, with color indicating feature value (blue for low, red for high). Vertical stacking shows the concentration of data points at a SHAP value, while horizontal spread indicates the magnitude and direction of a feature's impact on the model's output. At a scale of 100 µm (Fig. 17(a)) and 150 µm (Fig. 18(a)), the descending order of importance for input features in predicting $\mathbf{n}$ remains consistent. This consistency suggests that sufficient information is captured even at the 100, µm scale, with further enrichment at the 150 µm scale, as evidenced by the lower test MAPE observed previously. In contrast, the SHAP value beeswarm plots for predicting σ_y (Figs. 17(b) and 18(b)) show distinct variations in the descending order of feature importance. Specifically, ${}^{\mathbf{PCA}}f_3$ at the 100 µm scale does not capture significant information, whereas at the 150 µm scale, it emerges as the third most important feature in predicting σ_y.

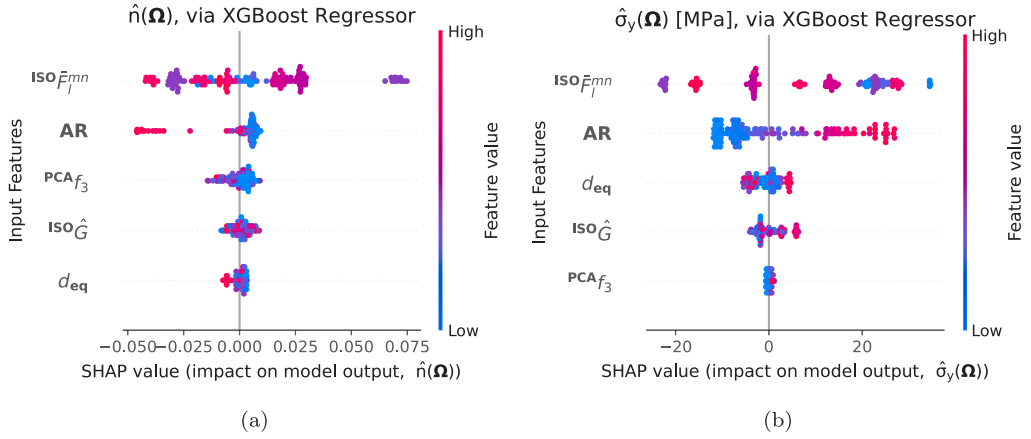

Fig. 17. Feature dependence in the order of importance for (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, for XGBoost regressor model based on SHAP value analysis, at scale = 100 μm.

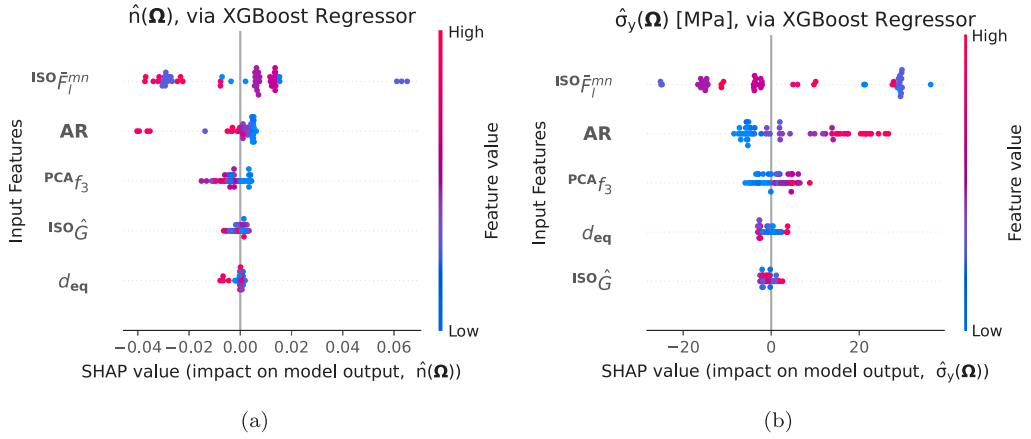

Fig. 18. Feature dependence in the order of importance for (a) Strain Hardening Exponent, **n** and (b) Yield Stress, σ_y, for XGBoost regressor model based on SHAP value analysis, at scale = 150 μm.

The texture descriptor, ${}^{\mathbf{ISO}}\bar{F}_l^{mn}$, is the most important feature in predicting both **n** and $\sigma_\mathbf{y}$ for this pipeline. This observation is well-supported by material science principles and literature on HCP materials, such as magnesium alloys. Magnesium alloys (e.g., ZM21, AZ31) exhibit strong basal texture due to high basal pole intensities aligned near the normal direction, which significantly influences yield stress. Small deviations from this principal alignment then result in higher yield stress in the transverse direction and lower yield stress in the rolling direction due to high activation ability of basal slip [67]. Additionally, the strong basal texture in Mg-alloys and tensile twinning leads to slower strain hardening rates across different orientations [68, 69]. The dependence of uniaxial properties, **n** and $\sigma_\mathbf{y}$, on the normal and material direction makes the statistics of aspect ratio distribution the second most important feature in both predictions. The influence of grain boundary features in restricting twin growth and associated reorientation of the grain orientation, as well as their role in cross slip and deformation-carrying dislocations, underscores the importance of microstructure descriptors. Specifically, ${}^{\mathbf{PCA}}f_3$ is shown to be significant in predicting **n**, as illustrated in Fig. 18 [70]. Both twin boundaries and grain boundaries affect tensile yield stress, with the Hall-Petch relationship serving as a fundamental principle. Dislocation pile-up and the activation of multislip at grain boundaries influence yield stress [71,72]. This is why ${}^{\mathbf{PCA}}f_3$ emerges as the third most important feature in predicting $\sigma_\mathbf{y}$. However, in our dataset, the strong basal texture of Mg-alloys means that the texture descriptor predominantly drives both predictions. Another important observation from Fig. 18 is that the microstructure descriptor, ${}^{\mathbf{ISO}}\hat{G}$, has a greater impact on predicting **n** and a lesser impact on predicting $\sigma_\mathbf{y}$. This can be attributed to the fact that the style and texture of grain boundaries influence slip more significantly, which in turn affects **n**, while having a lesser impact on $\sigma_\mathbf{y}$.

9. Conclusion

In this study, a machine learning pipeline for a generalized approach to property prediction from microstructure and texture descriptors is developed, implemented, and successfully tested. A data extraction and preprocessing pipeline is established to create a coherent database of extruded Mg-alloys, as detailed in Table 1. An image processing and deep learning-based workflow for microstructure binarization and grain statistic calculation is applied to OM microstructure images, which proved to be very useful for making the data more uniform and coherent for further machine learning operations. The XRD pole plots from the database are used to calculate the ODF, which is efficiently represented using GSH coefficients. The images, ODF, and GSH representation of ODF are further employed to compute statistical

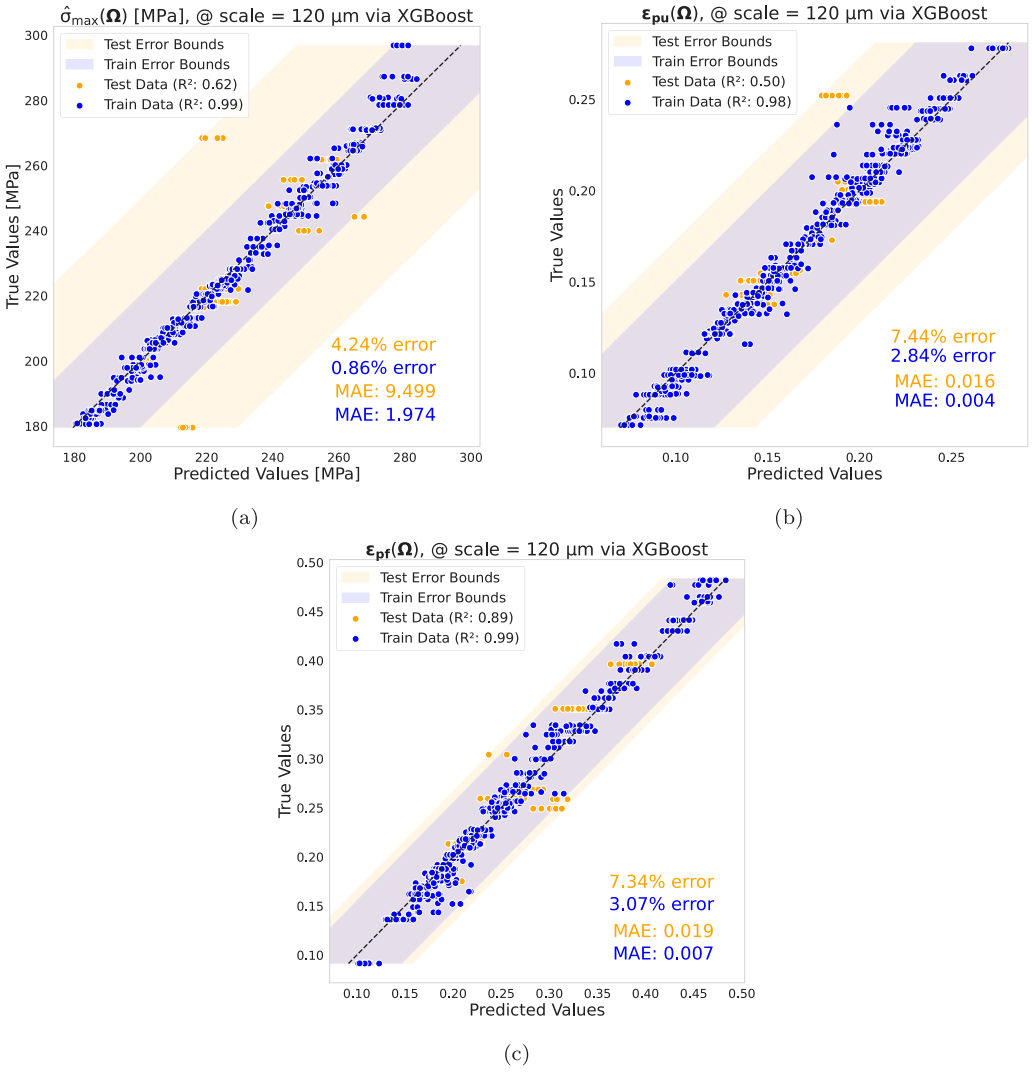

Fig. A.19. True and predicted values of (a) Ultimate tensile stress, σ_{max}, (b) Uniform strain, ε_{pu}, and (c) Fracture strain ε_{pf}, for XGBoost regressor using input set Ω at scale = 120 µm.

descriptors for microstructure and texture. Dimensionality reduction techniques, such as PCA, Isomap, and autoencoder-based vector embedding, are then applied to these descriptors according to their suitability. Finally, their performance in predicting **n** and σ_y is evaluated using GP, XGBoost, and MLP regressors. Several key findings are observed in this study, and they are as follows:

- The XGBoost model emerged as the best non-linear regressor according to test and train MAPE for the database. This performance can be attributed to the complexity of the input features and the sparse nature of the dataset. The GP model came close but suffered from underfitting.
- The combination of PCA-reduced three point spatial correlation $^{PCA}f_3$ and Isomap-reduced gram matrices $^{ISO}\hat{G}$ served as the best choice for microstructure descriptors.
- The Isomap-reduced GSH coefficient vector, $^{ISO}\bar{F}_l^{mn}$, served as the texture descriptor. Generalized Spherical Harmonics represented the ODF most efficiently for preserving information after dimensionality reduction.
- The texture vector and aspect ratio distribution (**AR**) contributed more prominently to the prediction of **n** and σ_y compared to the microstructure descriptor.
- Larger length scales (120 and 150 µm) are more suitable for effectively calculating microstructure statistics for property prediction (primarily for σ_y) because they capture more information.
- Finally, this structure–property pipeline achieved a high predictive performance for unseen material specimens of Mg-alloys, with a MAPE of 6.67% for predicting the strain hardening exponent (**n**) and 7.01% for predicting the tensile yield stress (σ_y).

The performance metrics achieved through this pipeline surpass the current state-of-the-art benchmarks for studies based on training and testing experimental data, rather than data generated via simulation. This paper not only presents a methodology for predicting the properties of extruded Mg-alloys but also provides a template for optimizing this pipeline. This approach helps generalize the methodology for other material types, revealing correlations between properties and microstructure or texture descriptors. Certain limitations were also present in this work, such as the insufficient representation of certain Mg-alloys in the training data. Additionally, including process parameters and alloy composition in the descriptors could enhance the robustness and accuracy of predicting mechanical properties. Increased efforts in collecting data specifically for the purpose to serve such pipeline would increase the quality and quantity of the training data. Another limitation of this study is the absence of any physics-based aspects in property prediction, as the study is entirely data-driven.

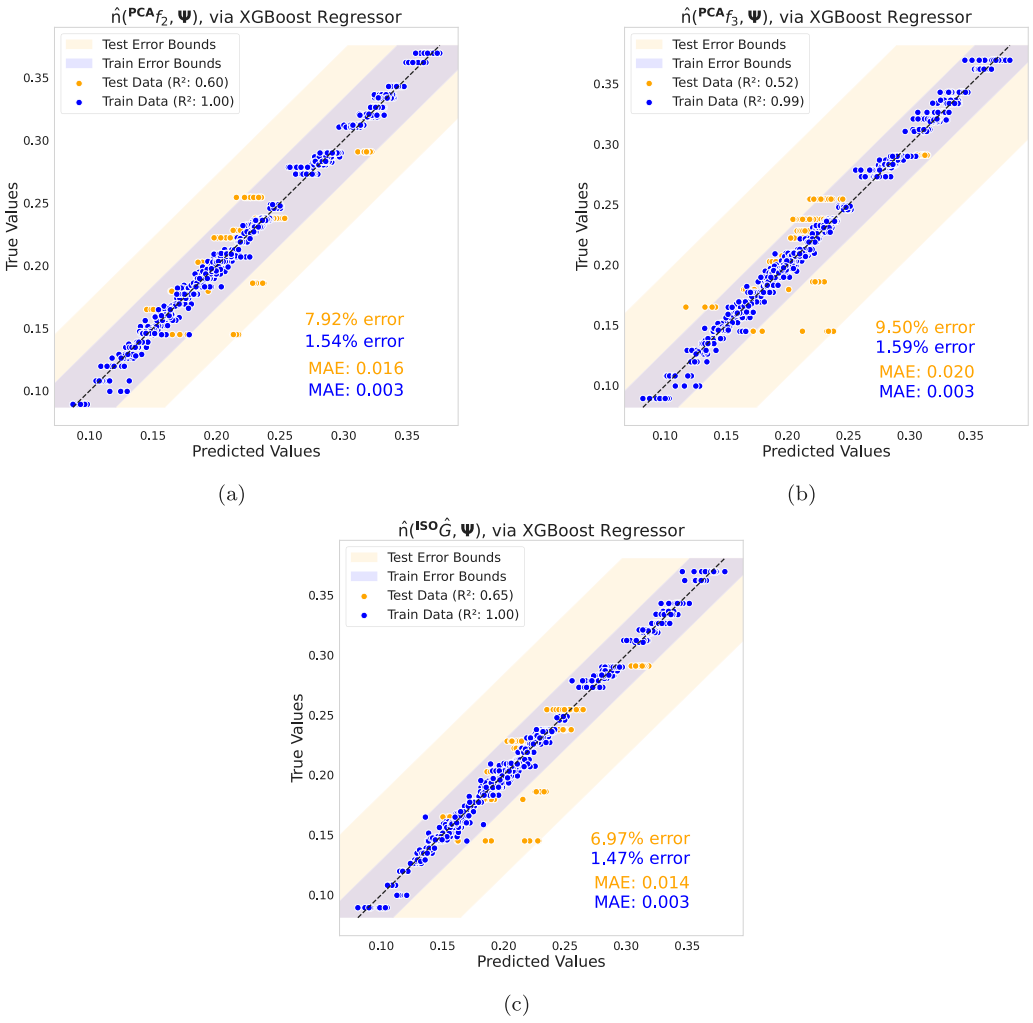

Fig. C.20. True and predicted values of Strain Hardening Exponent, **n**, for XGBoost regressor using input features (a) $^{PCA}f_2$, (b) $^{PCA}f_3$, and (c) $^{ISO}\hat{G}$ with input set Ψ at scale = 100 μm.

Another future scope of this study could involve enhancing the dataset with physics-based simulations to create a hybrid data-driven and physics-based property prediction pipeline.

CRediT authorship contribution statement

Mahish K. Guru: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jan Bohlen:** Resources, Supervision, Writing – review & editing, Project administration. **Roland C. Aydin:** Supervision, Writing – review & editing, Validation. **Noomane Ben Khalifa:** Funding acquisition, Project administration, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Helmholtz Center Hereon, Germany, Leuphana Universität Lüneburg, Germany, and Hamburg University of Technology (TUHH), Germany. The authors would like to thank Dr. Maria Nienaber for her invaluable guidance in magnesium alloys and mechanical property analysis, and Dr. Nowfal Al-Hamdany for his assistance in texture analysis of Magnesium alloys.

Appendix A. Prediction performance for σ_{max}, ε_{pu} and ε_{pf}

In this section, the prediction performance for ultimate tensile stress (σ_{max}), uniform strain (ε_{pu}), and fracture strain (ε_{pf}) is presented for the best-performing regressor, XGBoost. The prediction performance is demonstrated using the best-performing microstructure and texture descriptors in the input feature set **Ω** at a scale of 120, μm. Hyperparameter selection was conducted as discussed earlier in the paper, and the configuration with the most optimal MAPE and MAE is shown in Fig. A.19.

The prediction of σ_{max} shows a good test MAPE value of 4.24%, but the R^2 value suffers due to some material specimens in the test set not being adequately learned during training. A similar issue is observed for the prediction of ε_{pu}, as the data available at a scale of 120, μm is insufficient to capture the values of each material specimen in the dataset. While the test MAPE and R^2 values for ε_{pf} are relatively good, further enhancement of the dataset would still be beneficial.

Appendix B. Performance measures for all investigations

In this section all the performance measures are summarized for XGBoost, GP and MLP models in terms of train and test MAE and MAPE for prediction of **n** and σ_y.

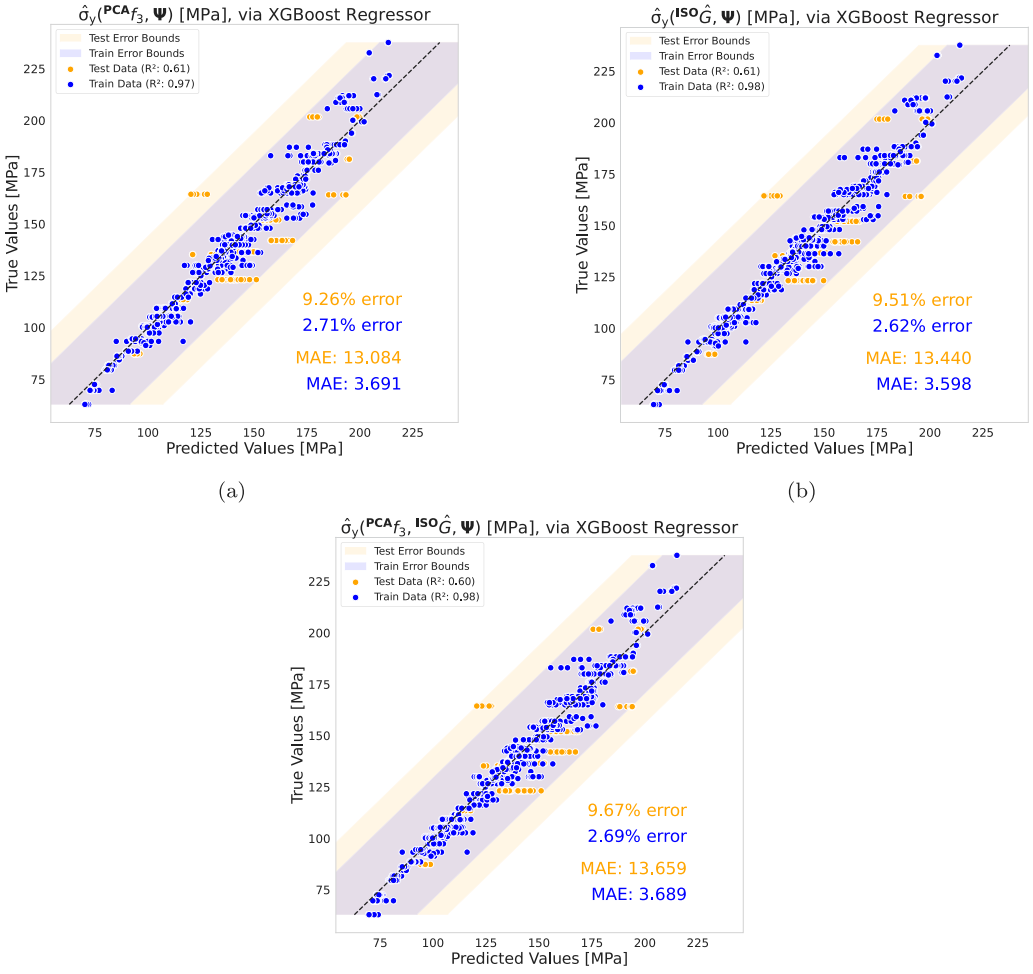

Fig. C.21. True and predicted values of Yield Stress, σ_y, for XGBoost regressor using input features (a) $^{PCA}f_3$, (b) $^{ISO}\hat{G}$, and (c) $^{PCA}f_3$, $^{ISO}\hat{G}$ with input set Ψ at scale = 100 μm.

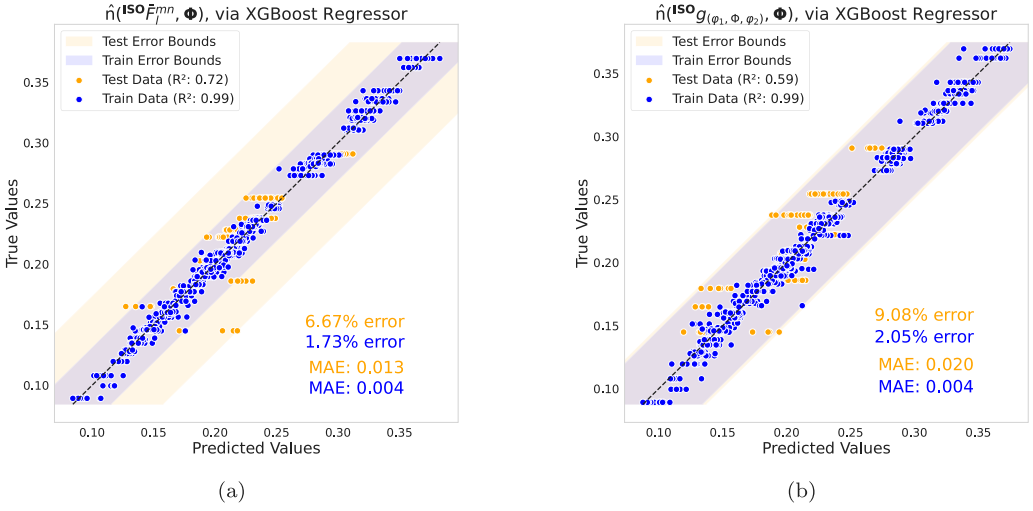

Fig. C.22. True and predicted values of Strain Hardening Exponent, **n**, for XGBoost regressor using input features (a) $^{ISO}\bar{F}_I^{mn}$, and (b) $^{ISO}g_{(\varphi_1,\Phi,\varphi_2)}$ with input set Φ at scale = 100 μm.

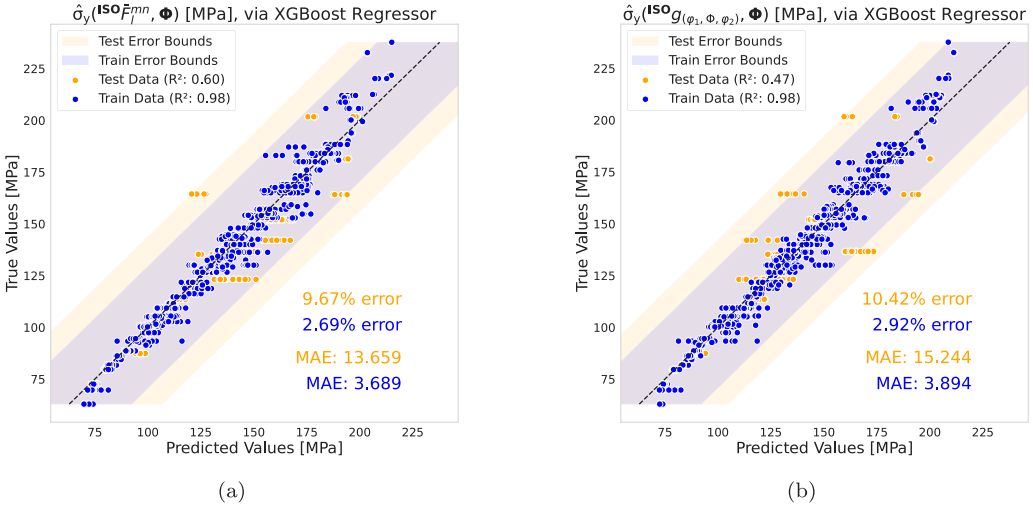

Fig. C.23. True and predicted values of Yield Stress, σ_y, for XGBoost regressor using input features (a) $^{ISO}\bar{F}_l^{mn}$, and (b) $^{ISO}g_{(\varphi_1,\Phi,\varphi_2)}$ with input set Φ at scale = 100 µm.

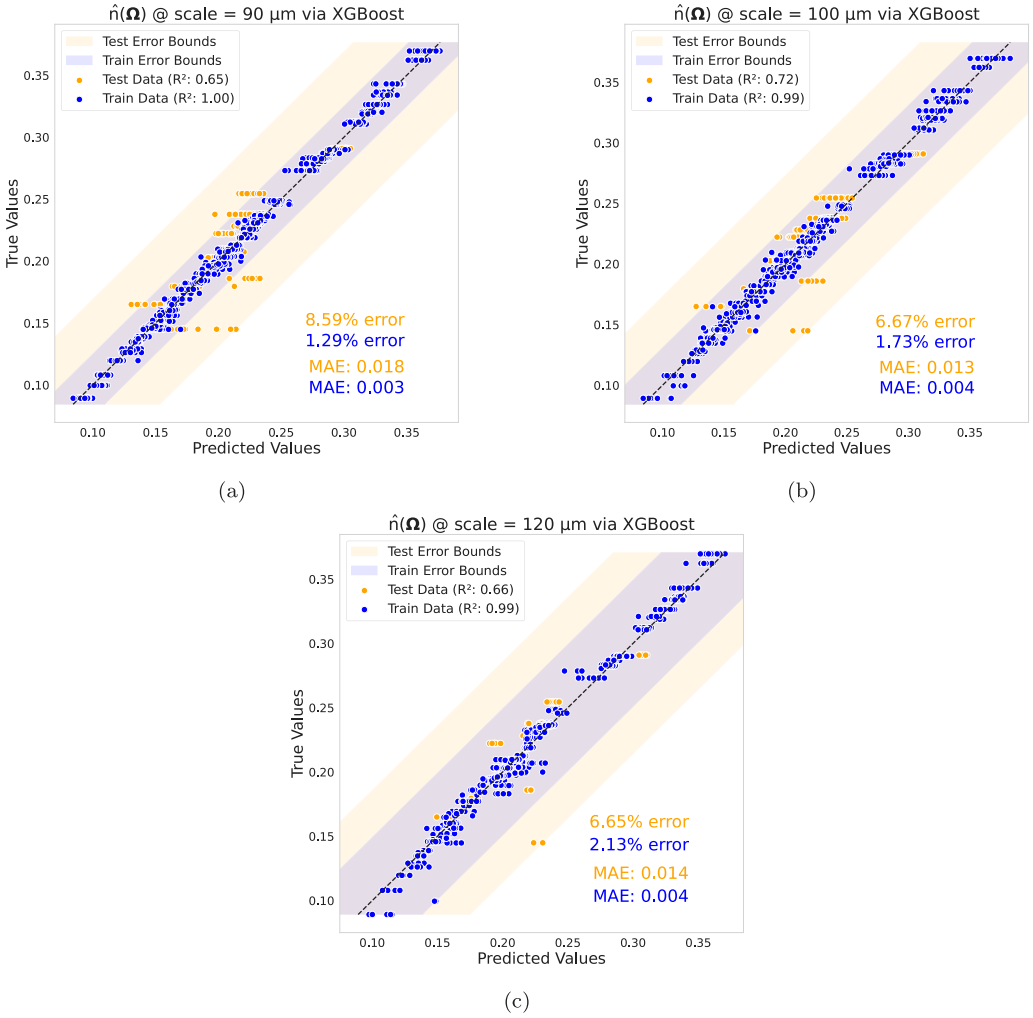

Fig. C.24. True and predicted values of Strain Hardening Exponent, **n**, for XGBoost regressor using with input set Ψ at scale (a) 90, (b) 100, and (c) 120 µm.

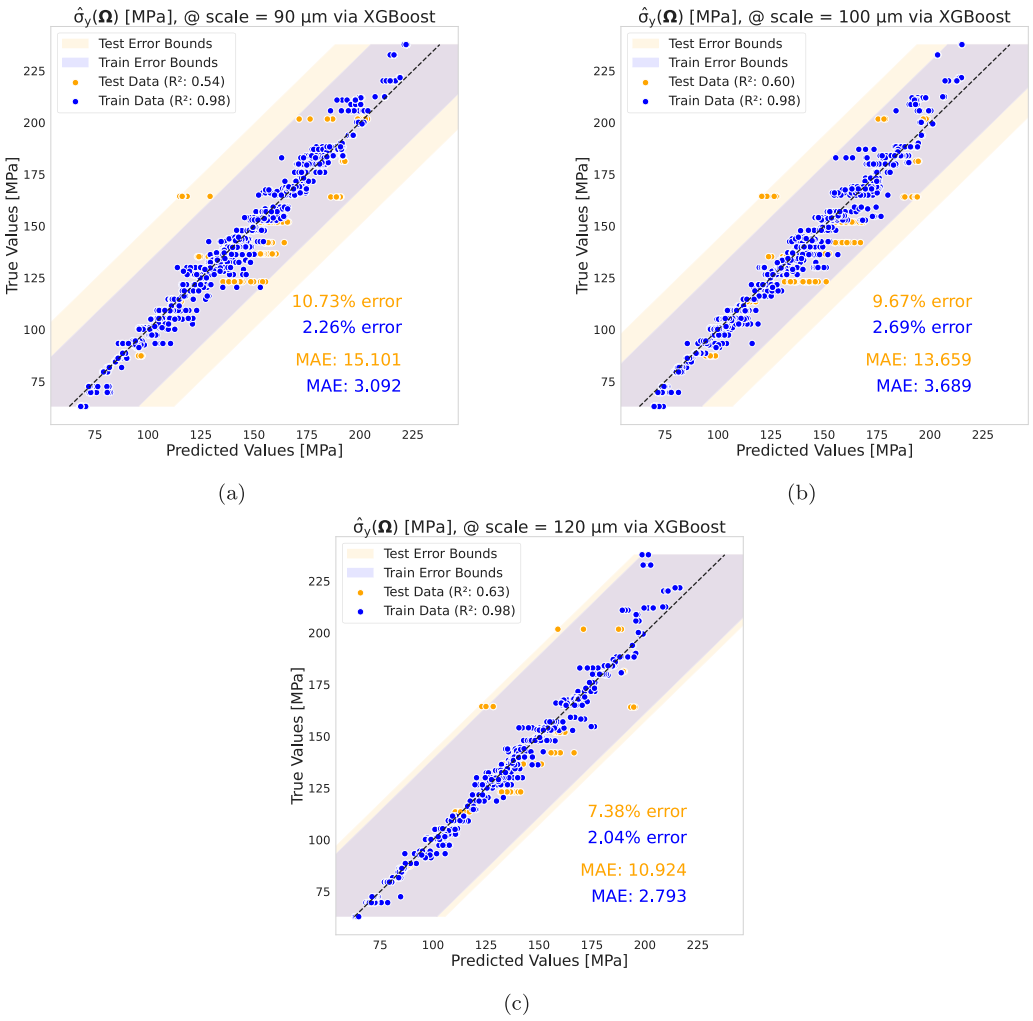

Fig. C.25. True and predicted values of Yield Stress, σ_y, for XGBoost regressor using with input set Ψ at scale (a) 90, (b) 100, and (c) 120 µm.

Table B.4
Input feature vectors and input sets along with their corresponding mean average percentage error (MAPE) and mean average error (MAE) of Strain Hardening Exponent, **n** and Yield Stress, σ_y for XGBoost regressor.

Configuration	Strain hardening exp. (**n**)				Yield stress (σ_y)			
	MAPE		MAE		MAPE		MAE	
	Train	Test	Train	Test	Train	Test	Train	Test
$({}^{\text{PCA}}f_2, \Psi)$ 100 µm	1.54%	7.92%	0.003	0.016	2.62%	9.36%	3.52 MPa	13.26 MPa
$({}^{\text{PCA}}f_3, \Psi)$ 100 µm	1.59%	9.50%	0.003	0.020	2.71%	9.26%	3.69 MPa	13.08 MPa
$({}^{\text{ISO}}\hat{G}, \Psi)$ 100 µm	1.47%	6.97%	0.003	0.014	2.62%	9.51%	3.59 MPa	13.44 MPa
$({}^{\text{PCA}}f_3, {}^{\text{ISO}}\hat{G}, \Psi)$ 100 µm	1.73%	6.67%	0.004	0.013	2.69%	9.67%	3.68 MPa	13.65 MPa
$({}^{\text{VEC}}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 µm	1.44%	8.17%	0.003	0.018	2.69%	10.28%	3.54 MPa	14.09 MPa
$({}^{\text{ISO}}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 µm	2.05%	9.08%	0.004	0.020	2.92%	10.42%	3.89 MPa	15.24 MPa
$({}^{\text{PCA}}f_3, {}^{\text{ISO}}\hat{G}, \Psi)$ 90 µm	1.29%	8.59%	0.003	0.018	2.26%	10.73%	3.09 MPa	15.10 MPa
$({}^{\text{PCA}}f_3, {}^{\text{ISO}}\hat{G}, \Psi)$ 120 µm	2.13%	6.65%	0.004	0.014	2.04%	7.38%	2.79 MPa	10.92 MPa
$({}^{\text{PCA}}f_3, {}^{\text{ISO}}\hat{G}, \Psi)$ 150 µm	2.25%	5.35%	0.004	0.011	2.31%	7.01%	3.05 MPa	9.86 MPa

B.1. XGBoost regressor

See Table B.4.

B.2. GP regressor

See Table B.5.

B.3. MLP regressors

See Table B.6.

Appendix C. Remaining parity plots for the investigations

In this section, we include the comparison of true and predicted values of **n** and σ_y for the investigations conducted in this paper for microstructure, texture and length scales. These comparisons were not shown in the main paper for simplicity.

C.1. Microstructure descriptors

Strain Hardening Exponent, **n**
Yield Stress, σ_y (see Figs. C.20 and C.21).

Table B.5
Input feature vectors and input sets along with their corresponding mean average percentage error (MAPE) and mean average error (MAE) of Strain Hardening Exponent, **n** and Yield Stress, σ_y for GP regressor.

Configuration	Strain hardening exp. (**n**)				Yield stress (σ_y)			
	MAPE		MAE		MAPE		MAE	
	Train	Test	Train	Test	Train	Test	Train	Test
$({}^{PCA}f_2, \Psi)$ 100 μm	3.61%	9.34%	0.008	0.024	3.87%	12.74%	5.39 MPa	12.31 MPa
$({}^{PCA}f_3, \Psi)$ 100 μm	3.28%	9.42%	0.007	0.024	3.51%	12.44%	4.89 MPa	11.99 MPa
$({}^{ISO}\hat{G}, \Psi)$ 100 μm	4.12%	9.92%	0.009	0.025	4.43%	13.06%	6.14 MPa	12.77 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 100 μm	3.09%	10.60%	0.007	0.026	3.26%	13.15%	4.61 MPa	12.97 MPa
$({}^{VEC}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 μm	3.68%	8.68%	0.008	0.021	4.19%	11.96%	6.07 MPa	12.39 MPa
$({}^{ISO}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 μm	1.85%	10.13%	0.004	0.025	2.28%	16.14%	3.58 MPa	16.28 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 90 μm	3.03%	9.78%	0.007	0.023	3.35%	12.35%	4.91 MPa	12.51 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 120 μm	2.99%	10.18%	0.007	0.026	3.05%	14.25%	4.22 MPa	12.91 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 150 μm	2.81%	9.80%	0.006	0.026	2.70%	14.71%	3.68 MPa	12.21 MPa

Table B.6
Input feature vectors and input sets along with their corresponding mean average percentage error (MAPE) and mean average error (MAE) of Strain Hardening Exponent, **n** and Yield Stress, σ_y for MLP regressor.

Configuration	Strain hardening exp. (**n**)				Yield stress (σ_y)			
	MAPE		MAE		MAPE		MAE	
	Train	Test	Train	Test	Train	Test	Train	Test
$({}^{PCA}f_2, \Psi)$ 100 μm	6.63%	8.19%	0.015	0.017	3.49%	14.82%	5.04 MPa	20.91 MPa
$({}^{PCA}f_3, \Psi)$ 100 μm	2.95%	11.53%	0.006	0.023	4.36%	9.81%	6.09 MPa	13.50 MPa
$({}^{ISO}\hat{G}, \Psi)$ 100 μm	2.66%	8.49%	0.006	0.017	4.08%	14.88%	5.94 MPa	12.57 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 100 μm	4.47%	8.60%	0.010	0.018	4.36%	9.98%	6.18 MPa	13.78 MPa
$({}^{VEC}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 μm	4.85%	8.22%	0.010	0.018	6.69%	12.91%	9.40 MPa	17.01 MPa
$({}^{ISO}g_{(\varphi_1,\Phi,\varphi_2)}, \Phi)$ 100 μm	4.68%	8.82%	0.010	0.017	4.04%	13.18%	6.12 MPa	18.11 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 90 μm	4.33%	10.74%	0.009	0.022	3.00%	10.78%	4.73 MPa	15.26 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 120 μm	4.09%	10.59%	0.009	0.020	5.01%	12.40%	6.74 MPa	16.78 MPa
$({}^{PCA}f_3, {}^{ISO}\hat{G}, \Psi)$ 150 μm	3.76%	10.39%	0.008	0.021	5.10%	10.75%	6.74 MPa	14.26 MPa

C.2. Texture descriptors

Strain Hardening Exponent, **n**
Yield Stress, $\boldsymbol{\sigma_y}$ (see Figs. C.22 and C.23).

C.3. Length scale

Strain Hardening Exponent, **n**
Yield Stress, $\boldsymbol{\sigma_y}$ (see Figs. C.24 and C.25).

References

[1] J.M. Rickman, T. Lookman, S.V. Kalinin, Materials informatics: From the atomic-level to the continuum, Acta Mater. 168 (2019) 473–510.

[2] Md Hosne Mobarak, Mariam Akter Mimona, Md. Aminul Islam, Nayem Hossain, Fatema Tuz Zohura, Ibnul Imtiaz, Md Israfil Hossain Rimon, Scope of machine learning in materials research—A review, Appl. Surf. Sci. Adv. 18 (2023) 100523.

[3] William Yi Wang, Jinshan Li, Weimin Liu, Zi-Kui Liu, Integrated computational materials engineering for advanced materials: A brief review, Comput. Mater. Sci. 158 (2019) 42–48.

[4] Xiaoyang Zheng, Xubo Zhang, Ta-Te Chen, Ikumu Watanabe, Deep learning in mechanical metamaterials: From prediction and generation to inverse design, Adv. Mater. 35 (45) (2023) 2302530.

[5] Da Ren, Chenchong Wang, Xiaolu Wei, Qingquan Lai, Wei Xu, Building a quantitative composition-microstructure-property relationship of dual-phase steels via multimodal data mining, Acta Mater. 252 (2023) 118954.

[6] Qian Liu, Wenliang Chen, Vladislav Yakubov, Jamie J. Kruzic, Chun H. Wang, Xiaopeng Li, Interpretable machine learning approach for exploring process-structure-property relationships in metal additive manufacturing, Addit. Manuf. 85 (2024) 104187.

[7] Pingluo Zhao, Yangwei Wang, Bingyue Jiang, Hongmei Zhang, Xingwang Cheng, Qunbo Fan, Neural network modeling of titanium alloy composition-microstructure-property relationships based on multimodal data, Mater. Sci. Eng.: A 879 (2023) 145202.

[8] Xiaoyu Guo, Longjian Li, Gaojie Liu, Huijun Kang, Zongning Chen, Enyu Guo, Jinchuan Jie, Tongmin Wang, Microstructure and properties of cu-3ti-0.3cr-0.15mg alloy designed via machine learning, Mater. Sci. Eng.: A 916 (2024) 147344.

[9] S. Jayasathyakawin, M. Ravichandran, N. Baskar, C. Anand Chairman, R. Balasundaram, Mechanical properties and applications of magnesium alloy – review, Mater. Today: Proc. 27 (2020) 909–913.

[10] H. Wang, X.C. Luo, D.T. Zhang, C. Qiu, D.L. Chen, High-strength extruded magnesium alloys: A critical review, J. Mater. Sci. Technol. 199 (2024) 27–52.

[11] A.G. Beer, 8 - enhancing the extrudability of wrought magnesium alloys, in: Colleen Bettles, Matthew Barnett (Eds.), Advances in Wrought Magnesium Alloys, in: Woodhead Publishing Series in Metals and Surface Engineering, Woodhead Publishing, 2012, pp. 304–322.

[12] Jan Bohlen, Sangbong Yi, Dietmar Letzig, Karl Ulrich Kainer, Effect of rare earth elements on the microstructure and texture development in magnesium–manganese alloys during extrusion, Mater. Sci. Eng.: A 527 (26) (2010) 7092–7098.

[13] Maria Nienaber, Karl Kainer, Dietmar Letzig, Jan Bohlen, Processing effects on the formability of extruded flat products of magnesium alloys, Front. Mater. 6 (2019) 253.

[14] Björn Wiese, Regine Willumeit-Römer, Dietmar Letzig, Jan Bohlen, Alloying effect of silver in magnesium on the development of microstructure and mechanical properties by indirect extrusion, J. Magnes. Alloy. 9 (1) (2021) 112–122.

[15] Robert N. Saunders, Kirubel Teferra, Alaa Elwany, John G. Michopoulos, Dimitris Lagoudas, Metal AM process-structure-property relational linkages using Gaussian process surrogates, Addit. Manuf. 62 (2023) 103398.

[16] Patxi Fernandez-Zelaia, Yuksel C. Yabansu, Surya R. Kalidindi, A comparative study of the efficacy of local/global and parametric/nonparametric machine learning methods for establishing structure–property linkages in high-contrast 3D elastic composites, Integr. Mater. Manuf. Innov. 8 (2019) 67–81.

[17] Noah H. Paulson, Matthew W. Priddy, David L. McDowell, Surya R. Kalidindi, Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics, Acta Mater. 129 (2017) 428–438.

[18] Shuya Dong, Yingying Wang, Jinya Li, Yuanyuan Li, Li Wang, Jinglai Zhang, Machine Learning Aided Prediction and Design for the Mechanical Properties of Magnesium Alloys, Met. Mater. Int. 30 (3) (2023) 593–606.

[19] Björn Wiese, Sven Berger, Jan Bohlen, Maria Nienaber, Daniel Höche, Property design of extruded magnesium-gadolinium alloys through machine learning, Mater. Today Commun. 36 (2023) 106566.

[20] Frederic E. Bock, Roland C. Aydin, Christian J. Cyron, Norbert Huber, Surya R. Kalidindi, Benjamin Klusemann, A review of the application of machine learning and data mining approaches in continuum materials mechanics, Front. Mater. 6 (2019).

[21] A. Di Gianfrancesco, 8 - technologies for chemical analyses, microstructural and inspection investigations, in: Augusto Di Gianfrancesco (Ed.), Materials for Ultra-Supercritical and Advanced Ultra-Supercritical Power Plants, Woodhead Publishing, 2017, pp. 197–245.

[22] Claudia Gorynski, Max Frei, Frank Einar Kruis, Markus Winterer, Machine learning based quantitative characterization of microstructures, Acta Mater. 256 (2023) 119106.
[23] E.O. Hall, The deformation and ageing of mild steel: III discussion of results, Proc. Phys. Soc. Sect. B 64 (9) (1951) 747.
[24] Norman James Petch, The cleavage strength of polycrystals, 1953.
[25] James G. Berryman, Measurement of spatial correlation functions using image processing techniques, J. Appl. Phys. 57 (1985) 2374–2384.
[26] H. Singh, A.M. Gokhale, S.I. Lieberman, S. Tamirisakandala, Image based computations of lineal path probability distributions for microstructure representation, Mater. Sci. Eng.: A 474 (1) (2008) 104–111.
[27] S. Torquato, J. Beasley, Yee Chiew, Two-point cluster function for continuum percolation, J. Chem. Phys. 88 (1988) 6540–6547.
[28] Ahmet Cecen, Tony Fast, Surya Kalidindi, Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure, Integr. Mater. Manuf. Innov. 5 (2016).
[29] Nicholas Lubbers, Turab Lookman, Kipton Barros, Inferring low-dimensional microstructure representations using convolutional neural networks, Phys. Rev. E 96 (2016).
[30] Paul Seibert, Alexander Raßloff, Yichi Zhang, Karl A. Kalina, Paul Reck, Daniel Peterseim, Markus Kästner, Reconstructing microstructures from statistical descriptors using neural cellular automata, Integr. Mater. Manuf. Innov. (2023).
[31] F. Wagner, Texture determination by using x ray diffraction, in: Jesus Ma. Rincon, Maximina Romero (Eds.), Characterization Techniques of Glasses and Ceramics, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999, pp. 169–186.
[32] Fan Xiong, B.A. Parker, The determination of complete pole figures using the reflection method, Texture, Stress. Microstruct. 6 (2) (1984) 431654.
[33] H.J. Bunge, Fabric analysis by orientation distribution functions, Tectonophysics 78 (1) (1981) 1–21.
[34] Adnan Eghtesad, Timothy J. Barrett, Marko Knezevic, Compact reconstruction of orientation distributions using generalized spherical harmonics to advance large-scale crystal plasticity modeling: Verification using cubic, hexagonal, and orthorhombic polycrystals, Acta Mater. 155 (2018) 418–432.
[35] Eamonn Keogh, Abdullah Mueen, Curse of dimensionality, in: Claude Sammut, Geoffrey I. Webb (Eds.), Encyclopedia of Machine Learning, Springer US, Boston, MA, 2010, pp. 257–258.
[36] Jingran Wang, Yi Liu, Peigong Li, Zhenxing Lin, Stavros Sindakis, Sakshi Aggarwal, Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality, J. Knowl. Econ. 15 (2023) 1–20.
[37] M. Nienaber, Einfluss der thermomechanischen behandlung auf die mikrostruktur und textur von magnesiumflachhalbzeugen, 2023, Online available at: (DOI). Nienaber, M.: Einfluss der thermomechanischen Behandlung auf die Mikrostruktur und Textur von Magnesiumflachhalbzeugen. Technische Universität Berlin, 2023..
[38] Guadalupe Cano-Castillo, José Victoria-Hernández, Jan Bohlen, Dietmar Letzig, Karl Ulrich Kainer, Effect of Ca and nd on the microstructural development during dynamic and static recrystallization of indirectly extruded Mg–Zn based alloys, Mater. Sci. Eng.: A 793 (2020) 139527.
[39] G. Kurz, M. Nienaber, J. Bohlen, D. Letzig, K.U. Kainer, Variation of extrusion process parameter for the magnesium alloy ME21, in: J. Brian Jordon, Victoria Miller, Vineet V. Joshi, Neale R. Neelameggham (Eds.), Magnesium Technology 2020, Springer International Publishing, Cham, 2020, pp. 181–188.
[40] Jochen Harmuth, Björn Wiese, Jan Bohlen, Thomas Ebel, Regine Willumeit, Wide range mechanical customization of mg-gd alloys with low degradation rates by extrusion, Front. Mater. (2019).
[41] Björn Wiese, Jochen Harmuth, Regine Willumeit-Römer, Jan Bohlen, Property variation of extruded mg-gd alloys by mn addition and processing, Crystals 12 (8) (2022).
[42] Alexandre Bordas, Jingchao Zhang, Juan C. Nino, Application of deep learning workflow for autonomous grain size analysis, Molecules 27 (15) (2022).
[43] Juwon Na, Jaejun Lee, Seong-Hoon Kang, Se-Jong Kim, Seungchul Lee, Label-free grain segmentation for optical microscopy images via unsupervised image-to-image translation, Mater. Charact. 206 (2023) 113410.
[44] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, 2, 2005, pp. 60–65 vol. 2.
[45] Karel Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics Gems IV, Academic Press Professional, Inc., USA, 1994, pp. 474–485.
[46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick, Segment anything, 2023.
[47] Bing Wang, Shaosheng Fan, An improved CANNY edge detection algorithm, 2009 Second. Int. Work. Comput. Sci. Eng. 1 (2009) 497–500.
[48] Yuyu Li, Tingzhuang Han, Zhibing Chu, Chun Xue, Qianhua Yang, Xiaodong Zhao, Hong Gao, Effect of loading direction on the tensile properties and texture evolution of AZ31 magnesium alloy, Crystals 11 (10) (2021).
[49] David Brough, Daniel Wheeler, Surya Kalidindi, Materials knowledge systems in python - a data science framework for accelerated development of hierarchical materials, Integr. Mater. Manuf. Innov. 6 (2016) 1–18.
[50] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks, CoRR (2015).
[51] Paul Seibert, Alexander Raßloff, Karl Kalina, Marreddy Ambati, Markus Kästner, Microstructure characterization and reconstruction in python: MCRpy, Integr. Mater. Manuf. Innov. 11 (3) (2022) 450–466.
[52] Florian Bachmann, Ralf Hielscher, Helmut Schaeben, Grain detection from 2d and 3d EBSD data—Specification of the MTEX algorithm, Ultramicroscopy 111 (12) (2011) 1720–1733.
[53] Timothy J. Barrett, Adnan Eghtesad, Rodney J. McCabe, Bjørn Clausen, Donald W. Brown, Sven C. Vogel, Marko Knezevic, A generalized spherical harmonics-based procedure for the interpolation of partial datasets of orientation distributions to enable crystal mechanics-based simulations, Materialia 6 (2019) 100328.
[54] Surya R. Kalidindi, 4 - reduced-order representations of spatial correlations, in: Surya R. Kalidindi (Ed.), Hierarchical Materials Informatics, Butterworth-Heinemann, Boston, 2015, pp. 111–127.
[55] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.
[56] Yang Jiaoyun, Haipeng Wang, Huitong Ding, Ning An, Gil Alterovitz, Nonlinear dimensionality reduction methods for synthetic biology biobricks' visualization, BMC Bioinformatics 18 (2017).
[57] Daniel Engel, Lars Hüttenberger, Bernd Hamann, A survey of dimension reduction methods for high-dimensional data analysis and visualization, in: Visualization of Large and Unstructured Data Sets, 2011.
[58] Olivier Lezoray, Leo Grady (Eds.), Image Processing and Analysis with Graphs: Theory and Practice, first ed., CRC Press, 2012.
[59] Iman Aganj, Christophe Lenglet, Guillermo Sapiro, Essa Yacoub, Kamil Ugurbil, Noam Harel, Reconstruction of the orientation distribution function in single- and multiple-shell q-ball imaging within constant solid angle, Magn. Reson. Med. 64 (2) (2010) 554–566.
[60] Tianqi Chen, Carlos Guestrin, XGBoost: A scalable tree boosting system, CoRR (2016).
[61] Hervé A. Bourlard, Nelson Morgan, Multilayer perceptrons, in: Connectionist Speech Recognition: A Hybrid Approach, Springer US, Boston, MA, 1994, pp. 59–80.
[62] Jie Wang, An intuitive tutorial to Gaussian process regression, Comput. Sci. Eng. 25 (4) (2023) 4–11.
[63] Jens Schreiter, Duy Nguyen-Tuong, Marc Toussaint, Efficient sparsification for Gaussian process regression, Neurocomputing 192 (2016).
[64] Ian Covert, Scott Lundberg, Su-In Lee, Understanding global feature contributions with additive importance measures, 2020.
[65] Scott Lundberg, Su-In Lee, A unified approach to interpreting model predictions, 2017.
[66] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, Su-In Lee, Explainable AI for trees: From local explanations to global understanding, 2019.
[67] Jan Bohlen, Marcus R. Nürnberg, Jeremy W. Senn, Dietmar Letzig, Sean R. Agnew, The texture and anisotropy of magnesium–zinc–rare earth alloy sheets, Acta Mater. 55 (6) (2007) 2101–2112.
[68] Feilong Guo, Huihui Yu, Chenyu Wu, Yunchang Xin, Cong He, Qing Liu, The mechanism for the different effects of texture on yield strength and hardness of Mg alloys, Sci. Rep. 7 (2017).
[69] Xiaoyuan Lou, R. Boger, S.R. Agnew, R. Wagoner, Hardening evolution of AZ31B Mg sheet, Int. J. Plast. 23 (2007) 44–86.
[70] Babak Ravaji, Shailendra Joshi, A crystal plasticity investigation of grain size-texture interaction in magnesium alloys, Acta Mater. 208 (2021) 116743.
[71] Huihui Yu, Yunchang Xin, Adrien Chapuis, Xiaoxu Huang, Renlong Xin, Qing Liu, The different effects of twin boundary and grain boundary on reducing tension-compression yield asymmetry of Mg alloys, Sci. Rep. 6 (2016) 29283.
[72] H.-H. Fu, D.J. Benson, M.A. Meyers, Analytical and computational description of effect of grain size on yield stress of metals, Acta Mater. 49 (13) (2001) 2567–2582.