Modeling Grain Size Using Grain Orientation in Polycrystalline Materials

By Andrew Dahlstrom
M.Sc. Candidate
Data Science & Artificial Intelligence
San Francisco State University
5/25/2024

Executive Summary

This research investigates using machine learning techniques to predict grain sizes in polycrystalline materials from grain orientation data. Polycrystalline materials have properties significantly influenced by grain size. Smaller grains typically enhance material strength, while larger grains improve thermal and electrical conductivity. The study aimed to explore the relationship between grain orientation and grain size, and develop predictive models to uncover patterns and correlations to estimate grain size accurately.

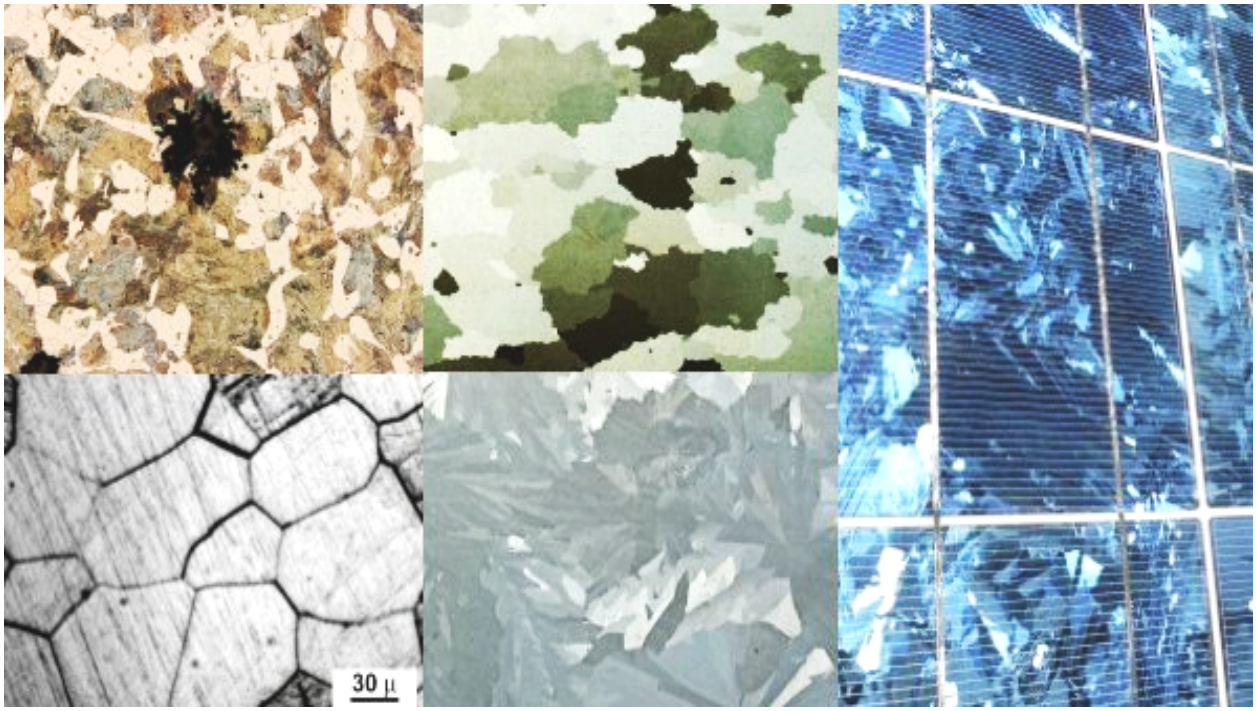
Data was extracted from .grn2 and .ori files to construct a 4D dataset of grain IDs and Euler angles, resulting in 341 samples. Initial exploration included statistical analyses and visualizations of grain sizes and orientations. K-Means clustering revealed clearer distinctions in grain sizes when based on maximum Euler angles. Two-point correlation statistics quantified spatial relationships within the grain structure, with lags of 4, 8, and 16 chosen for their effectiveness in capturing spatial relationships and corresponding closely to grain sizes.

Various regression models, including Ridge, Lasso, Random Forest, and KNN, were tested. KNN achieved the best performance with an MAE of 2.55. Classification models, including Random Forest, KNN, and SVM, showed SVM as the best performer with an MAE of 2.4 due to its capability in handling non-linear classification boundaries. The study demonstrated that classification models, especially SVM, outperformed regression models in predicting grain sizes, likely due to the problem's nature aligning better with classification tasks. The results highlight the importance of model selection and hyperparameter tuning.

In conclusion, this research showed the potential of machine learning in predicting grain sizes from microstructural data, with SVM excelling in handling complex data. SVM performed particularly well with smaller grain sizes versus larger grain sizes.

Table of Contents

1. Introduction



(https://en.wikipedia.org/wiki/Crystallite)

1.1 Background

Polycrystalline materials, composed of numerous small crystals or grains, are fundamental to various industries, including aerospace, automotive, electronics, and construction. The microstructure of these materials, specifically the size and arrangement of their grains, significantly influences their mechanical properties. Grain size plays a crucial role in determining the strength, thermal stability, and electrical conductivity of polycrystalline materials. For instance, smaller grain sizes generally enhance the material's strength and resistance to deformation due to grain boundary strengthening. Conversely, larger grains can improve thermal and electrical conductivity due to reduced grain boundary scattering.

Understanding and predicting grain size is essential for material scientists and engineers aiming to tailor materials for specific applications. Accurate prediction of grain

size from microstructural data allows for the design of materials with desired properties, optimizing performance and durability in practical applications. This predictive capability is particularly important in industries where material performance is critical, such as in the manufacturing of jet engines, microelectronics, and structural components for buildings and bridges.

The research problem addressed in this study is the need to predict grain sizes based on grain orientation data, specifically Euler angles. Grain orientation affects the anisotropic properties of materials, meaning their properties vary with direction. By leveraging machine learning techniques, this research aims to establish a reliable method for predicting grain size from orientation data, thereby advancing the field of materials science.

1.2 Objective

The main objectives of this research are to:

1. Extract and preprocess grain orientation data from polycrystalline material samples.
2. Explore the relationship between grain orientation and grain size using statistical and machine learning techniques.
3. Develop predictive models to estimate grain size from grain orientation data.
4. Evaluate the performance of various regression and classification models in predicting grain size.
5. Identify the most effective model and hyperparameter configurations for accurate grain size prediction.

The hypothesis underlying this research is that grain orientations, represented by Euler angles, can predict grain sizes accurately. By analyzing the spatial distribution of these orientations within the grains, it is possible to uncover patterns and correlations that can be used to estimate grain size. This hypothesis is grounded in the understanding that

the physical properties of polycrystalline materials are direction-dependent, and thus, the orientation of grains should provide valuable insights into their size and overall material properties.

2. Literature Review

2.1 Previous Studies

Predicting grain size in polycrystalline materials has garnered significant attention in the field of materials science. Several studies have focused on leveraging machine learning techniques to establish structure-property linkages within these materials.

In the study "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" by Fernandez-Zelaia, Yabansu, and Kalidindi, the authors compare various machine learning methods to predict material properties from microstructural data. They demonstrate the efficacy of both local/global and parametric/nonparametric machine learning techniques in capturing complex relationships within high-contrast 3D elastic composites. This study underscores the potential of advanced machine learning methods in materials science but also highlights the need for further exploration into different types of microstructural data, such as grain orientation information .

Another significant contribution is the paper "Microstructure Informatics Using Higher-Order Statistics and Efficient Data-Mining Protocols" by Kalidindi et al. This research emphasizes the use of higher-order statistical methods and efficient data-mining protocols to analyze microstructural data. By employing advanced statistical techniques, the authors successfully identify and quantify key features within the microstructure that correlate with material properties. This work provides a strong foundation for using statistical methods to analyze complex microstructural datasets and demonstrates the importance of efficient data-mining protocols in the field.

While these studies have advanced the understanding of structure-property linkages using various machine learning and statistical methods, gaps remain. Notably, there is

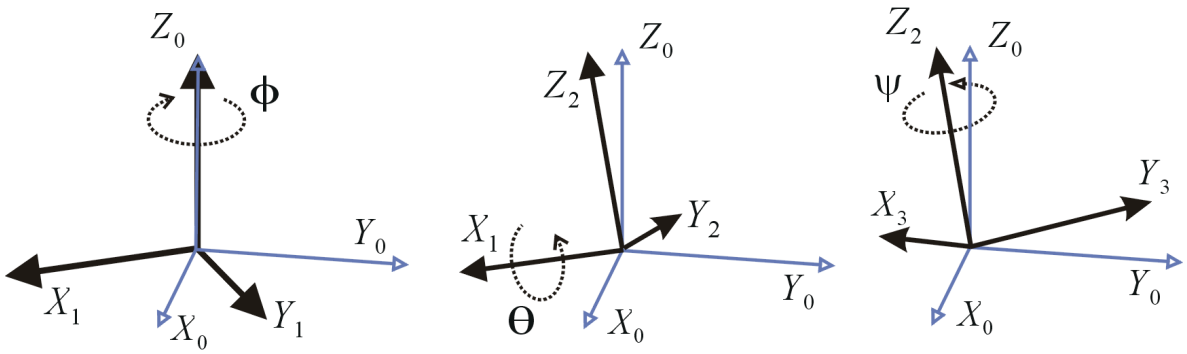
limited research on the application of grain orientation data, specifically Euler angles, for predicting grain size. Additionally, the use of interpretable machine learning models, as opposed to more complex deep learning techniques, is underexplored. This research aims to address these gaps by focusing on the predictive power of grain orientation data using interpretable regression and classification models.

2.2 Theoretical Framework

The theoretical framework of this research is grounded in several key concepts: Euler angles, two-point correlation statistics, and Principal Component Analysis (PCA).

Euler Angles

Euler angles are a set of three angles that describe the orientation of a rigid body relative to a fixed coordinate system. In the context of polycrystalline materials, Euler angles are used to define the orientation of individual grains. These angles are crucial for understanding the anisotropic properties of materials, as the orientation of grains can significantly impact the material's mechanical and thermal behavior. By analyzing the Euler angles, we can gain insights into the spatial distribution of grain orientations and their relationship to grain size.



Two-Point Correlation Statistics

Two-point correlation statistics are a powerful tool for quantifying spatial relationships within a material. These statistics measure the probability of finding pairs of points at a certain distance with specific properties, such as grain orientation. In this research, two-point correlation statistics are computed for different lags to capture the spatial correlation of Euler angles within the grains. This approach extends traditional image analysis techniques to 3D, providing a more comprehensive understanding of the material's microstructure.

Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while retaining most of the variance in the data. In this research, PCA is applied to the two-point correlation statistics to reduce the complexity of the dataset and focus on the most informative features. By capturing the principal components, PCA helps in improving the efficiency and performance of the predictive models, making it easier to identify patterns and correlations in the data.

Together, these theoretical concepts provide a robust framework for analyzing grain orientation data and developing predictive models for grain size. By combining Euler angles, two-point correlation statistics, and PCA, this research aims to uncover meaningful relationships within the data that can be used to accurately predict grain size in polycrystalline materials.

3. Methodology

3.1 Data Extraction and Preparation

The data for this study was obtained from .grn2 and .ori files, which provide detailed information about the microstructure of polycrystalline materials. The .grn2 files contain grain IDs, representing distinct regions within the material, while the .ori files include corresponding Euler angles that describe the orientation of these grains in three-dimensional space.

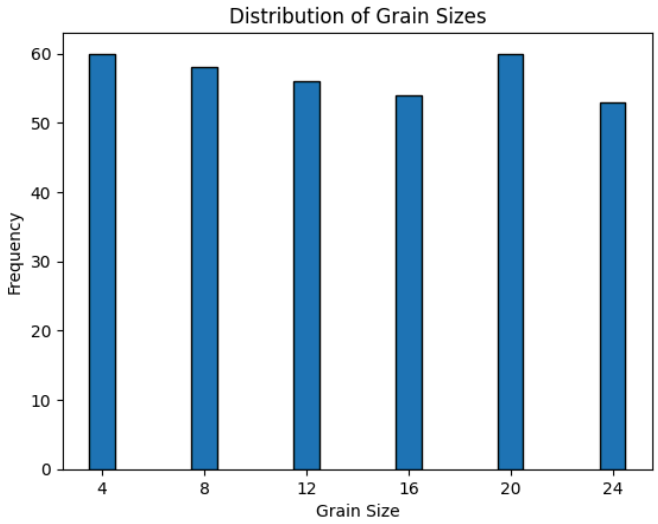
The initial step involved extracting the grain ID data from the .grn2 files and reshaping them into 3D voxel arrays. Each 3D voxel array represented a 32x32x32 grid, capturing the spatial distribution of grains within the material. Subsequently, the Euler angles from the .ori files were mapped to these grain IDs, resulting in a comprehensive 4D dataset (32x32x32x3) that included the orientation of each grain.

After processing, the final dataset comprised 341 complete samples, each containing detailed grain orientation and size information. This dataset served as the foundation for subsequent analysis and model development.

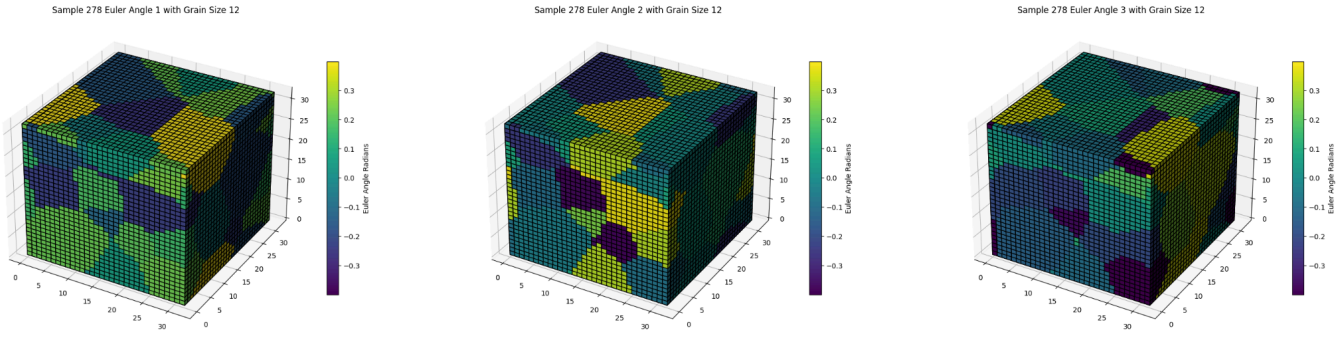
3.2 Data Exploration

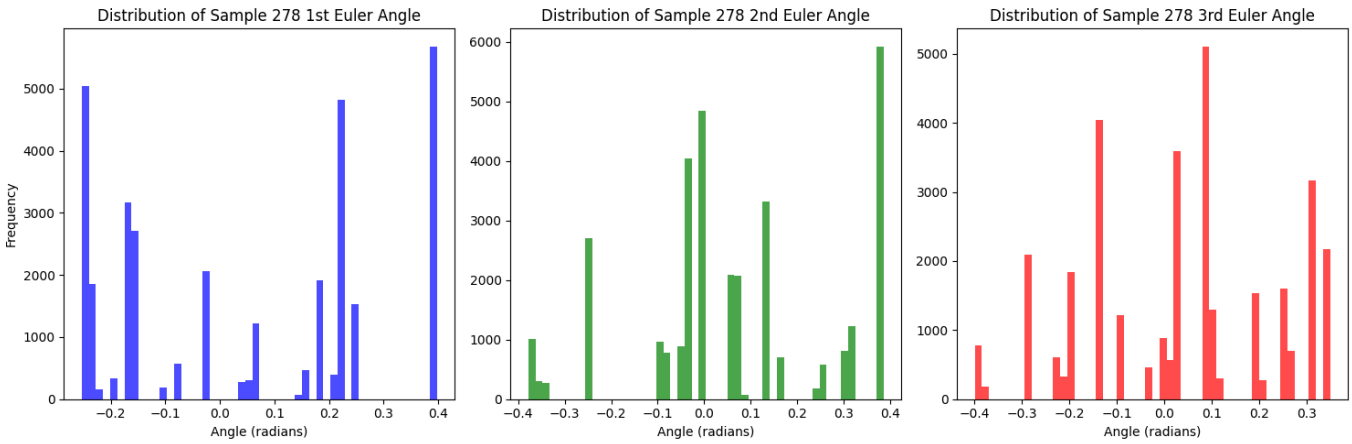
Initial data exploration was conducted to gain insights into the distribution of grain sizes and the orientation of grains within the material. Statistical analyses, including the calculation of mean and standard deviation, were performed on the dataset to summarize the key characteristics.

Histograms were employed to visualize the frequency distribution of grain sizes, providing a clear overview of the range and common grain size values within the samples.



Additionally, 3D voxel plots were created to illustrate the spatial configuration of Euler angles, offering a visual representation of grain orientations. These visualizations helped identify patterns and anomalies in the dataset, informing the development of predictive models.

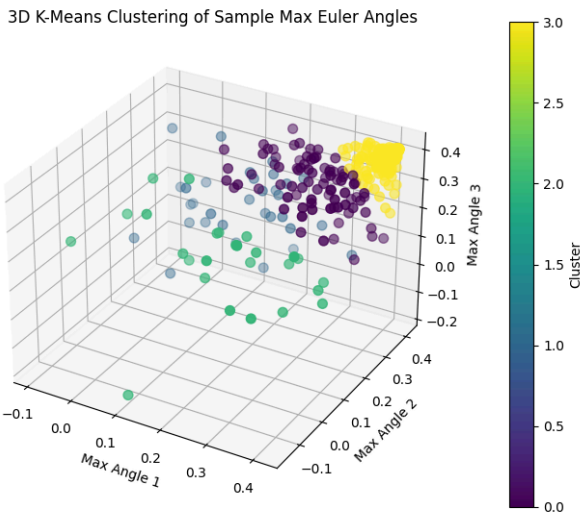


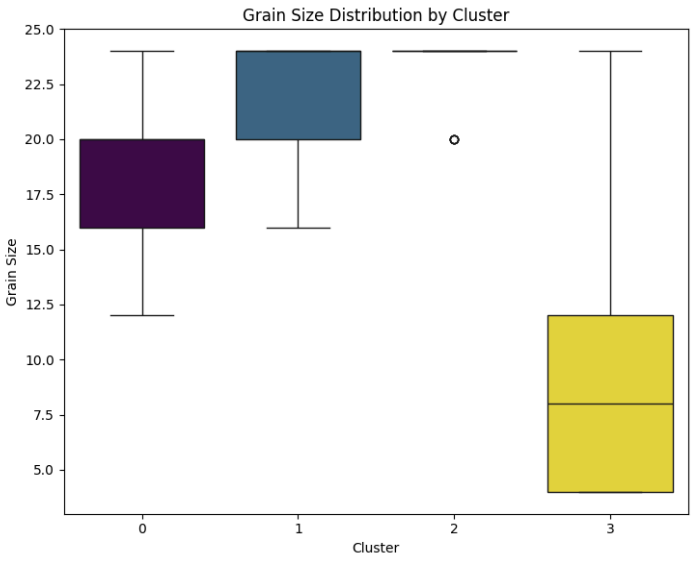


3.3 Unsupervised Learning: K-Means Clustering

K-Means clustering was utilized as an unsupervised learning technique to explore inherent patterns in the dataset without predefined labels. The rationale for using K-Means clustering was to identify natural groupings within the data that could provide insights into the relationship between grain orientation and size.

Clustering was performed using both the mean and maximum Euler angles for each sample. The mean Euler angles provided an average orientation, while the maximum Euler angles captured the most extreme orientations within each sample. By comparing the results of these two approaches, it was observed that clustering based on maximum Euler angles offered clearer distinctions among different grain sizes. For example, the yellow cluster below of grains with high positive rotations for all three Euler angles, corresponded with a small grain size between four and twelve. This insight was crucial for refining feature engineering in subsequent predictive modeling.

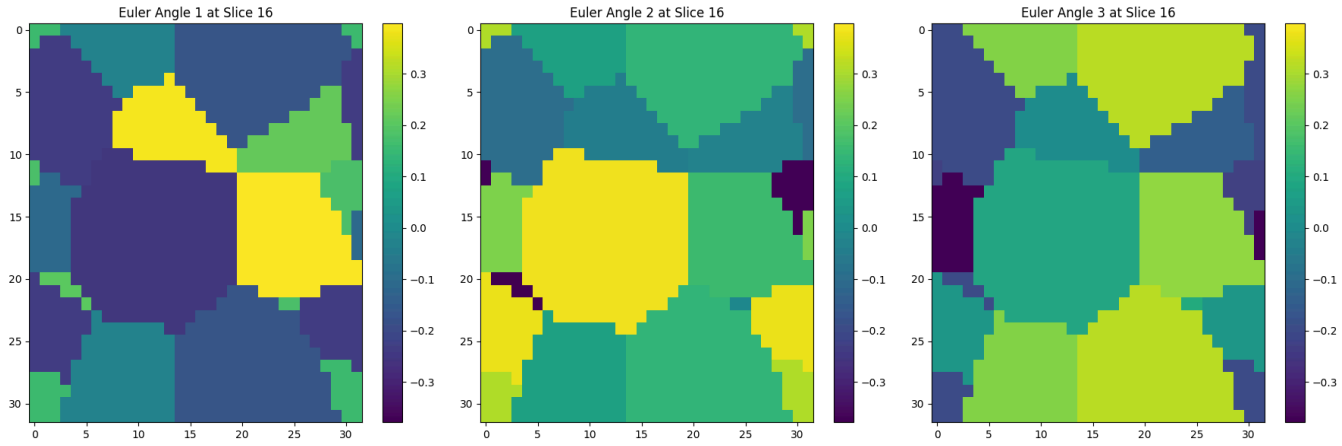




3.4 Two-Point Correlation Statistics

Two-point correlation statistics were computed to quantify the spatial relationship between points within the grain structure. These statistics measure how the orientation of grains varies relative to each other at specific distances, providing a detailed characterization of the material's microstructure.

The computation involved selecting specific lags (distances) and calculating the correlation between points at these lags in three dimensions. Initially, many lag values were experimented with to determine the most effective distances for capturing spatial relationships. Ultimately, lags of 4, 8, and 16 were chosen to balance the length of training time with the most effective lag distances, which closely corresponded to grain sizes. These selected lags aimed to capture both immediate and longer-range spatial relationships, providing a comprehensive analysis of how grain orientations vary within the material. By analyzing these correlations, the study aimed to uncover patterns that could be predictive of grain size.

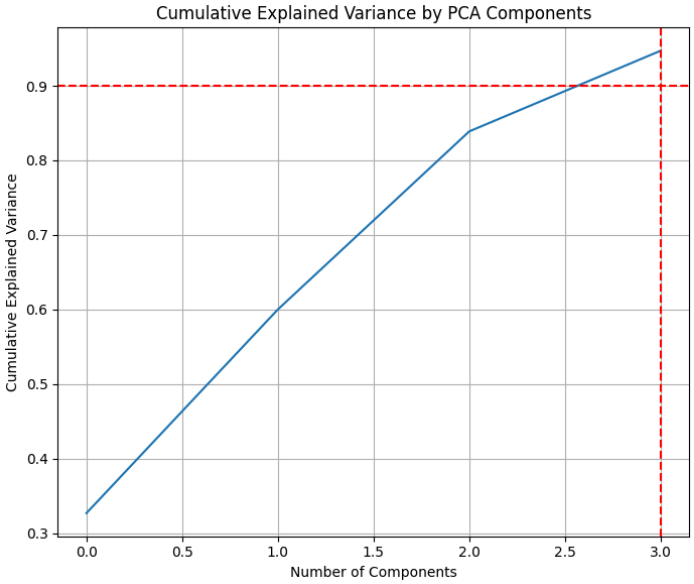


3.5 Principal Component Analysis (PCA)

Given the high-dimensional nature of the two-point correlation statistics, Principal Component Analysis (PCA) was employed to reduce dimensionality while preserving the most significant variance in the data. The rationale for using PCA was to streamline

the computational process and enhance model performance by focusing on the principal components that captured the most critical information.

PCA was initially applied to the two-point correlation statistics to transform the data into a set of uncorrelated components. By retaining components that explained 90% of the variance, the dimensionality of the dataset was reduced from nine features to four principal components, making it more manageable for machine learning models. However, this step did not significantly improve the accuracy of the predictive models and, in some cases, actually decreased it. As a result, PCA was not utilized in the final training data for the models developed in this study.



4. Regression Modeling

4.1 Model Selection

In this study, four regression models were chosen to predict grain sizes from the extracted features: Ridge Regression, Lasso Regression, Random Forest, and K-Nearest Neighbors (KNN).

- Ridge Regression: This linear model includes a regularization term to prevent overfitting, making it suitable for datasets with multicollinearity.
- Lasso Regression: Similar to Ridge, Lasso Regression performs regularization but can also perform feature selection by driving some coefficients to zero.
- Random Forest: This ensemble learning method constructs multiple decision trees and merges them to improve the predictive performance and control overfitting.
- K-Nearest Neighbors (KNN): A non-parametric method used for classification and regression that predicts the output based on the k-nearest neighbors in the feature space.

The selection of these models was based on their diverse approaches to handling regression tasks, which provided a comprehensive comparison of linear and non-linear methods, as well as parametric and non-parametric techniques.

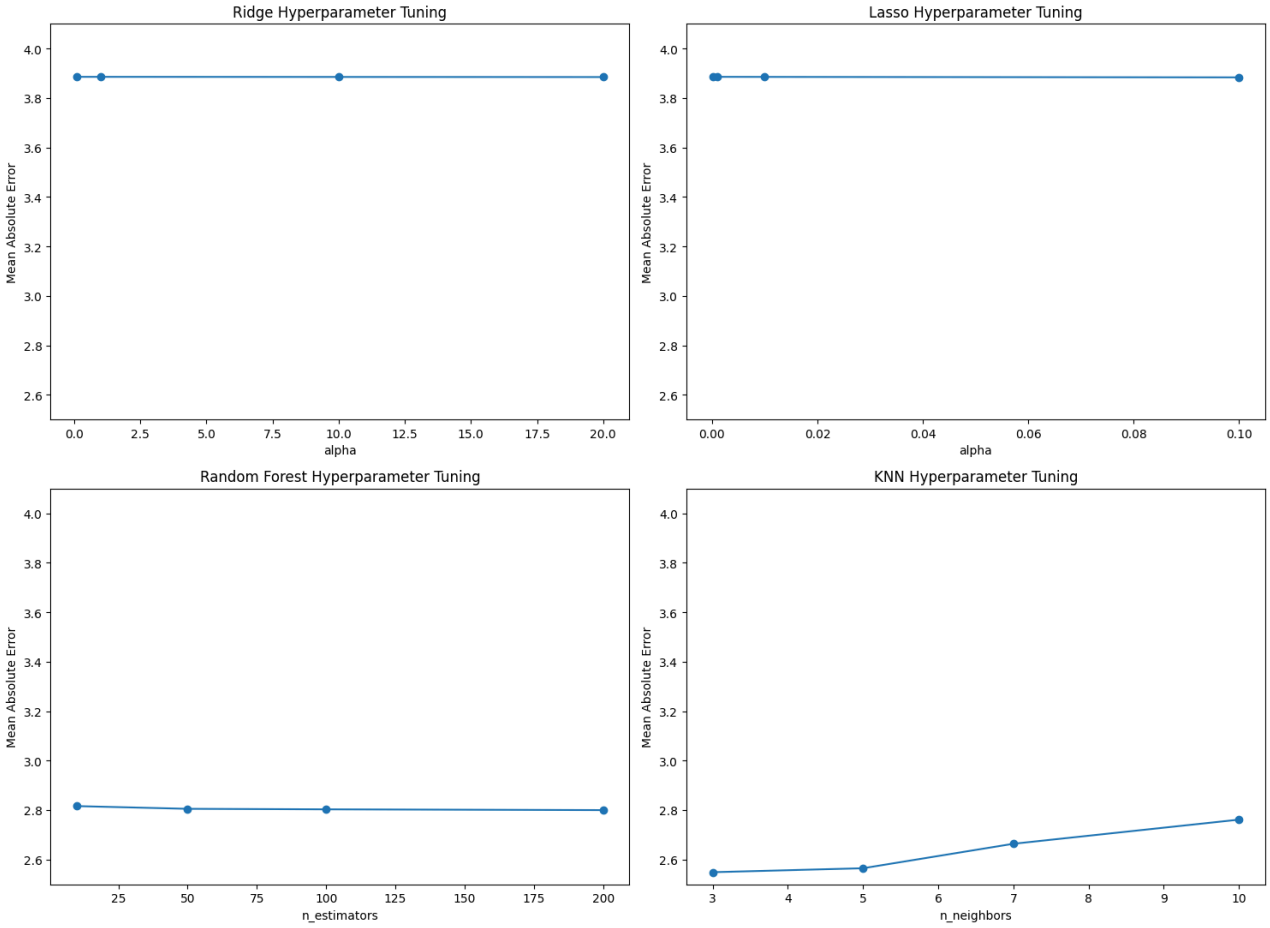
4.2 Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the performance of each model. The following hyperparameters were selected for tuning:

- Ridge Regression:
    - alpha: Regularization strength, tested values: [0.1, 1, 10, 20]

- Lasso Regression:
    - alpha: Regularization strength, tested values: [0.001, 0.01, 0.1, 1]
- Random Forest:
    - n_estimators: Number of trees in the forest, tested values: [10, 50, 100, 200]
- K-Nearest Neighbors (KNN):
    - n_neighbors: Number of neighbors to use, tested values: [3, 5, 7, 10]

GridSearchCV was employed to systematically search for the best hyperparameters. This method performs an exhaustive search over the specified parameter grid, evaluating the model using cross-validation to identify the hyperparameters that yield the best performance.

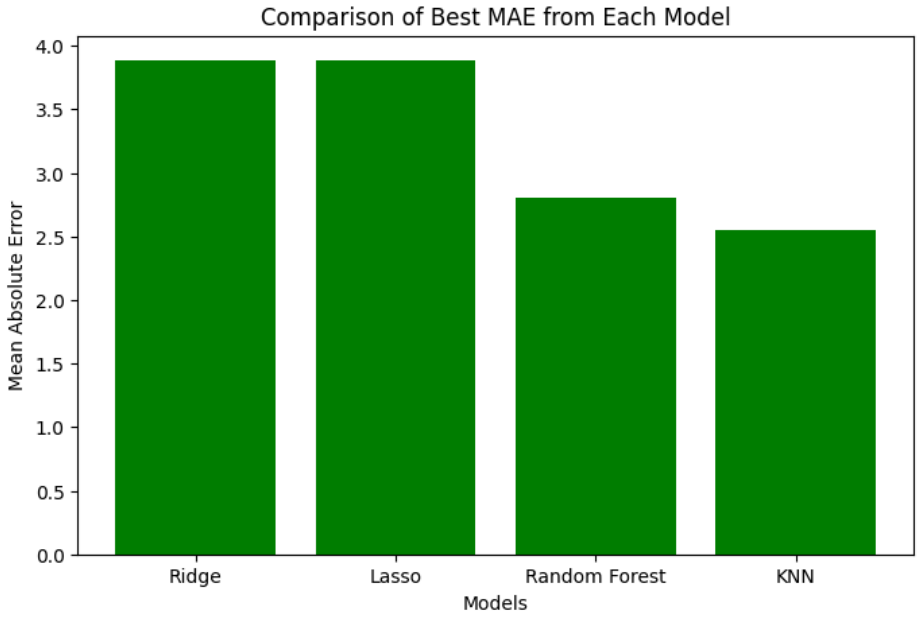


4.3 Performance Evaluation

The performance of each model was evaluated using the Mean Absolute Error (MAE), which measures the average magnitude of the errors in a set of predictions, without considering their direction. MAE is particularly useful in this study due to its straightforward interpretation and relevance to predicting grain sizes.

The results of the regression models are summarized below:

- Ridge Regression:
    - Minimum MAE: 3.88 with alpha: 20
- Lasso Regression:
    - Minimum MAE: 3.88 with alpha: 0.01
- Random Forest:
    - Minimum MAE: 2.8 with n_estimators: 200
- KNN:
    - Minimum MAE: 2.55 with n_neighbors: 3



The KNN model achieved the best performance with the lowest MAE, indicating its superior ability to capture the relationships in the data. KNN works by comparing a sample to its nearest neighbors in the feature space, making it highly effective for datasets where similar samples have similar responses.

The hyperparameters that provided the best results for each model are highlighted, demonstrating the effectiveness of the GridSearchCV optimization process. These results underscore the importance of model selection and hyperparameter tuning in predictive modeling tasks. The strong performance of KNN also highlights the value of non-parametric methods in capturing complex, non-linear patterns within the data.

5. Classification Modeling

5.1 Model Selection

In this study, three classification models were chosen to predict discrete grain size categories from the extracted features: Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

- Random Forest: This ensemble learning method constructs multiple decision trees and merges them to improve predictive performance and control overfitting. It is robust to noise and capable of handling large datasets with higher dimensionality.
- K-Nearest Neighbors (KNN): A non-parametric method that predicts the class of a sample based on the majority class among its k-nearest neighbors. It is simple and effective for small datasets but can be computationally intensive for large datasets.
- Support Vector Machine (SVM): A powerful classification algorithm that finds the optimal hyperplane separating different classes in the feature space. Using a radial basis function (RBF) kernel, SVM can handle non-linear decision boundaries, making it suitable for complex classification tasks.

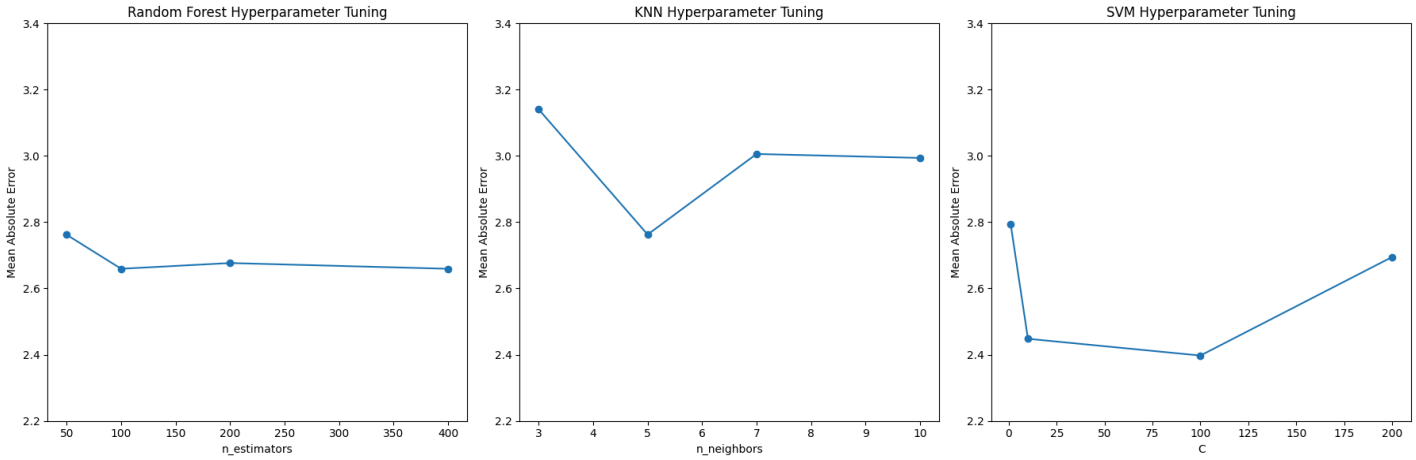
The selection of these models was based on their proven effectiveness in classification tasks, providing a mix of tree-based, instance-based, and kernel-based approaches to capture different aspects of the data.

5.2 Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the performance of each classification model. The following hyperparameters were selected for tuning:

- Random Forest:
    - n_estimators: Number of trees in the forest, tested values: [50, 100, 200, 400]
- K-Nearest Neighbors (KNN):
    - n_neighbors: Number of neighbors to use, tested values: [3, 5, 7, 10]
- Support Vector Machine (SVM):
    - C: Regularization parameter, tested values: [1, 10, 100, 100]

GridSearchCV was employed to systematically search for the best hyperparameters. This method performs an exhaustive search over the specified parameter grid, evaluating the model using cross-validation to identify the hyperparameters that yield the best performance.

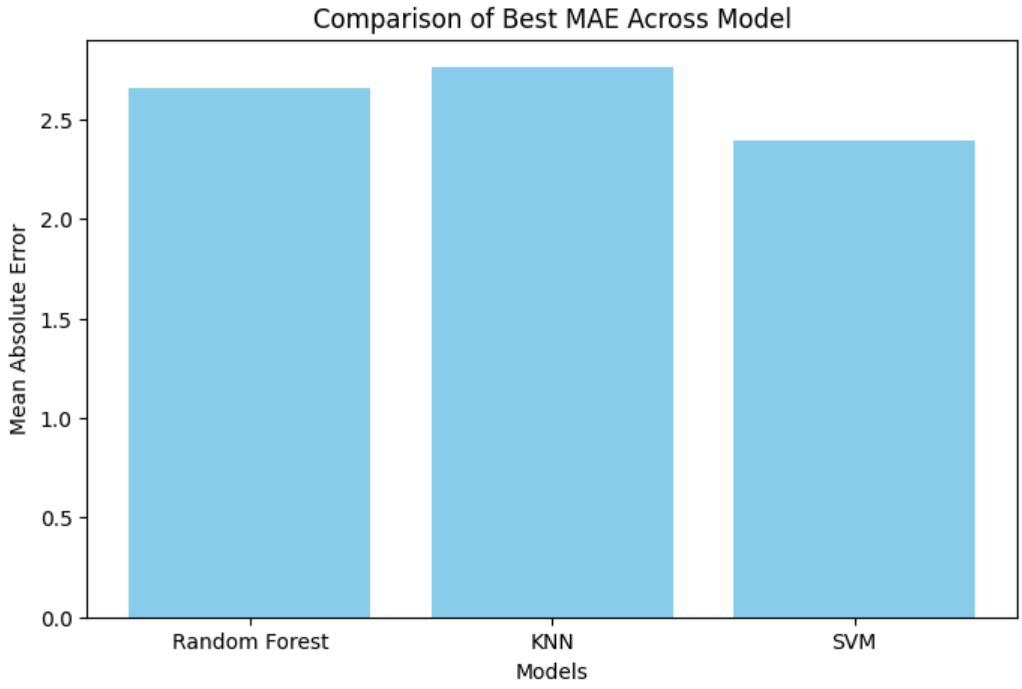


5.3 Performance Evaluation

The performance of each model was evaluated using the Mean Absolute Error (MAE), which measures the average magnitude of errors in a set of predictions without considering their direction. MAE is particularly useful in this study due to its straightforward interpretation and relevance to predicting grain sizes.

The results of the classification models are summarized below:

- Random Forest:
    - Minimum MAE: 2.66 with n_estimators: 400
- KNN:
    - Minimum MAE: 2.76 with n_neighbors: 5
- SVM:
    - Minimum MAE: 2.4 with C: 100



The Support Vector Machine (SVM) model achieved the best performance with the lowest MAE, indicating its strong capability in handling non-linear classification boundaries. SVM operates by finding the hyperplane that best separates the classes in the feature space, even in high-dimensional settings. The use of the RBF kernel in this project allowed the SVM to map the input features into a higher-dimensional space where a linear separation is possible, enhancing its ability to manage complex, non-linear relationships in the data. This result underscores the effectiveness of SVMs in classification tasks involving intricate and multi-faceted data patterns.

6. Results and Discussion

6.1 Performance Results

In this study, both regression and classification models were employed to predict and categorize grain sizes based on their orientations. The performance of these models was evaluated using the Mean Absolute Error (MAE) metric.

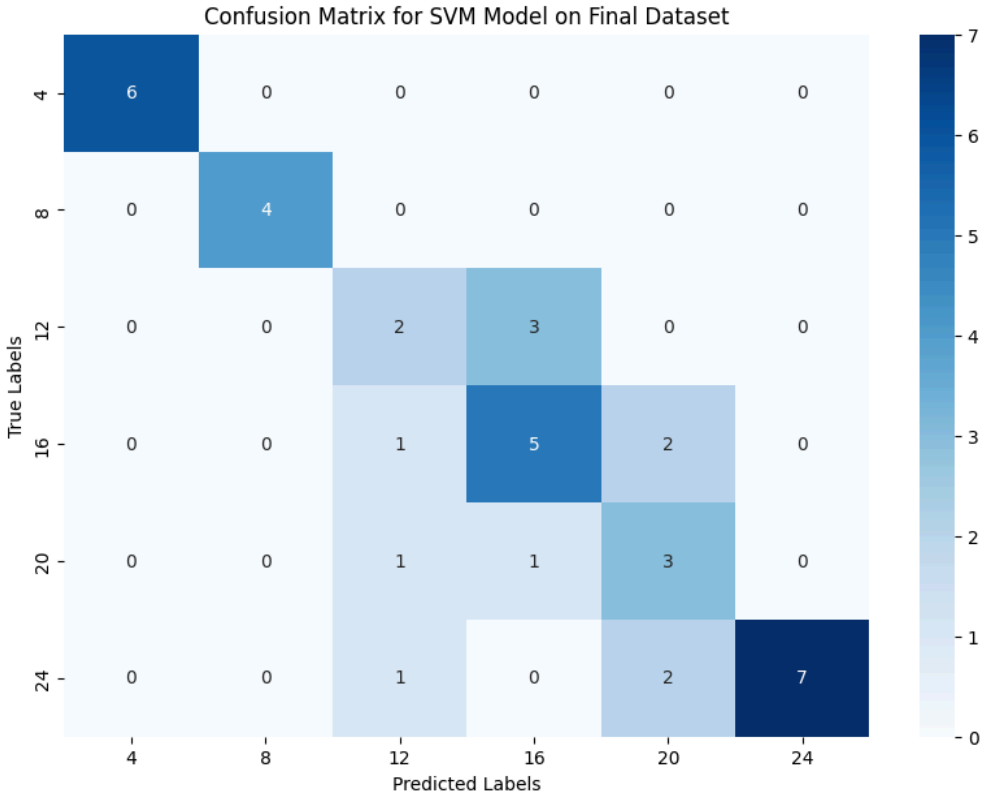
The regression models demonstrated varying degrees of success. The Ridge and Lasso regression models both achieved a minimum MAE of 3.88 with different alpha values, indicating their limited ability to capture the complex relationships in the data. The Random Forest regression model performed better, achieving a minimum MAE of 2.8 with 200 estimators, highlighting its capability to model non-linear relationships. However, the KNN regression model outperformed all others with a minimum MAE of 2.55 using 3 neighbors, demonstrating its strength in capturing local patterns in the data through distance-based predictions.

In comparison, the classification models showed superior performance overall. The Random Forest classification model achieved a minimum MAE of 2.66 with 400 estimators, indicating its robustness and flexibility in handling complex data. The KNN classification model followed closely with a minimum MAE of 2.76 using 5 neighbors, reaffirming the effectiveness of distance-based methods in this context. The Support Vector Machine (SVM) model outperformed all others, achieving the lowest MAE of 2.4 with a C value of 100. The SVM's ability to handle non-linear classification boundaries by finding the optimal hyperplane that separates the classes in a higher-dimensional space proved highly effective.

The classification models, particularly the SVM, outperformed the regression models in this study. This could be attributed to the nature of the problem: predicting discrete grain sizes may inherently align better with classification tasks than regression. The SVM's superior performance can be attributed to its use of the RBF kernel, which maps input

features into a higher-dimensional space, allowing for a more effective separation of classes even when the data is non-linear and complex. The kernel trick utilized by SVM enhances its ability to manage intricate patterns and relationships in the data, making it a powerful tool for this type of classification task.

These results underscore the importance of selecting appropriate models and fine-tuning their hyperparameters for predictive modeling tasks. The classification models, particularly the SVM, demonstrated a significant advantage in predicting grain sizes based on their orientations, providing valuable insights for the field of materials science. On the final test set of 38 samples, the best SVM model achieved an accuracy of .71 and a minimum MAE of 1.47 improving even on the validation set performance. The model performed much better on smaller grain sizes 4 and 6 as well as the largest size 24, than it did on grain sizes in the middle such as 12, 16 and 20.



6.3 Interpretation of MAE

The Mean Absolute Error (MAE) was chosen as the primary evaluation metric for its clear and interpretable measure of prediction accuracy. MAE calculates the average magnitude of errors between predicted and actual values, providing a straightforward understanding of model performance.

In the context of this study, MAE is particularly relevant due to its applicability to both regression and classification tasks. For regression, MAE quantifies how close the predicted grain sizes are to the actual sizes, offering insight into the model's precision. For classification, MAE measures the average discrepancy between the predicted and true grain size categories, making it a valuable metric for assessing classification accuracy.

Additionally, using MAE aligns with the goal of minimizing prediction errors in practical applications, where accurate grain size predictions are crucial for material properties and performance. The consistent use of MAE across different modeling approaches ensures a uniform basis for comparing model performance, enhancing the reliability and interpretability of the results.

7. Conclusion

7.1 Summary of Findings

This research focused on predicting grain sizes in polycrystalline materials using grain orientation data. By applying both regression and classification models, the study demonstrated that machine learning techniques could effectively model the relationship between grain orientations and sizes. Key findings include:

1. Regression Modeling: Among the regression models, the KNN regression outperformed other models, achieving the lowest MAE of 2.5492 with 3 neighbors. This indicates that KNN's ability to capture local patterns through distance-based predictions was particularly effective for this dataset. While the Random Forest regression also showed strong performance with an MAE of 2.8, it was slightly less effective than KNN in this context.
2. Classification Modeling: The SVM model demonstrated the best performance among classification models, achieving the lowest MAE of 2.4 with a C value of 100. This underscores SVM's superior capability to handle non-linear classification boundaries in high-dimensional spaces, thanks to its use of the RBF kernel. The Random Forest and KNN classification models also performed well, with MAEs of 2.66 and 2.76, respectively, further supporting the robustness of classification approaches for this problem.
3. Data Exploration and Feature Engineering: The use of two-point correlation statistics was crucial in capturing spatial relationships within the grain orientation data. However, PCA was not utilized in the final modeling due to its lack of improvement in predictive accuracy. The detailed exploration and feature engineering steps were vital in managing the high-dimensional data and improving the overall model performance.

These findings underscore the potential of advanced machine learning techniques in accurately predicting grain sizes from microstructural data. The successful application of SVM, in particular, highlights its value in handling complex, non-linear relationships, contributing valuable insights to the field of materials science and advancing the development of predictive models for material properties.

7.2 Future Work

Based on the findings and limitations of this study, several areas for future research are suggested:

1. Expand the Dataset: Future studies could include a larger and more diverse dataset to validate the generalizability of the models. Incorporating data from different types of polycrystalline materials could also provide broader insights.
2. Explore Additional Features: Investigating other microstructural features, such as grain boundary characteristics and texture, could provide a more comprehensive understanding of grain size prediction. Additionally, exploring features derived from higher-order statistics may further enhance model performance.
3. Refine Feature Engineering: Given the success of two-point correlation statistics, further refinement and exploration of spatial relationship metrics could improve predictive accuracy. This could involve developing new methods for quantifying grain orientation relationships.
4. Advanced Machine Learning Techniques: Exploring other advanced machine learning techniques, such as deep learning models, could potentially enhance prediction accuracy further. Techniques like convolutional neural networks (CNNs) might be particularly well-suited for capturing spatial patterns in 3D data.
5. Hybrid Models: Combining different machine learning approaches, such as hybrid models that integrate SVMs with other techniques like ensemble learning, could potentially leverage the strengths of multiple methods for improved accuracy and robustness.

These suggestions aim to build on the current research, addressing its limitations and exploring new avenues for enhancing the prediction and analysis of grain sizes in polycrystalline materials. By expanding the scope and depth of future studies, it is possible to advance the field further and develop more accurate and reliable predictive models.

8. References

1. 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 Patxi Fernandez-Zelaia, Yuksel C. Yabansu & Surya R. Kalidindi https://par.nsf.gov/servlets/purl/10147978
2. Microstructure informatics using higher-order statistics and efficient data-mining protocols https://link.springer.com/article/10.1007/s11837-011-0057-7