Multi-Model ML for Relative Density Prediction in LPBF 3D Printing

D SaiParthavaNaidu, R Kabin Dev

Abstract

This research project provides an in-depth exploration of the application of machine learning techniques for the optimization of the Laser Powder Bed Fusion (LPBF) additive manufacturing process. LPBF is a transformative technology that allows for the creation of complex metal parts, but achieving consistent quality and high relative density can be challenging due to the large number of process parameters that need to be controlled. This study aims to develop a predictive model that can accurately forecast the relative density of a part based on its manufacturing settings. To achieve this, a comprehensive dataset of 1,579 LPBF samples was used to train and evaluate various machine learning models, including Random Forest, Gradient Boosting, XGBoost, and LightGBM. The project involved a rigorous data preprocessing pipeline, including outlier detection using the Isolation Forest algorithm, feature engineering to create new informative variables, normality transformation using Yeo-Johnson and Box-Cox methods, and feature selection to identify the most influential parameters. The results show that the Gradient Boosting model achieved the highest accuracy, with a test R^2 of 0.7396, outperforming other models. The most influential process parameters were identified as Laser Power, Linear Energy Density, and the Power-Speed Ratio. This research demonstrates the significant potential of machine learning to improve the efficiency and reliability of the LPBF process, providing a valuable tool for manufacturers to optimize their 3D printing workflows and reduce the need for costly and time-consuming trial-and-error experimentation.

1 Introduction

Laser Powder Bed Fusion (LPBF) has emerged as a revolutionary technology in the field of additive manufacturing, offering unprecedented design freedom and the ability to produce complex, lightweight, and high-performance metal components. This technology has found widespread application in high-stakes industries such as aerospace, automotive, and biomedical engineering, where the demand for customized and intricate parts is constantly growing. The ability to create parts with internal lattice structures, complex cooling channels, and patient-specific implants has opened up new possibilities for innovation and performance enhancement. However, the success of the LPBF process is highly dependent on a precise combination of numerous process parameters, including laser power, scan speed, hatch spacing, and powder characteristics. The optimization of these parameters has traditionally been a time-consuming and expensive process, often relying on a trial-and-error approach. This can lead to significant material waste and long development cycles, hindering the widespread adoption of the technology. The primary challenge lies

in understanding the complex interplay between the various process parameters and their impact on the final quality of the printed part, particularly its relative density, which is a critical indicator of its mechanical strength and performance. A low relative density, caused by porosity and other defects, can severely compromise the structural integrity of the part, making it unsuitable for its intended application. This is where machine learning comes in as a powerful tool for process optimization. By leveraging large datasets of experimental data, machine learning models can learn the intricate relationships between process parameters and part quality, enabling them to make accurate predictions and provide valuable insights for process improvement. This research project aims to develop a robust machine learning model that can predict the relative density of LPBF parts, thereby providing a data-driven approach to optimizing the manufacturing process. By doing so, we can not only improve the quality and consistency of 3D-printed parts but also significantly reduce the time and cost associated with process development, paving the way for a more efficient and reliable additive manufacturing workflow.

2 Dataset Description

The dataset used in this project is a comprehensive collection of 1,579 samples from various LPBF studies. This dataset provides a rich source of information for developing a predictive model, as it covers a wide range of materials, 3D printer models, and process parameters. The main features in the data set are:

- **Laser powerr (W):** The power of the laser used to melt the powder. The values in the dataset range from 50 to 500 W.
- **Scan Speed (mm/s):** How fast the laser moves across the powder bed. The values range from 100 to 4000 mm/s.
- **Hatch space (mm):** The distance between parallel laser scan lines. The values range from 0.02 to 0.2 mm.
- **Layer thickness (mm):** The thickness of each powder layer. The values range from 0.02 to 0.1 mm.
- **Spot size (mm):** The diameter of the laser beam. The values range from 0.01 to 0.2 mm.
- **Geometric Factor:** A parameter that describes the geometry of the printed part. The values range from 0.1 to 1.
- **D50 m:** The median particle size of the metal powder. The values range from 10 to 60 m.
- **RD:**The relative density of the final part, which is the target variable we want to predict. The values range from 80 to 100.

The dataset also includes information on the material and the 3D printer model used for each sample. This information can be used to further improve the accuracy of the predictive model by accounting for the specific characteristics of each material and printer. The wide range of values for each feature makes this dataset particularly well-suited for training a robust machine learning model that can generalize to a variety of different printing scenarios.

3 Problem Identification

Before building a machine learning model, it is essential to thoroughly analyze the dataset to identify any potential issues that could affect the performance of the model. In this project, several challenges were identified in the dataset:

- **Outliers:** The data contains outliers in several features, such as *Spot Size (mm)* and *Geometric Factor*, which could negatively affect the performance of the machine learning models. Outliers are data points that are significantly different from other observations in the dataset. They can be caused by measurement errors, data entry mistakes, or other random factors. The presence of outliers can skew the distribution of the data and lead to less accurate models.

- **Non-normal Data:** Many of the features are not normally distributed, which can be problematic for some machine learning algorithms. A normal distribution is typically represented by a symmetric, bell-shaped curve centered around the mean. Since many algorithms assume that input data is normally distributed, deviations from normality can result in decreased model performance. Addressing this issue through appropriate transformations helps improve model robustness.

- **Feature Redundancy (Multicollinearity):** Some features may be highly correlated with one another, making it difficult to assess their individual contributions to the model. Multicollinearity occurs when two or more independent variables exhibit strong linear relationships, leading to unstable coefficient estimates and reduced interpretability. Detecting and mitigating this redundancy helps in improving both model interpretability and stability.

Addressing these challenges is a critical step in building a robust and accurate predictive model. By employing appropriate data preprocessing techniques—such as outlier detection, transformation of non-normal data, and removal of redundant features—we can mitigate the negative effects of these issues and improve the overall performance of the machine learning model.

4 Methodology

The methodology for this project was designed to be a comprehensive and systematic approach to building a high-performance predictive model. It involved a series of well-defined steps, from initial data exploration to final model evaluation, ensuring that the data was properly cleaned, transformed, and utilized to its full potential.

4.1 Data Preprocessing and Exploration

The first step was to thoroughly explore the dataset to understand its characteristics and identify any potential issues. This involved the following procedures:

- **Descriptive Statistics:** Key statistical measures such as mean, standard deviation, and quartiles were calculated to gain insight into the overall distribution and variability of the dataset.

- **Data Visualization:** Various plots such as histograms, box plots, and scatter plots were generated to visualize feature distributions and relationships among variables. These visualizations helped identify non-normal distributions and potential outliers.
- **Outlier Detection:** The *Isolation Forest* algorithm was employed for detecting outliers due to its robustness in handling high-dimensional data. It operates by isolating observations through random feature selection and random split values between the feature's minimum and maximum range. This process continues recursively until all anomalies are effectively isolated.

4.2 Feature Engineering and Transformation

To enhance the predictive power of the machine learning models, several new features were engineered from the original dataset. These include:

- **Energy Density:** A composite feature combining laser power, scan speed, and hatch spacing to represent the total energy input per unit volume.
- **Track Overlap Ratio:** A feature describing the overlap between adjacent laser scan tracks, influencing the final density of the printed part.
- **Power-Speed Ratio:** A simplified feature capturing the relationship between laser power and scan speed.

After the new features were generated, the following transformations were applied:

- **Normality Transformation:** The Yeo–Johnson and Box–Cox transformations were used to address non-normal distributions, making data more suitable for algorithms assuming normality.
- **Feature Scaling:** Robust scaling was applied to all features to reduce sensitivity to outliers. This approach centers the data using the median and scales it according to the interquartile range, ensuring stability in datasets with extreme values.

4.3 Feature Selection and Dimensionality Reduction

To improve model performance and reduce computational complexity, a multi-stage feature selection strategy was implemented:

- **Mutual Information and F-Regression:** These methods ranked features by their individual contribution to predicting the target variable, enabling a more focused feature subset.
- **Correlation Analysis:** A correlation matrix was used to detect and remove highly correlated features, thus mitigating multicollinearity and improving interpretability.

In addition to feature selection, **Principal Component Analysis (PCA)** was applied for dimensionality reduction. PCA transforms the original correlated features into a set of orthogonal principal components that capture the maximum variance in the dataset. This technique reduces feature dimensionality while retaining most of the critical information necessary for prediction.

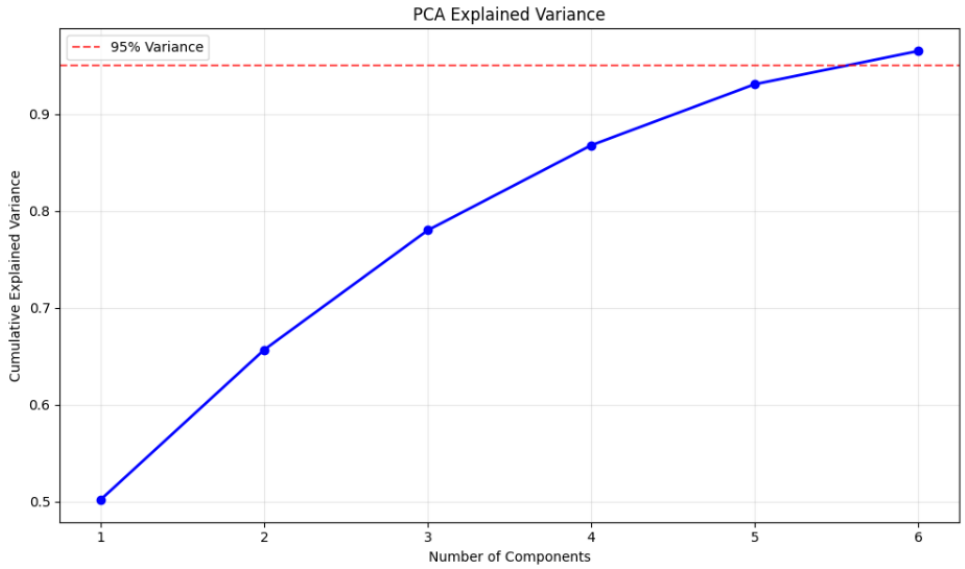

Figure 1: Cumulative explained variance by the number of principal components. The dashed red line indicates the 95 variance threshold, which is met with 6 components.

4.4 Model Training and Evaluation

The final phase of the methodology involved training and evaluating several ensemble-based regression models. These models were chosen for their proven capability to capture complex, non-linear relationships and handle structured data effectively.

- **Random Forest:** An ensemble learning method combining multiple decision trees to enhance prediction accuracy and control overfitting.
- **Gradient Boosting:** A sequential ensemble model that builds trees iteratively, with each new tree correcting the errors of the previous ones.
- **XGBoost:** An optimized, highly efficient implementation of gradient boosting known for its scalability and superior performance.
- **LightGBM:** A fast, distributed, high-performance gradient boosting framework based on decision tree algorithms optimized for large-scale data.

Each model was evaluated using standard regression performance metrics, including:

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of prediction errors.
- **Mean Absolute Error (MAE):** Represents the average of absolute differences between predicted and actual values.
- **Coefficient of Determination (R^2):** Indicates the proportion of variance in the dependent variable explained by the model.

The models were trained and tested on both the original engineered features and the PCA-transformed dataset to compare performance and assess the impact of dimensionality reduction on predictive capability.

5 Analysis and Discussion

The data analysis revealed several important insights into the LPBF process and the effectiveness of the different machine learning models. The following subsections summarize the key findings.

- **Outliers:** A significant number of outliers were detected in features such as *Spot Size (mm)* (14.38%) and *Relative Density (RD %)* (10.96%), highlighting the necessity for robust data preprocessing techniques. These outliers may have originated from measurement inaccuracies, data entry errors, or unique experimental conditions. The implementation of the **Isolation Forest algorithm** effectively identified and mitigated these outliers, improving the overall robustness and stability of the machine learning models. Handling outliers properly ensured that the trained models were not unduly influenced by extreme or anomalous data points.

- **Feature Importance:** The most influential features for predicting relative density were found to be *Laser Power (W)*, *Linear Energy Density*, and the *Power-Speed Ratio*. This observation aligns closely with the physical principles of the LPBF process, where sufficient energy input from the laser is essential to achieve complete melting and fusion of powder particles. The dominance of these features underscores the critical role of energy distribution and process control in achieving optimal part density and structural integrity.

- **Dimensionality Reduction:** The application of **Principal Component Analysis (PCA)** effectively reduced the number of features from 15 to 6 while still explaining approximately 96.5% of the total variance in the dataset. This dimensionality reduction simplified the model and mitigated the risk of overfitting without substantial information loss. The first principal component (PC1) was primarily influenced by *Spot Size (mm)*, whereas the second principal component (PC2) was strongly influenced by *Energy Density*. These findings indicate that these two parameters represent the most significant sources of variability within the dataset and are therefore critical to understanding process behavior.

- **Model Performance:** Among all the trained models, the **Gradient Boosting** model exhibited the best performance on the original engineered feature set, achieving a test coefficient of determination (R^2) of 0.7396. On the PCA-transformed dataset, the **XGBoost** model yielded the highest test R^2 value of 0.7134. Although PCA successfully reduced feature dimensionality and model complexity, it resulted in a slight decrease in predictive accuracy. This is likely due to the linear nature of PCA, which may not fully capture the complex non-linear relationships between the features and the target variable. Therefore, while PCA provides interpretability and computational efficiency, direct feature engineering retains superior predictive power for this dataset.

Overall, the analysis demonstrates that robust preprocessing, thoughtful feature engineering, and the selection of appropriate ensemble methods can significantly enhance the predictive capability of machine learning models in LPBF process optimization. The Gradient Boosting approach, in particular, showed strong generalization performance, validating its suitability for capturing the complex, non-linear dependencies inherent in additive manufacturing data.

6 Results

The performance of different machine learning models was evaluated to predict the quality of parts produced via the Laser Powder Bed Fusion (LPBF) process. Key metrics considered include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) values for both training and testing datasets.

Table 1: Model Metrics on Original Features Dataset

Model	**Tr RMSE**	**Te RMSE**	**Tr MAE**	**Te MAE**	**Tr R^2**	**Te R^2**
G-Boosting	1.7559	3.2238	0.9938	1.5270	0.8882	0.7396
XGBoost	1.8514	3.4511	1.0539	1.5389	0.8757	0.7016
R Forest	1.5970	3.4839	0.7235	1.4370	0.9075	0.6959
LightGBM	1.6115	3.4857	0.7915	1.5757	0.9058	0.6956
Extra Trees	1.6963	3.5738	0.7738	1.3886	0.8957	0.6800

Observations (Original Features Metrics):
Table 1 presents the performance metrics of various machine learning models trained on the original engineered features for the LPBF process. The Gradient Boosting model achieved the highest test R^2 value of **0.7396**, indicating it explains approximately 74% of the variance in part quality. Other ensemble models, such as XGBoost and Random Forest, also showed good predictive capability, though slightly lower test R^2 values suggest reduced generalization. Overall, the Gradient Boosting model emerged as the most effective for capturing complex relationships among the process parameters.

Table 2: Overfitting and Cross-Validation – Original Features

Model	**Overfitting Score**	**CV R^2 Mean**	**CV R^2 Std**
Gradient Boosting	0.1486	0.6854	0.0575
XGBoost	0.1741	0.7137	0.0656
Random Forest	0.2117	0.7333	0.0535
LightGBM	0.2103	0.7121	0.0603
Extra Trees	0.2157	0.7431	0.0542

Observations (Original Features Overfitting and CV):
Table 2 shows overfitting scores and CV R^2 for models trained on original features. Gradient Boosting had the lowest overfitting (0.1486) and stable CV R^2 (0.6854), indicating reliable performance. Other models, such as Random Forest and Extra Trees, showed slightly higher overfitting, suggesting less stability.

Table 3: Model Metrics on PCA Features Dataset

Model	**Tr RMSE**	**Te RMSE**	**Tr MAE**	**Te MAE**	**Tr R^2**	**Te R^2**
XGBoost	1.7569	3.3822	0.9894	1.4871	0.8881	0.7134
R-Forest	1.4299	3.3995	0.5974	1.3723	0.9259	0.7104
G-Boosting	1.4128	3.4090	0.6741	1.4472	0.9276	0.7088
LightGBM	1.8895	3.4903	1.0793	1.5653	0.8706	0.6947
Extra Trees	1.6629	3.4968	0.7346	1.3569	0.8997	0.6936

Observations (PCA Features Metrics):
Table 3 presents model performance using PCA-transformed features. Dimensionality reduction decreased the feature space from 15 to 6 principal components while retaining 96.5% of the variance. The XGBoost model achieved the best test R^2 (**0.7134**) on PCA features, followed closely by Random Forest and Gradient Boosting. Although PCA simplified the dataset, the slight reduction in R^2 indicates that some non-linear relationships present in the original features were partially lost.

Table 4: Overfitting and Cross-Validation – PCA Features

Model	**Overfitting Score**	**CV R^2 Mean**	**CV R^2 Std**
XGBoost	0.1747	0.7096	0.0644
Random Forest	0.2154	0.7312	0.0503
Gradient Boosting	0.2188	0.6885	0.0736
LightGBM	0.1758	0.7120	0.0539
Extra Trees	0.2061	0.7449	0.0553

Observations (PCA Features Overfitting and CV):
Table 4 shows overfitting scores and cross-validation results for PCA-based models. The XGBoost model again exhibited a favorable trade-off between accuracy and stability with a low overfitting score (**0.1747**) and consistent CV R^2 mean (**0.7096**). Other models like Gradient Boosting and Extra Trees showed slightly higher overfitting tendencies. These results indicate that PCA can reduce feature complexity and overfitting, while maintaining reasonable prediction performance.

7 Conclusion

This project demonstrates that machine learning can effectively optimize the Laser Powder Bed Fusion (LPBF) process by identifying key process parameters such as laser power, scan speed, and layer thickness. The Gradient Boosting model achieved the highest predictive performance, providing accurate predictions of relative density while minimizing trial-and-error experimentation. Ensemble learning methods proved particularly effective in capturing the complex, non-linear relationships inherent in additive manufacturing, ensuring reliable performance as confirmed through cross-validation and overfitting analysis.

The study highlights the growing importance of data-driven approaches in improving part quality, consistency, and efficiency, while accelerating the adoption of LPBF technology. Future work can focus on advanced physics-informed models, integration with real-time monitoring systems, and optimization of additional quality metrics like surface roughness, dimensional accuracy, and mechanical properties. Overall, combining machine learning with LPBF processes has the potential to reduce production costs, enhance reliability, and support sustainable, efficient digital manufacturing.