(IJRST) 2025, Vol. No. 15, Issue No. 4, Oct-Dec **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

Material Lifespan Predictor

Kisna Patel
Grade 12, National Senior Certificate (NSC) Candidate
Worcester Gymnasium High School
Worcester, South Africa

DOI:10.37648/ijrst.v15i04.004

[1] *Received: 18/09/2025; Accepted: 11/10/2025; Published: 25/10/2025*

Abstract

Predicting the operational lifespan of industrial components remains a critical challenge in manufacturing, where early failure leads to costly downtime and waste. This study addresses that problem by applying machine learning to estimate the lifespan of materials based on manufacturing process parameters and alloy composition. The dataset comprises 1,000 samples containing attributes such as component type, microstructure, cooling and heat-treatment parameters, alloy percentages, and defect counts. Seven regression models were developed and compared—Linear Regression, Polynomial Regression, Random Forest, XGBoost, Support Vector Regression, CatBoost, and LightGBM—using standardized numeric features and one-hot encoded categorical variables. Model performance was evaluated through mean squared error (MSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). Tree-based ensemble methods achieved superior results, with LightGBM delivering the best performance (MSE = 3757.36, MAPE = 4.13%, R^2 = 0.964). SHAP explainability analysis revealed that cooling rate, alloy composition, and defect counts were the most influential features. These findings demonstrate that gradient boosting ensembles, combined with explainability techniques, can provide accurate and interpretable predictions for material lifespan optimization in manufacturing environments.

Keywords: *Machine Learning; Material Lifespan; Regression Models; SHAP; Manufacturing*

I. Introduction

Accurate prediction of component lifespan is essential for manufacturing reliability, maintenance scheduling, and cost optimization. Traditional lifecycle estimates are often conservative and based on limited empirical rules. Machine learning provides a data driven approach to estimate lifespan from many interacting process variables. This study aims to predict component lifespan in hours using manufacturing parameters and alloy composition. The work trains multiple regression models and applies explainability analysis to identify which process and material features most strongly influence predicted life. The goal is both accurate prediction and interpretable insight to guide manufacturing improvements.

II. Related Work

Several studies have applied machine learning to predict material properties and performance under various manufacturing conditions. Prior research has focused on corrosion prediction, fatigue life estimation, and process optimization using both statistical and modern machine learning approaches.

A. Predicting Material Fatigue and Degradation The authors in applied machine learning methods to estimate fatigue life in metals based on microstructural and environmental variables. Their study employed linear regression and support vector regression to model stress– strain relationships and demonstrated that nonlinear methods can

[1] How to cite the article: Patel K. (October, 2025); Material Lifespan Predictor; *International Journal of Research in Science and Technology*; Vol 15, Issue 4; 31-41, DOI: http://doi.org/10.37648/ijrst.v15i04.004

significantly outperform traditional empirical equations. However, this work was limited to a single model type and did not include explainability analysis.

Additionally, a comparative study on **multi-region real estate price prediction [1]** demonstrated how combining regression and ensemble learning methods with SHAP explainability can uncover the most important predictive features across datasets. Drawing methodological inspiration from that approach, the present work extends similar modelling and interpretability techniques to a new domain, predicting material lifespan from process and alloy data in manufacturing.

B. Using Ensemble Learning in Manufacturing Prediction

A separate study introduced the use of Random Forest and Gradient Boosting algorithms to predict material strength and hardness based on thermal and mechanical parameters. These ensemble models achieved high predictive accuracy but lacked transparency regarding feature contributions. Inspired by this, our work incorporates multiple ensemble models, including XGBoost [2], CatBoost [3], and LightGBM [4], and complements them with SHAP explainability [5] to interpret feature influence.

C. Integrating Explainable AI in Material Science The work of explored explainable artificial intelligence for process optimization in manufacturing, employing SHAP and LIME to identify the most critical process variables. Although their focus was processing control rather than lifespan prediction, their methodology validated the importance of interpretability in industrial applications.

Building on these studies, our research contributes a comparative evaluation of seven regression algorithms trained on a unified dataset describing component manufacturing and metallurgical parameters. Unlike previous works, we focus on both prediction accuracy and model interpretability, using SHAP to identify how variables such as cooling rate, alloy composition, and defect counts affect predicted lifespan.

III. Implementation

A. The Dataset

Material Lifespan Prediction Dataset [6] Attributes:

Feature	Description
PredictedHours	Target variable, lifespan in hours
ComponentType	Type of component
StructureType	Microstructural grain configuration
CoolRate	Cooling rate during manufacture
QuenchDuration	Quenching duration in seconds
ForgeDuration	Forging duration in seconds
HeatProcessTime	Heat treatment time in minutes
NickelComposition	Nickel percentage
IronComposition	Iron percentage
CobaltComposition	Cobalt percentage
ChromiumComposition	Chromium percentage
MinorDefects	Count of minor defects
MajorDefects	Count of major defects

EdgeDefects	Count of edge defects
InitialPosition	Initial component position
FormationMethod	Casting/formation method

Table 1

Correlation matrix: A correlation matrix is a table displaying the correlation coefficients for every pair of variables in a dataset. These coefficients always fall within the range of -1 and 1. A strong correlation exists when the coefficient is close to either positive one, indicating a positive relationship where variables move in the same direction, or negative one, indicating a negative relationship where variables move in opposite directions. a coefficient approaching zero indicates a weak correlation.

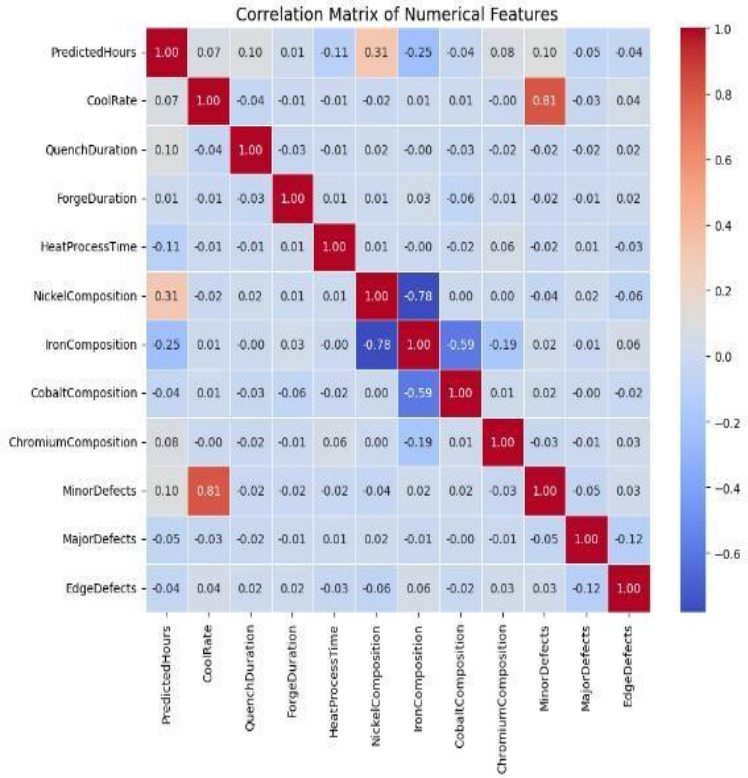


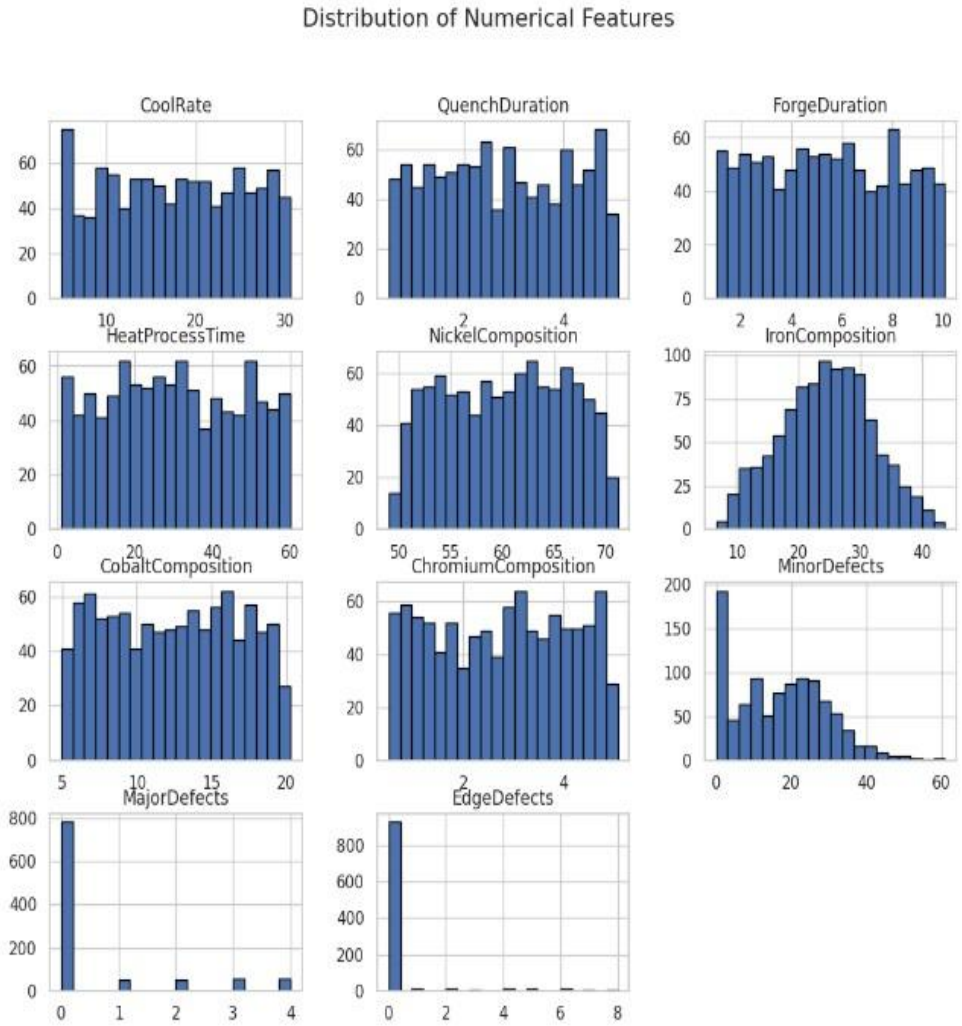
Figure 1

The correlation matrix in *Figure 1* illustrates the relationships between numerical features in the dataset.
Nickel Composition showed a moderate positive correlation (0.31) with PredictedHours, indicating that higher nickel content generally leads to longer material lifespan. Conversely, IronComposition displayed a weak negative correlation (–0.25), suggesting that excessive iron content may slightly reduce durability.

Distribution of numerical features:

Figure 13 shows the distributions of the numerical input features used to predict component lifespan. These histograms summarize the range, central tendency, and skewness for each continuous predictor and highlight count distributions for the defect features.

Figure 2



The distributions reveal several important characteristics of the data that guided our preprocessing and modeling choices. First, many continuous process variables such as CoolRate, QuenchDuration, ForgeDuration and HeatProcessTime span distinct ranges and do not all follow the same distributional form; some appear fairly uniform while others show mild skew. Composition features such as Nickel Composition and ChromiumComposition concentrate within a limited band, indicating most samples share similar alloy ratios. Notably, the defect variables are highly zero-inflated: MinorDefects has a long right tail with many small counts, whereas MajorDefects and EdgeDefects are dominated by zeros. This zero-inflation indicates that defect counts are sparse events and may require special handling.

B. Data Refining and Standardising The dataset was first inspected for completeness and basic consistency. No missing values were detected across the 1,000 samples, and duplicate rows were absent. Numeric features were standardized to zero mean and unit variance using ***StandardScaler*** to ensure comparability across measurements. Categorical features were encoded with one-hot encoding using ***handle_unknown='ignore'*** so the preprocessing remains robust to unseen categories in the test set. A random 80/20 train/test split with a fixed seed was used to ensure reproducible evaluation. Additionally, we inspected the target distribution and considered logtransforming highly skewed targets where appropriate to stabilise variance; MAPE was computed only when actual values were non-zero to avoid division by zero issues.

C. Model Training

Seven regression models were trained using identical preprocessing pipelines to ensure fair comparison. Models used in this study were:

- Linear Regression
- Polynomial Regression (degree 2)
- Random Forest Regressor
- XGBoost Regressor
- Support Vector Regression
- CatBoost Regressor
- LightGBM Regressor

No hyperparameters were tuned for this study. Models were trained using default (library) settings unless otherwise noted. For reproducibility, the preprocessing and model pipelines were encapsulated using scikit-learn Pipelines [6].

IV. Experiments and Results

A. Evaluation metrics and methodology

Models were evaluated on this test using:

- Mean Squared Error (MSE) [8]
- Mean Absolute Percentage Error (MAPE)
- R squared (R^2)

Each model was trained on the training set and scored on the hold-out test set. The final numeric results are summarized in *Table 2*

Model	MSE	MAPE (%)	R^2
Linear Regression	91234.997	21.697	0.119
SVR	36668.623	12.424	0.646
Polynomial Regression	30087.187	11.867	0.709
Random Forest	7647.216	6.163618	0.926
XGBoost	5754.633	5.155	0.944
LightGBM	3757.36	4.13	0.964
Catboost	2729.909	3.338	0.973

Table 2

Notes: SVR underperformed relative to tree based ensembles on this data

B. Visualisation & Eplainability

Figure 2. Linear Regression Model

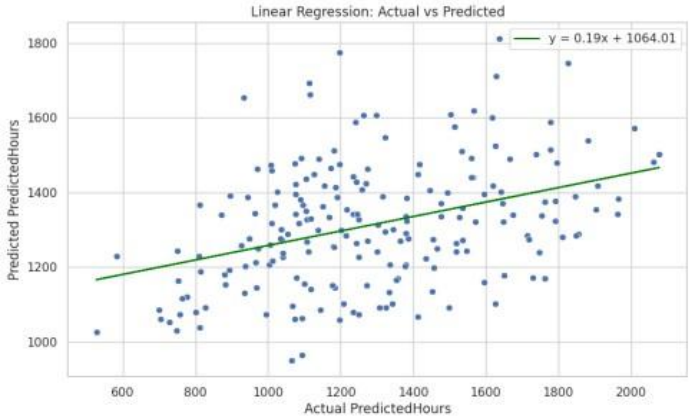


Figure 3. Support Vector Regression Model

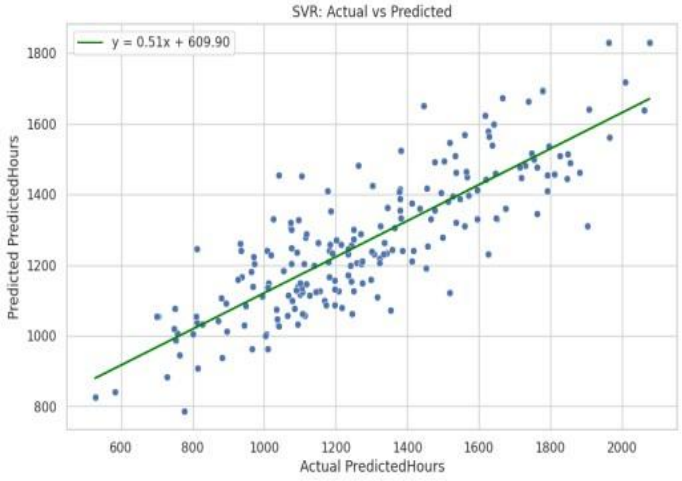


Figure 4. Polynomial Regression Model

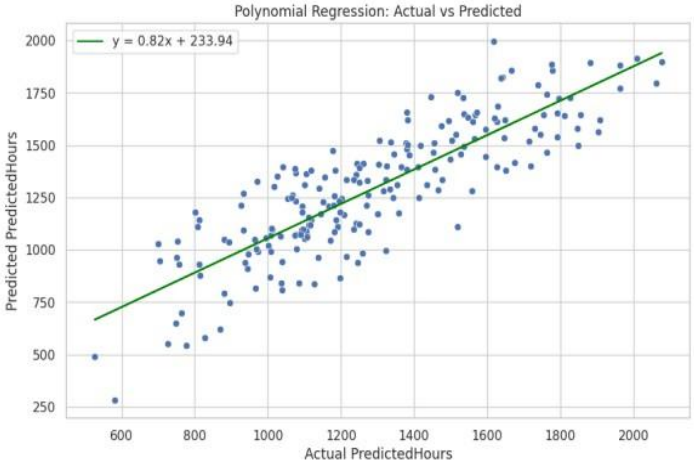


Figure 5. Random Forest Model

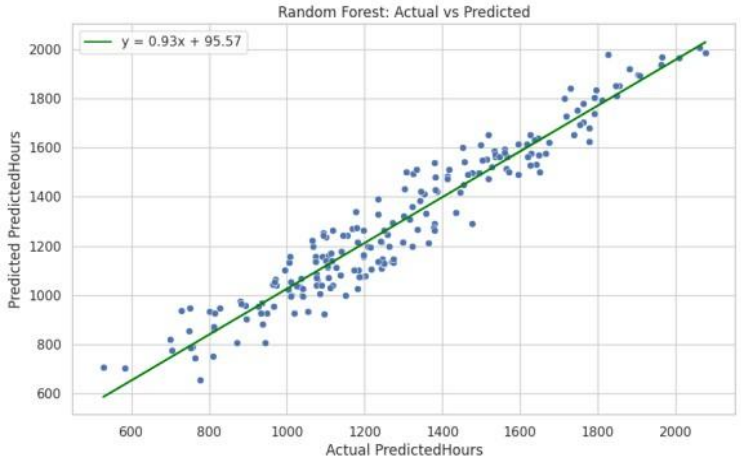


Figure 6. XGBoost Model

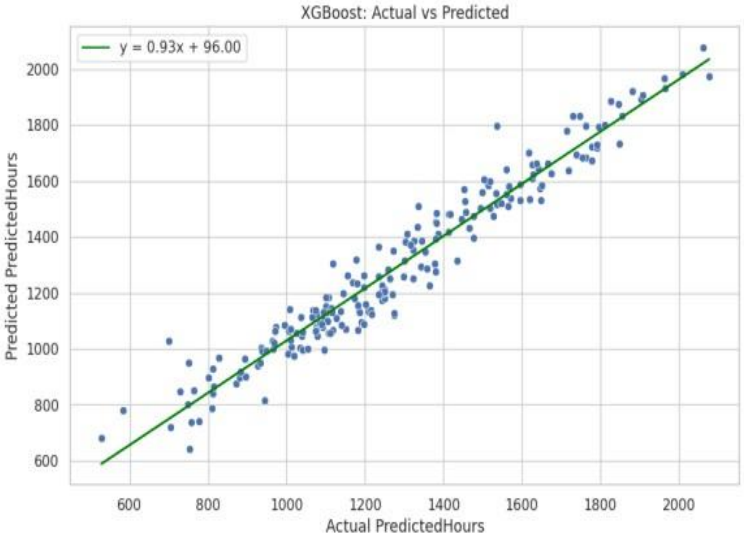


Figure 7: LightGBM Model

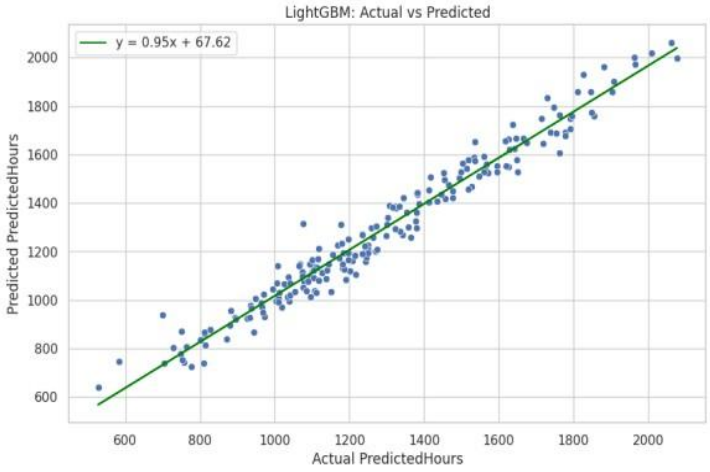
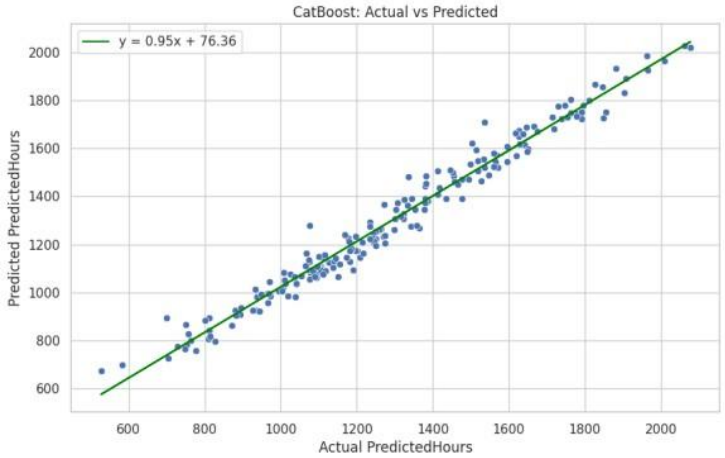


Figure 8. CatBoost Model



These results clearly demonstrated the superiority of tree-based ensemble methods (*Figure 5,6,7 & 8*) for this task compared to non tree based ensemble methods (Figure 2,3 & 4)

The CatBoost Regressor (Figure 8) emerged as the topperforming model, achieving the lowest Mean Squared Error (MSE) and the highest coefficient of determination (R^2) at 0.974. It also recorded the lowest mean percentage prediction error (MAPE).

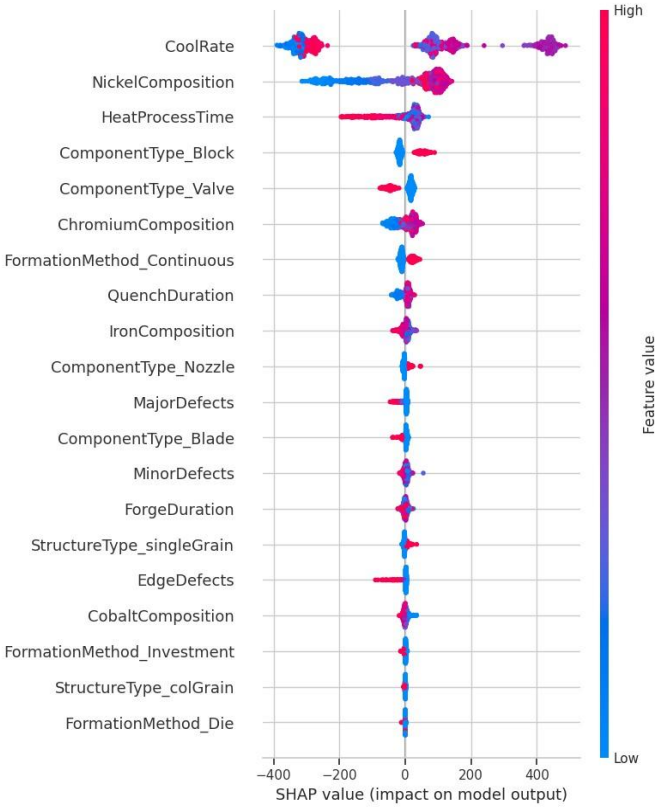
SHAP Analysis:

For interpretability, SHAP summary plots were generated for the leading tree-based models.

SHAP Analysis:

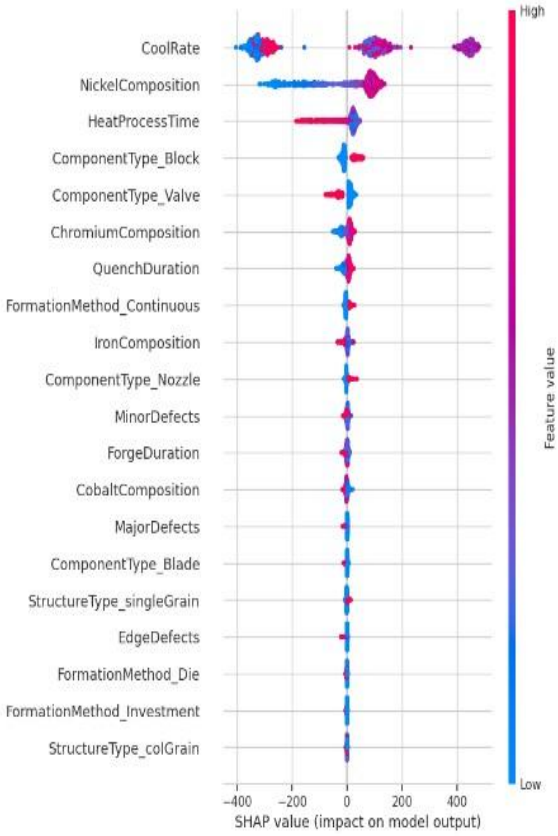
Random forest:

Figure 9: Random Forest SHAP Summary Plot



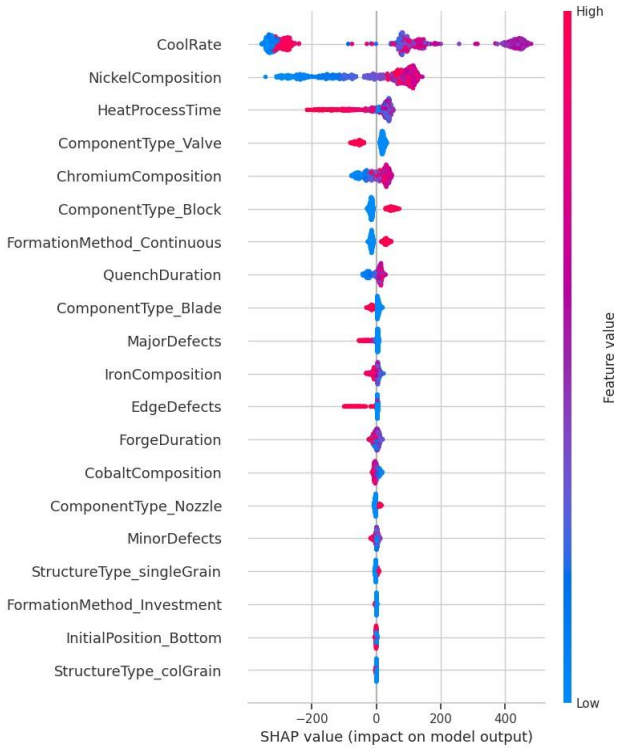
XGBoost:

Figure 10: XGBoost SHAP Summary Plot



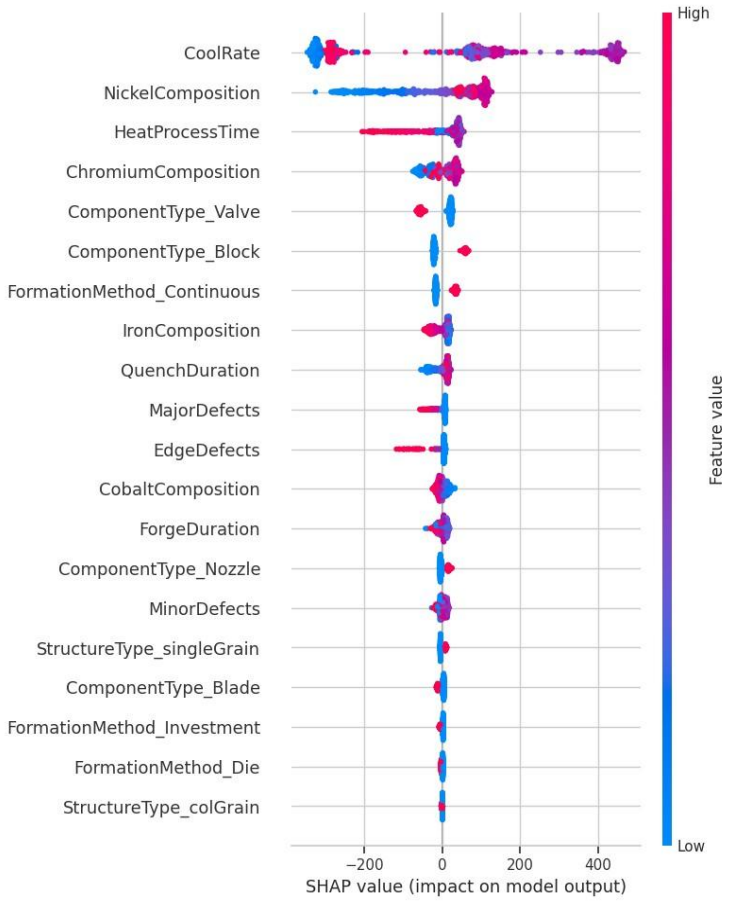
LightGBM:

Figure 11: LightGBM SHAP Summary Plot



CatBoost:

Figure 12: CatBoost SHAP Summary Plot



Key finding:

The SHAP explainers consistently ranked cooling related features, nickel and chromium composition, and defect counts among the top contributors to predicted lifespan. Place the SHAP figure and then a brief paragraph interpreting the plot, e.g., high cooling rates decrease predicted life, or increased major defects reduce predicted life, etc., based on the observed SHAP sign and magnitude.

V. Conclusion

This study demonstrates that machine learning models can accurately estimate component lifespan using manufacturing process variables and alloy composition. Gradient boosting ensembles, specifically LightGBM and CatBoost, delivered the most accurate predictions in this analysis. LightGBM produced the strongest numeric performance and, together with SHAP explainability, identified cooling rate, alloy composition, and defect counts as the principal drivers of predicted lifespan. Linear and polynomial regression provided baseline performance and interpretability, while SVR did not generalize well on this dataset under default settings. The combination of high predictive accuracy and explainability makes these models valuable for data driven process improvement and maintenance planning.

These models are not just predictive engines but are foundational for data-driven process improvement, enabling manufacturers to optimize specific production parameters (like cooling rate) to maximize longevity and integrate predictive lifespan estimates directly into maintenance and reliability planning.

VI. Future Work

In this research we trained several baseline and ensemble models without performing extensive hyperparameter optimization, so there is clear scope for further tuning using grid search, randomized search, or Bayesian optimization to potentially improve predictive accuracy. Additional machine learning approaches, including neural networks, zero-inflated or two-stage models for sparse defect counts, and time-aware models if longitudinal data become available, could also be explored to broaden the analysis. While we used SHAP to evaluate feature contributions, the work could be extended to include alternative interpretability methods and model-specific importance plots to corroborate the explanations. Finally, collecting more data across different component types and manufacturing conditions would strengthen generalizability and allow comparative studies that identify whether the same process and composition drivers hold across parts and plants, enabling more robust, actionable recommendations for production engineers.

Acknowledgment

This research was undertaken independently under the supervision of Aakash Shanbhag in pursuit of higher education. It has no institutional affiliation and is not part of any school curriculum.
The authors wish to acknowledge the use of ChatGPT in the writing of this paper. This tool was used to assist with improving the language and formatting of the paper. **The paper remains an accurate representation of the authors' underlying work and novel intellectual contributions**. This submission is the result of the authors' independent work.

References

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv. https://doi.org/10.48550/arXiv.1809.03006

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Ghatnekar, A., & Shanbhag, A. D. (2021). Explainable, multi-region price prediction. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1–7). IEEE. https://doi.org/10.1109/ICECET52533.2021.9698641

Kaggle. (n.d.). Material Lifespan Prediction Dataset. Retrieved October 27, 2025, from https://www.kaggle.com/code/talhafazal07/data-analysis-project-1/input

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.09516

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R. J., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31. https://arxiv.org/abs/1706.09516