

Machine Learning Approaches for Understanding Water Droplet Erosion in Metals

Khaled Al Hammad

A Thesis

In

The Department

Of

Mechanical, Industrial and Aerospace Engineering

Presented in Partial Fulfilment of the Requirements for the Degree of Doctorate in
Philosophy (PhD) (Mechanical Engineering)

Concordia University

Montreal, Quebec, Canada

November 2025

© Khaled Al Hammad, 2025

**CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Khaled Al Hammad

Entitled: **Machine Learning Approaches for Understanding Water Droplet Erosion in Metals**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (MECHANICAL ENGINEERING)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____	Chair
Dr. Patricia Comeau	
_____	External Examiner
Dr. Carlo Antonini	
_____	Arms-Length Examiner
Dr. Sixu Deng	
_____	Examiner
Dr. Christian Moreau	
_____	Examiner
Dr. Ryan O'Neil	
_____	Thesis Co-Supervisor
Dr. Mamoun Medraj	
_____	Thesis Co-Supervisor
Dr. Moussa Tembely	

Approved by _____

Dr. Ramin Sedaghati, Graduate Program Director

October, 2025 _____

Dr. Mourad Debbabi, Dean

Gina Cody School of Engineering and Computer Science

Abstract

Machine Learning Approaches for Understanding Water Droplet Erosion in Metals

Khaled Al Hammad

Concordia University, 2025

Water droplet erosion (WDE) significantly compromises the durability and performance of critical components in aerospace, power generation, and wind energy industries. Accurate prediction of WDE behavior, particularly the incubation period and maximum erosion rate, is vital for effective material selection and component design. Traditional empirical and analytical models often struggle to capture the complex, non-linear interactions of material properties and impact conditions, limiting their predictive accuracy and generalizability. To overcome the limitations of existing models, this thesis introduces a comprehensive framework that leverages machine learning (ML) techniques for robust and interpretable predictions of WDE. Through a series of interconnected investigations, this work develops and validates an ML approach to accurately predict the WDE incubation period and maximum erosion rate (R_e), demonstrating the effectiveness of this methodology. This includes evaluating various ML algorithms and investigating the impact of data transformation techniques on prediction accuracy. Findings revealed that statistical data transformations, such as Box-Cox and Yeo-Johnson, substantially enhanced model performance, with a linear regression model achieving an R^2 value above 90%. Feature importance analysis, particularly through SHAP (SHapley Additive exPlanations) values, identified impact velocity and surface hardness as the most influential factors, providing valuable physical insights into the erosion mechanisms. To predict the maximum erosion rate, this research addresses the challenge of limited experimental datasets in WDE studies. It showcases the transformative impact of data augmentation strategies, such as Synthetic Minority Over-sampling Technique for Regression (SMOTER). These strategies enabled highly accurate predictions, with a K-Nearest Neighbors (KNN) model achieving an R^2 exceeding 95% on training and testing data and maintaining that accuracy on unseen validation data, despite the small initial datasets. This independently validates the effectiveness of data augmentation in enhancing predictive accuracy and generalization for maximum erosion rate prediction. This study also explores the interpretability and generalization of data-driven models for predicting the WDE incubation

period. Traditional methods, such as Multiple Linear Regression (MLR) with dimensionless analysis and Principal Component Analysis combined with a Neural Network (PCA-NN), are compared with advanced tabular architectures such as the Feature Tokenizer Transformer (FT-Transformer), Tabular Neural Network (TabNet), Self-Attention and Intersample (SAINT), and Tabular Prior-Data Fitted Network (TabPFN). While advanced transformer-based models achieve strong predictive performance on training and testing datasets, their generalization varies when applied to an unseen material dataset. Notably, TabPFN and SAINT demonstrated more robust extrapolative capabilities, underscoring the importance of balancing accuracy, interpretability, and generalization.

A significant contribution of this work is developing a physics-informed model for the WDE incubation period (N_o) using Buckingham- Π dimensional analysis. This framework incorporates key parameters, including droplet impact velocity, droplet diameter, material yield strength, fracture toughness, speed of sound in both material and water, material density, water density, Vickers hardness, and Poisson's ratio. The resulting dimensionless power-law model highlights the effectiveness of physics-informed approaches in predicting WDE incubation period and capturing the complex interplay among impact conditions and material properties.

Collectively, this thesis bridges the gap between traditional empirical approaches and modern ML-based techniques, providing robust, accurate, and interpretable predictive models for WDE. The findings contribute to improved material selection, component design, and maintenance strategies in systems prone to WDE by offering a data-driven framework that enhances both predictive capability and mechanistic understanding.

Keywords: Water droplet erosion, Machine learning, Incubation period, Maximum erosion rate, Materials degradation, Prediction models, Data transformation, Data augmentation, Dimensionless analysis, Tabular foundation models.

Dedication

I dedicate this work wholeheartedly to the cherished memory of my dear parents, whose lasting love and wisdom continue to guide me even after their passing. I also thank my brothers and sisters for their steadfast support, my friends for their continuous encouragement, and my beloved wife and her family for their enduring patience, care, and unwavering love throughout this journey.

Acknowledgements

I want to express my sincere and heartfelt gratitude to my supervisors, Professor Mamoun Medraj and Professor Moussa Tembely, for their unwavering moral, financial, and intellectual support. Their guidance and constructive feedback throughout my PhD journey at Concordia University have been invaluable.

I extend my appreciation to my collaborators, and I am deeply grateful to my family and relatives for their constant support and prayers, which have been a source of strength throughout this endeavour. Special thanks go to my wife, Manal Theyab, and my family for their continuous encouragement and steadfast support.

I also gratefully acknowledge the financial support provided by the Hashemite University, Jordan.

Co-Authorship

To comply with the requirements of the manuscript-based thesis format, this thesis has been prepared in accordance with those regulations. Co-authored works are therefore included as Chapters 3, 4, and 5. The specific contributions of each co-author are described below:

Chapter 3: Article 1 “Application of Machine Learning for Predicting the Incubation Period of Water Droplet Erosion in Metals”, Reprinted from *Discover Applied Sciences*, Volume 7, Article 712, 2025, with DOI : <https://doi.org/10.1007/s42452-025-07268-8>.

Khaled Al Hammad: Conceptualization, methodology, data curation and modeling, analysis of the results, validation, and writing the manuscript.

Dr. Mamoun Medraj: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Dr. Moussa Tembely: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Chapter 4: Article 2 “Data-Augmented Learning for Estimating the Maximum Water Droplet Erosion Rate in Metallic Alloys”, currently submitted for publication to the *Journal of Engineering for Gas Turbines and Power*.

Khaled Al Hammad: Conceptualization, methodology, data curation and modeling, analysis of the results, validation, and writing the manuscript.

Dr. Mamoun Medraj: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Dr. Moussa Tembely: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Chapter 5: Article 3 “Toward Interpretable Data-Driven Models for Predicting Water Droplet Erosion Incubation Period”, currently submitted for publication to *Computational Materials Science (Journal)*.

Khaled Al Hammad: Conceptualization, methodology, data curation and modeling, analysis of the results, validation, and writing the manuscript.

Dr. Mamoun Medraj: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Dr. Moussa Tembely: Conceptualization, methodology, review and editing of the manuscript, supervision, and funding acquisition.

Dr. Julio Cesar Franco-Correa: Performed water droplet erosion experiments on austenitic stainless steel 304L; these experimental results were used as an external validation dataset for machine learning models.

Table of Contents

Chapter 1: Introduction	1
1.1 Background	1
1.2 Limitations of Traditional WDE Prediction Models	4
1.3 The Potential of Machine Learning in WDE	5
1.4 Thesis layout	5
1.5 Thesis Objectives	6
Chapter 2: Literature Review	8
2.1 Water Droplet Erosion: Mechanisms and Stages.....	8
2.2 Parameters Affecting WDE	12
2.2.1 Impact Speed.....	12
2.2.2 Droplet Size	13
2.2.3 Impact Angle.....	14
2.2.4 Mechanical Properties of the Target Material	15
2.3 Modeling of Water Droplet Erosion	16
2.3.1 Analytical and Empirical Models	16
2.4 Machine Learning Approaches in Material Science	17
2.4.1 Data Acquisition and Preprocessing	19
2.4.2 Exploratory Data Analysis (EDA) and Data Transformation.....	19
2.4.3 Data Augmentation	20
2.4.4 Dimensional Analysis	21
2.4.5 PCA-Based Models.....	21
2.4.6 Advanced Tabular Data Methods	21
2.5 Critical Gaps and Need for the Study	22
Chapter 3: Application of Machine Learning for Predicting the Incubation Period of Water Droplet Erosion in Metals	25
Abstract	25
3.1 Introduction.....	27
3.2 Methodology	31
3.2.1 Data Collection	31
3.2.2 Data Preprocessing and Statistical Analysis	33
3.2.3 Feature Importance and Correlation Matrix Analysis	34
3.2.4 Incubation Period Prediction by Machine Learning	35
3.2.5 Algorithm Selection Rationale.....	36
3.2.6 Model Performance Evaluation	38

3.2.7 Data Transformation	39
3.3 Results and Discussion	41
3.3.1 Correlation Matrix and Feature Impact.....	41
3.3.2 Engineering Implications of SHAP Analysis	44
3.3.3 Baseline Models.....	45
3.3.4 ML Model Improvements	47
3.3.5 Artificial Neural Network.....	58
3.4 Conclusion	62
Chapter 4: Data-Augmented Learning for Estimating the Maximum Water Droplet Erosion Rate in Metallic Alloys	66
Abstract.....	66
4.1 Introduction.....	68
4.2 Methodology	71
4.2.1 Maximum Erosion Rate Data Acquisition and Preparation.....	72
4.2.2 Analysis of Input Features, Contributions and Correlations.....	76
4.2.3 Maximum Erosion Rate Prediction by Machine Learning	76
4.2.4 Data Augmentation Methods in Machine Learning.....	79
4.3. Results and Discussion	82
4.3.1 Exploratory Data Analysis of Maximum Erosion Rate	82
4.3.2 Performance of Baseline Models with Hyperparameter Optimization.....	83
4.3.3 ML Model Improvements	84
4.4 Conclusion	100
Chapter 5: Toward Interpretable Data-Driven Models for Predicting Water Droplet Erosion Incubation Period	104
Abstract.....	104
5.1 Introduction.....	105
5.2 Literature Review.....	106
5.3 Methodology	110
5.3.1 Addressing Challenges of Limited Dataset in WDE	110
5.3.2 Dataset Description and Characteristics	110
5.3.3 Dataset Preprocessing and Feature Scaling	112
5.3.4 Dataset Augmentation.....	113
5.3.5 Dimensionless Analysis for Incubation Period Estimation	113
5.3.6 Multiple Linear Regression with Dimensionless Analysis.....	117
5.3.7 Principal Component Analysis for Neural Network.....	117

5.3.8 Advanced Deep Learning Architectures for Tabular Data	119
5.3.9 Model Evaluation and Validation	124
5.4. Results and Discussion	125
5.4.1 Physics-Informed Modeling and Dimensionless Analysis	125
5.4.2 ML Model with Dimensionless Groups.....	126
5.4.3 Principal Component Analysis for Improving Neural Network Performance.....	128
5.4.4 Advanced Tabular Learning Architectures.....	131
5.4.5 Model Comparison.....	134
5.4.6 Validation of Incubation Period Prediction for Austenitic Stainless Steel 304L.....	135
5.4.7 Performance Analysis and Model Ranking	137
5.5 Conclusion	137
Chapter 6: Synthesis and Discussion	139
6.1 WDE Prediction Insights	139
6.2 Balancing Accuracy, Interpretability, and Generalization.....	141
6.3 Practical Implications for Engineering Applications.....	142
6.3.1 Improved Material Selection.....	143
6.3.2 Optimized Maintenance Strategies	143
6.3.3 Enhanced Component Design.....	143
6.3.4 Mitigation of Data Scarcity:.....	143
Chapter 7: Conclusion.....	145
7.1 Summary of Research	145
7.2 Contributions to Knowledge.....	146
7.2.1 Effective Handling of Skewed Data	146
7.2.2 Use of Data Augmentation to Address Scarcity	146
7.2.3 Integration of Interpretability and Generalization	147
7.2.4 Development of a Unified ML Framework for WDE	147
7.3 Suggestions for Future Work	147
7.3.1 Hybrid Physics-ML Modeling	147
7.3.2 Expansion of Experimental Datasets	147
7.3.3 Deployment and Real-World Validation	147
7.3.4 Advanced Interpretability Techniques.....	148
7.3.5 Transfer Learning Across Rigs and Degradation Mechanisms	148
7.4 Final Remarks	148
References.....	149

List of Figures

Figure 1.1. Applications affected by WDE damage	2
Figure 1.2. Annual energy production loss over turbine lifetime	3
Figure 1.3. Schematic of the water droplet erosion rig	4
Figure 2.1. Shock wave behavior during the liquid droplet-solid surface interaction	8
Figure 2.2. Schematic of erosion damage initiation on a ductile solid surface proposed	9
Figure 2.3. Schematically, the different stages of water droplet erosion.....	10
Figure 2.4. Non-dimensional volume loss of stainless steel as a function of impact angle	14
Figure 2.5. Machine learning process	18
Figure 3.1. Data distributions and Q-Q plot for the incubation period.....	34
Figure 3.2 Workflow for predicting WDE incubation period.....	35
Figure 3.3. Spearman's rank correlation matrix	42
Figure 3.4. Features impact on the incubation period using mean absolute SHAP values.....	43
Figure 3.5. Baseline model predictions of the incubation period using various ML algorithms...	46
Figure 3.6. Model predictions after hyperparameters tuning using various ML algorithms	49
Figure 3.7 Transformed data distributions and Q-Q plot for the incubation period	52
Figure 3.8. Model predictions with logarithmic transformations using various ML models	53
Figure 3.9. Model predictions with Box-Cox transformation using various ML algorithms	54
Figure 3.10. Model predictions with Yeo-Johnson transformation using various ML models	55
Figure 3.11. Actual vs. predicted incubation period: without data transformation (LR and Ridge) and with data transformation using Ridge with Box-Cox and LR with Yeo-Johnson	57
Figure 3.12. Predicted vs. actual values of N_0 for the LR model with Box-Cox transformation..	58
Figure 3.13. Predicted vs. actual values of N_0 for the ANN model with Yeo-Johnson	59
Figure 3.14. Actual vs. predicted N_0 for PH 17-4 SS comparing ANN model and LR with Box-Cox transformation	60
Figure 4.1. Typical stages in water droplet erosion damage curves	68
Figure 4.2. WDE mass loss rate curve for Ti64 at $V_o = 250$ m/s, and $d = 463$ μ m.....	73
Figure 4.3. Impact speed ranges for the tested materials at two different droplet sizes and maximum erosion rates for different materials and droplet sizes	74
Figure 4.4. Mechanical properties of the tested materials	75
Figure 4.5. Workflow for predicting WDE maximum erosion rate.....	77

Figure 4.6. Correlation matrix of input features and target R_e and features impact using an absolute SHAP algorithm	82
Figure 4.7. Baseline tuned model predictions	84
Figure 4.8. ML models with Yeo-Johnson transformation.....	85
Figure 4.9. Predicted vs. actual R_e values of different models (Multivariate Interpolation)	87
Figure 4.10. Predicted vs. actual R_e values of different models (Mixup)	89
Figure 4.11. Predicted vs. actual R_e values of different models (SMOTER).....	90
Figure 4.12. Predicted vs. actual R_e values of different models (Noise Injection):	91
Figure 4.13. Predicted vs. actual R_e values of different models (CTGAN).....	92
Figure 4.14. Predicted unseen vs. actual values of R_e (SMOTER, Multivariate Interpolation, Mixup, and Noise Injection).....	94
Figure 4.15. ANN model predictions vs. actual R_e values and loss function with different data augmentation methods	97
Figure 4.16. ANN model predicted unseen vs. actual values of maximum erosion rate with different data augmentation methods	100
Figure 5.1. Distributions of input parameters and output variables in the dataset	112
Figure 5.2. PCA-NN architecture for incubation period prediction	119
Figure 5.3. Feature tokenization-based transformer for incubation period prediction	121
Figure 5.4. TabNet method for incubation period prediction	122
Figure 5.5. SAINT method for incubation period prediction	122
Figure 5.6. TabPFN method for incubation period prediction.....	123
Figure 5.7. Comparison between measured and predicted N_o on a logarithmic scale	126
Figure 5.8. MLR model performance for N_o using dimensionless groups from the Buckingham PI theorem vs. the original features	127
Figure 5.9. Cumulative explained variance ratio as a function of the number of principal components	128
Figure 5.10. Training and testing process for the PCA-NN model on the WDE dataset	129
Figure 5.11. PCA-NN model: Actual vs. predicted N_o values for the test set	130
Figure 5.12. FT-Transformer: Actual vs. predicted N_o values	131
Figure 5.13. TabNet: Actual vs. predicted N_o values	132
Figure 5.14. SAINT: Actual vs. predicted N_o values	133
Figure 5.15. TabPFN: Actual vs. predicted N_o values	134
Figure 5.16. Predicted unseen data results for an Austenitic Stainless Steel 304L	136

List of Tables

Table 3.1. Impact speed ranges for the materials tested with different droplet sizes.	31
Table 3.2. Mechanical properties of tested materials	32
Table 3.3. Statistics of the incubation period.....	33
Table 3.4. Summary of machine learning model performance evaluation metrics	39
Table 3.5. Summary of the different transformations employed on N_o in the present work	40
Table 3.6. Tuned hyperparameters for predictive regression algorithms	48
Table 3.7. Best ML transformed models' evaluation metrics from 5-fold cross-validation.....	56
Table 3.8. ML model performance metrics using SNN.....	59
Table 4.1. Summary of different data augmentation methods	81
Table 4.2. ML models' performance metrics with data transformation.....	85
Table 4.3. ML models' performance before and after applying data augmentation (Multivariate Interpolation)	88
Table 4.4. ML models' evaluation metrics (different data augmentations)	93
Table 4.5. ML evaluation metrics based on an unseen dataset from different data augmentation techniques	95
Table 4.6. ANN-based ML models' evaluation metrics from different data augmentations	99
Table 5.1. Physical variables and dimensions involved in the application of the Pi-theorem.....	115
Table 5.2. Performance comparison of different models.....	135

List of abbreviations and symbols

Symbol	Definition
AdaBoost	Adaptive Boosting Regressor
ANN	Artificial Neural Networks
d	Droplet Diameter
C_s	Speed of Sound
DT	Decision Trees
E	Elastic Modulus
EDA	Exploratory Data Analysis
EN	Elastic Net regression
ET	Extra Trees Regressor
FEM	Finite Element Method
GBR	Gradient Boosting Regressor
H_V	Vickers Hardness
K_{IC}	Plane Strain Fracture Toughness
KNN	K-Nearest Neighbors Regressor
LightGBM	Light Gradient Boosting Machine
LR	Linear Regression
ML	Machine Learning
MLR	Multiple Linear Regression
N_o	Incubation Period
PCA	Principal Component Analysis
R_e	Maximum Erosion Rate
RF	Random Forests

RMSE	Root Mean Square Error
R^2	Coefficient of determination
SPH	Smoothed Particle Hydrodynamics
TabPFN	Tabular Prior-data Fitted Networks
WDE	Water Droplet Erosion
XGBoost	Extreme Gradient Boosting Regressor
ρ	Density
σ_y	Yield Strength
ν	Poisson's Ratio

Chapter 1: Introduction

1.1 Background

Material degradation presents a critical challenge to the durability and performance of engineering components, often leading to progressive material loss and compromised structural integrity. Among various degradation phenomena, water droplet erosion (WDE) stands out due to its prevalence and severity in aerospace, power generation, and other industrial applications [1–3]. WDE represents a pervasive form of material damage initiated by the repetitive, high-velocity impact of discrete water droplets on solid surfaces, resulting in material removal and surface deterioration over time.

Unlike other forms of erosion, such as cavitation or solid particle erosion, WDE is characterized by the unique dynamics of liquid-solid interaction, where the impact of individual droplets generates localized high pressures that can deform and eventually remove material from the surface [2]. This process is cumulative, meaning that even seemingly minor impacts can, over time, lead to significant surface deterioration and the formation of pits, cracks, and eventually, large-scale material loss [3]. The consequences of WDE can range from reduced operational efficiency and increased maintenance costs to critical component failure, posing significant safety and economic challenges [1–3]. Figure 1.1 shows applications affected by WDE in aerospace and power generation industries, highlighting its importance across different engineering domains.

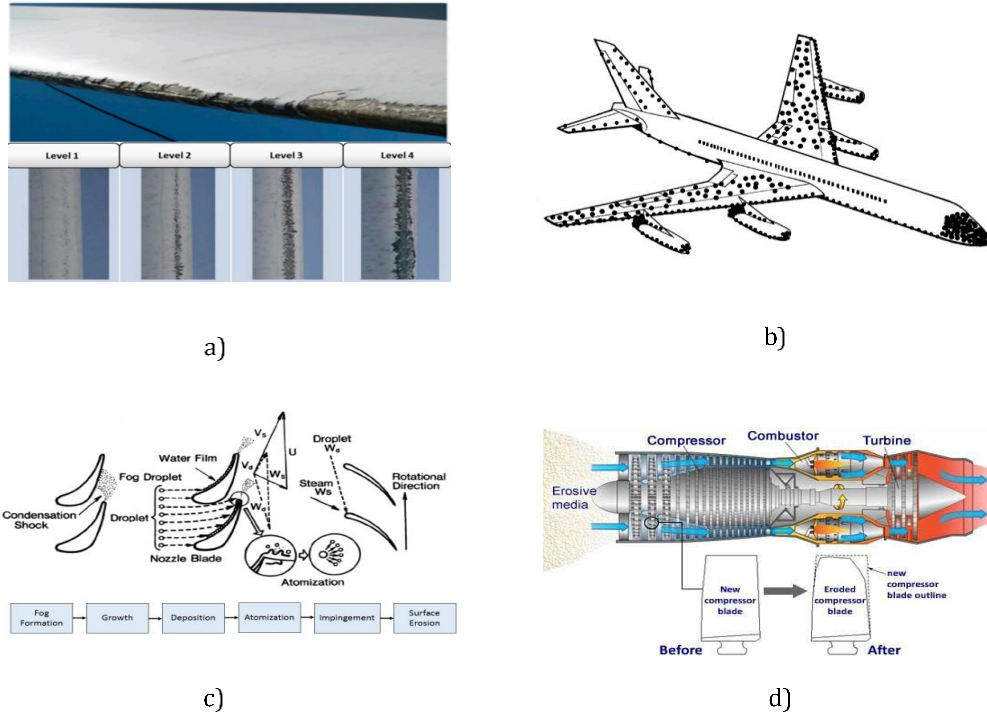


Figure 1.1. Applications affected by WDE damage: a) Leading edge erosion (LEE) in wind turbines [4, 5], b) Rain erosion [6], c) Steam turbine blades [7], d) Gas turbine compressor [3]

The process of WDE involves several stages. Initially, the impact of a water droplet creates a shock wave that propagates through the material, leading to localized stress concentrations [1]. Subsequent impacts on the same area exacerbate the damage, causing fatigue and the eventual removal of material through plastic deformation, microcracking, and fracture [2]. The severity of WDE is influenced by several factors, including the velocity and size of the water droplets, the frequency of impacts, the material properties (such as hardness, toughness, and fatigue resistance), and the surface finish of the component [3]. For instance, higher impact velocities and larger droplet sizes generally lead to more severe erosion [1]. Materials with lower hardness or poor fatigue resistance are also more susceptible to WDE [2]. Furthermore, the presence of surface irregularities or pre-existing defects can accelerate the degradation process, acting as stress concentrators and initiation sites for damage [3]. Effective mitigation strategies for WDE often involve material selection, surface coatings, and design modifications to minimize droplet impingement [1–3]. Research continues to explore novel materials and surface treatments to

enhance the resistance of engineering components to this challenging degradation mechanism [2, 3].

Industries such as aerospace, power generation (including gas and steam turbines), nuclear energy, and wind energy are common places for WDE, which can cause substantial economic losses through reduced efficiency, increased maintenance costs, and eventual component failure [2, 3]. For instance, in steam turbines, WDE can lead to severe damage of turbine blades operating under wet steam conditions. Similarly, in wind energy applications, turbine blades' leading-edge erosion (LEE) significantly diminishes aerodynamic efficiency and annual energy production [8, 9]. Figure 1.2 illustrates the progressive decline in Annual Energy Production (AEP) as leading-edge erosion (LEE) advances, transitioning from the incubation period through the initial stages of erosion to advanced blade damage [5].

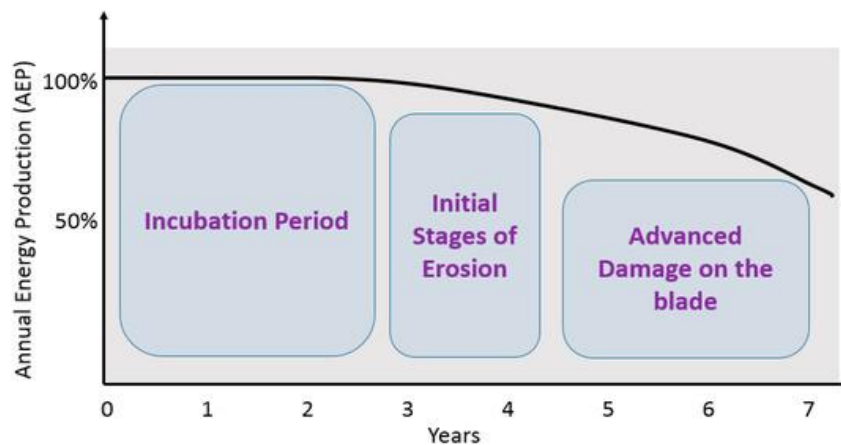


Figure 1.2. Annual energy production loss over turbine lifetime [5]

The erosion process is a progressive phenomenon that develops through multiple stages, each dominated by different damage mechanisms [10, 11]. It is influenced by a complex interplay of fluid dynamics and material properties, including impact velocity, droplet geometry, impact angle, surface roughness, material hardness, and fracture toughness. Among these stages, the incubation period—marking the onset of subsurface damage without visible material loss—and the maximum erosion rate—indicating the phase of most rapid material removal—are essential for characterizing and predicting WDE behavior.

WDE experiments expose specimens mounted on rotating arms or disks to water jets or simulated raindrops, with periodic weighing to quantify material loss. The Concordia University test rig uses a vacuum-sealed rotating disk to achieve impact speeds of 100–500 m/s, minimizing friction and droplet evaporation (Figure 1.3) [3]. It is possible to vary the impact speed, droplet size, and impact angle in this rig.

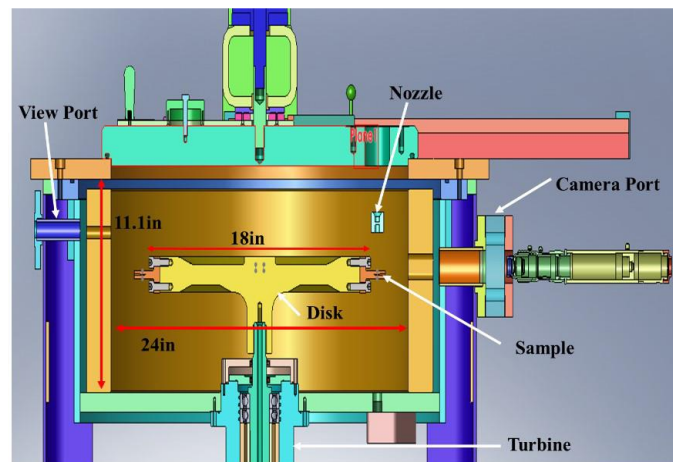


Figure 1.3. Schematic of the water droplet erosion rig [3]

1.2 Limitations of Traditional WDE Prediction Models

While WDE controlled experiments are essential for isolating key parameters in WDE behavior, the resulting datasets remain limited in scope due to the time-intensive and costly nature of testing. Moreover, assumptions embedded in both experimental configurations (rotating arm or disk) and in many traditional models—such as fixed droplet angles and uniform velocities—restrict their applicability. As a result, existing models in the literature often fail to generalize across different materials, geometries, and loading conditions. Traditional modeling approaches for WDE have predominantly relied on empirical and semi-empirical correlation frameworks, such as those proposed by Springer and Heymann [12–14], and analytical models [15–17]. While these models have provided valuable insights, particularly concerning the influence of impact conditions, they often depend on material-specific constants and adjustable parameters that lack general physical meaning. Consequently, their predictive capability is limited when applied to new materials or operating conditions. Moreover, numerical simulation techniques offer detailed physical modeling, but are computationally intensive and typically constrained to specific boundary conditions and material systems [18].

1.3 The Potential of Machine Learning in WDE

To address these limitations, machine learning (ML) has emerged as a powerful alternative. ML models can capture complex, non-linear relationships in high-dimensional datasets without the need for explicit physical equations, enabling them to generalize across diverse conditions and materials. Recent advancements in ML have shown significant promise in materials degradation applications [19–21]. Recently, ML techniques have been successfully applied to predict cavitation erosion in stainless steels [21]. Given the parallels between erosion and cavitation damage mechanisms, ML is well-positioned to advance WDE prediction, particularly in accurately identifying the incubation period and estimating maximum erosion rates.

This study seeks to bridge the gap between traditional empirical approaches and modern ML-based techniques for predicting WDE. It implements a comprehensive ML framework on experimental data obtained from a controlled test rig for various metallic materials. The methodology includes data preprocessing, hyperparameter optimization, statistical transformations such as Box-Cox and Yeo-Johnson transformations, data augmentation, and advanced tabular data techniques, which were found to significantly enhance the predictive accuracy of the models and performance evaluation across multiple algorithms.

1.4 Thesis layout

This thesis addresses the challenges in modeling water droplet erosion (WDE) by combining both traditional and advanced machine-learning (ML) methods with dimensional analysis. Chapter 2 reviews WDE mechanisms, existing models, and recent ML developments in materials engineering. Chapter 3 focuses on preprocessing techniques—such as power and log transformations—to reduce data skewness and improve incubation period prediction. Chapter 4 explores data augmentation to enhance maximum erosion rate prediction despite limited experimental data. Chapter 5 performs a comprehensive dimensional analysis and compares traditional methods—Multiple Linear Regression and PCA—NN—with advanced tabular ML architectures (FT-Transformer, TabNet, SAINT, TabPFN), with an emphasis on performance on predicting unseen data. Chapter 6 summarizes key findings and trade-offs, and Chapter 7 concludes with the study’s contributions and implications for future research and practical applications.

1.5 Thesis Objectives

This work aims to advance the predictive modeling of water droplet erosion (WDE) in metallic materials through a machine learning (ML)-based framework, addressing some challenges in predictive accuracy, data limitations, model interpretability, and mechanistic understanding. The objectives are directly aligned with the outcomes of three interrelated research studies (Paper 1 published and Paper 2–3 under review):

1. The first objective is to develop and validate ML models capable of predicting the incubation period of WDE with high precision and interpretability. To achieve this, a suite of algorithms—linear regression (LR), decision trees (DT), random forests (RF), gradient boosting regressors (GBR), k-nearest neighbors (KNN) and artificial neural networks (ANNs)—has been applied in combination with advanced data transformation techniques such as Box–Cox and Yeo–Johnson transformations. Furthermore, SHAP (Shapley Additive Explanations) analysis has been employed to quantify the dominant influence of variables, including impact velocity and surface hardness, thereby offering physically meaningful insights into the mechanisms that govern WDE. (Chapter 3)
2. While these models established a reliable framework for predicting the incubation period, their effectiveness was constrained by the limited availability of experimental datasets. The second objective, therefore, addressed the challenge of data scarcity by introducing and systematically evaluating several data-augmentation strategies, including multivariate interpolation, noise injection, Mixup, the Synthetic Minority Over-Sampling Technique for Regression (SMOTER), Gaussian Process Sampling (GP), and the Conditional Tabular Generative Adversarial Network (CTGAN). By integrating these approaches and applying the resulting models to unseen data, the study demonstrates that augmented datasets can significantly improve model robustness and extend predictive capability across diverse materials and erosion conditions. (Chapter 4)
3. With both a predictive framework and enriched datasets in place, the third objective is to systematically evaluate ML architectures in order to optimize the trade-offs between predictive accuracy, interpretability, and generalization. Following dimensional analysis to construct suitable dimensionless groups, Multiple Linear Regression (MLR) is benchmarked against PCA-based Neural Networks (PCA–NN) to assess the impact of

dimensionality reduction. These conventional methods are then directly compared with attention-based architectures, including FT-Transformer and TabNet, to test their ability to capture nonlinear dependencies in tabular WDE data. Finally, more advanced models such as SAINT and TabPFN are contrasted with all other approaches to determine their extrapolative performance on unseen material datasets. The results reveal that while simpler models such as MLR and PCA-NN remain advantageous in terms of interpretability, tabular deep learning architectures—particularly TabPFN—demonstrate superior extrapolative power, underscoring the critical need to balance model complexity with practical applicability in at least some materials science applications. (Chapter 5).

Chapter 2: Literature Review

When materials are exposed to fast-moving media such as solid particles, liquid droplets, or gas jets, their surfaces can slowly wear out over time. This type of surface damage, commonly known as erosion, is a serious concern in the field of materials science and tribology, especially in areas where components face repeated high-speed impact from such media. The nature of erosion depends on both the type of medium involved and how it interacts with the material surface. Among the different forms of erosion, water droplet erosion (WDE) has attracted increasing attention in recent years due to its growing relevance in critical sectors such as aerospace and power generation [2, 3].

2.1 Water Droplet Erosion: Mechanisms and Stages

Water Droplet Erosion (WDE) is a progressive material degradation process that occurs when high-velocity liquid droplets repeatedly impact solid surfaces. This phenomenon is commonly observed in steam and gas turbine blades, aircraft components exposed to rain, and offshore structures subjected to sea spray and mist [3, 22].

When a droplet impacts a solid surface, two main events are responsible for introducing stresses into the target material: (i) impact pressure (also known as water hammer pressure) and the stress waves it induces into the material, and (ii) the lateral outflow or jet of the droplet and the accompanying shearing effects [1, 2, 6]. Figure 2.1 shows the pressure and shock wave behavior during the liquid-solid interaction.

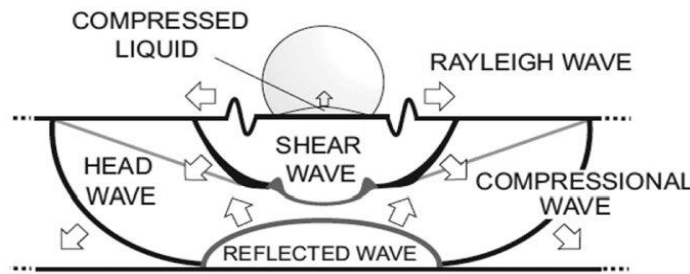


Figure 2.1. Schematic of shock wave behavior during the liquid droplet-solid surface interaction [6]

In most cases, stresses due to a single droplet impact event are insufficient to cause erosion damage. Rather, erosion damage is seen only after repetitive impacts during which stresses are being accumulated. So far, no erosion theory has successfully explained the interaction of the impacting droplets and the accumulation of stresses due to repetitive droplet impacts.

When droplets continuously strike a smooth surface, they slowly damage it. The step-by-step development of this damage is illustrated in Figure 2.2 [1]. At first, a small dent or depression appears on the surface due to the initial droplet impacts (Figure 2.2a). With repeated impacts, this depression becomes deeper. Eventually, the incoming droplet starts sitting fully within the depression, and the water gets compressed slightly longer. This increases pressure during impact, as shown in Figure 2.2b.

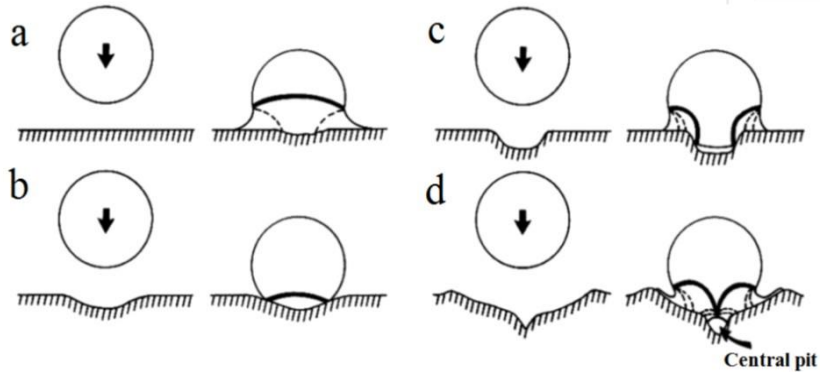


Figure 2.2. Schematic of erosion damage initiation on a ductile solid surface proposed [1]

The pressure waves from the sloped sides of the dent further compress the droplet, creating a narrow jet at the center (Figure 2.2c). This jet violently hits the bottom and causes a deeper central cavity (Figure 2.2d). The cavity keeps deepening until either air or water gets trapped inside, which acts like a cushion and slows down further erosion.

On the other hand, a long history of erosion experimentation has shown that WDE exhibits a strong time-dependency (i.e., it features different erosion rates at various time intervals), resulting in a nonlinear progression of damage [2]. This is usually presented in the so-called erosion graphs. Figure 2.3 shows a schematic of the typical (s-shaped) erosion curve observed during WDE of most bulk materials. The curve is usually divided into five distinct regions or stages [23]; stage A the incubation period (N_0) where there is only an increase in the roughness of the surface due to

the repetitive impact of the water droplets without measurable material loss; stage B the acceleration stage where micro-pits coalesce, which leads to an increase in the size of the crater; stage C the steady state maximum erosion rate where the erosion rate is maximum and remains constant for a relatively long period. This stage of constant erosion rate is known as steady state maximum erosion (R_e), and it is very important in characterizing the erosion damage; Attenuation or deceleration stage (stage D) where the erosion rate starts to decrease (usually 25 to 50% of the maximum erosion rate [2]) due mainly to the high roughness and irregularities encountered in the erosion crater, which effectively change the stress distribution, and therefore reduce the erosion rate. The deceleration of the erosion rate continues until it once again becomes constant in the terminal stage (stage E)

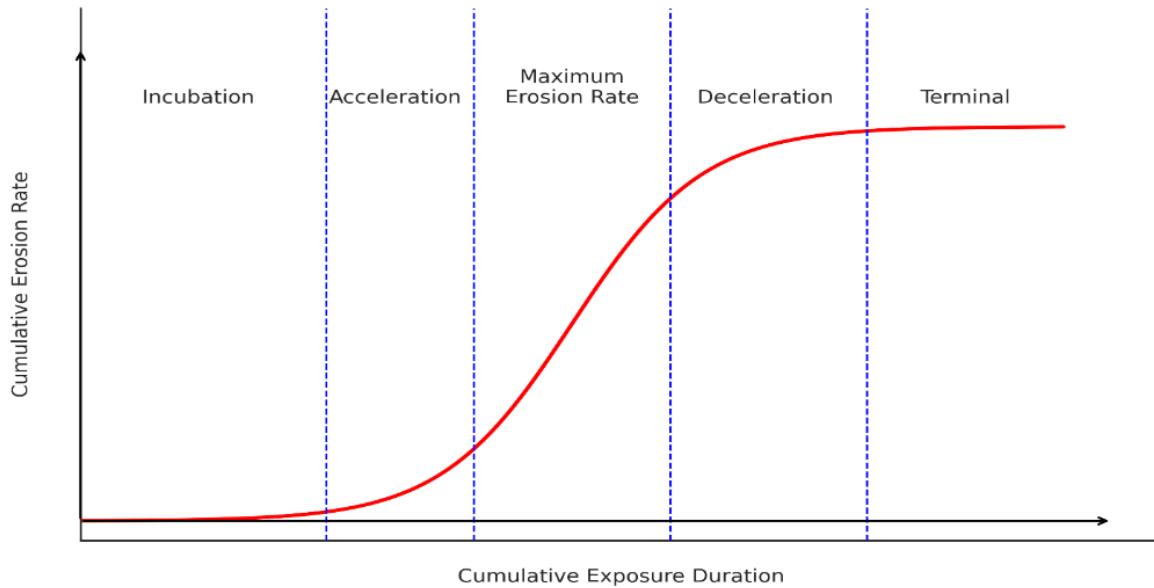


Figure 2.3. Schematic of the different stages of water droplet erosion

The stages are discussed briefly below [23]:

1. **Incubation Period (N_0):** This initial phase is marked by the absence of measurable material loss despite continuous exposure to droplet impact. However, significant subsurface activity occurs. Localized plastic deformation initiates within the near-surface layers, forming dislocation structures such as slip bands, dislocation pile-ups, and incipient microstructural alterations. This stage is essentially preparatory in nature, where the accumulation of cyclic

impact energy gradually compromises the surface integrity, setting the stage for subsequent damage evolution.

2. **Acceleration Phase:** Following the incubation period, the surface begins to exhibit early signs of degradation. The erosion rate increases as localized strain concentration leads to the initiation and propagation of micro-cracks. Pitting and the detachment of fine surface fragments become evident. This stage marks the transition from purely subsurface damage to observable surface deterioration, indicating that the material's resistance to further erosion is diminishing.
3. **Maximum Erosion Rate (R_e):** This stage represents the peak of material degradation, where the erosion rate reaches its maximum. Extensive coalescence of micro-cracks and material fatigue results in the rapid detachment of larger surface fragments. The surface morphology becomes increasingly rough and fragmented, and the depth of damage increases significantly. At this point, the material experiences the most severe mechanical and structural compromise, posing a critical threat to functional integrity, particularly in load-bearing or precision-critical components.
4. **Deceleration Phase:** After the peak, the erosion rate begins to slow down. This doesn't mean the damage stops, but the rate of new material loss reduces. This could be due to the formation of a rough or hardened surface or because the softer surface layers have already been removed. The remaining material is often more erosion-resistant than the original surface.
5. **Terminal Phase:** This is the final stage during which the material has undergone significant erosion, and the surface is heavily damaged. The remaining life of the component is minimal. Surface roughness is very high, cracks may have penetrated deeply, and in some cases, structural failure becomes a concern. Replacing or repairing the component is usually necessary at this point.

Among the five stages of WDE, the incubation period (N_0) and the steady-state maximum erosion rate (R_e) are the most critical [3, 23]. The incubation represents how long a material can resist damage before any visible erosion starts. In contrast, the steady-state erosion rate indicates how fast material is lost once erosion occurs. Being able to predict the incubation period is especially useful. It helps in planning timely maintenance so that parts can be repaired or replaced before severe damage occurs. Depending on the working environment, it also helps to choose the right

material during the design phase. Additionally, it gives insight into what happens at the very beginning of erosion and how the damage builds up [2, 3, 13]. A material with a more extended incubation period and a lower erosion rate is generally considered more erosion resistant. For these reasons, the present study focuses on predicting both the incubation time and the maximum erosion rate in water droplet erosion.

2.2 Parameters Affecting WDE

Several interconnected factors influence water droplet erosion. Droplet size, speed, frequency of impact, and shape play a significant role among droplet-related parameters [3, 6, 24, 25, 26]. Larger and faster droplets carry more energy and cause greater impact damage, while higher droplet impact frequency accelerates material degradation. The material properties of the target surface are equally important [3, 26]. More complex and stronger materials resist erosion better, though ductile materials may absorb some energy through plastic deformation [1–3]. Impact conditions such as the angle of droplet strike, temperature during impact, and the number of impact cycles contribute to how damage evolves over time [2, 3]. Environmental factors such as air, humidity, surrounding pressure, and even the liquid's chemical nature can influence droplet behaviour and surface response. In some cases, simultaneous corrosion may worsen erosion depending on the liquid composition. These parameters determine how fast and severe erosion occurs in practical situations.

2.2.1 Impact Speed

Impact speed is a critical parameter governing the initiation and progression of WDE [2, 11, 12]. The impact velocity determines whether degradation mechanisms are predominantly mechanical, chemical, or combined. At low impact velocities, degradation is mainly chemical (e.g., corrosion), with minimal mechanical damage [25, 26]. Once droplet velocity exceeds a threshold—commonly termed the critical velocity—mechanical effects dominate, producing observable surface erosion [25, 27, 28]. This velocity-sensitive transition underscores impact speed as a principal determinant of erosion behavior.

Impact speed also governs several physical phenomena associated with WDE, including water hammer pressure and radial outflow velocity, both of which scale strongly with velocity [2].

Consequently, impact speed directly influences stress wave intensity and material removal rates. The erosion rate has become a key indicator of WDE severity. Numerous studies [27, 29, 30] have consistently identified impact speed as the most influential factor in determining erosion rate and mechanisms. Heymann [31] summarized several widely used models, each expressing the erosion rate as a function of impact speed:

$$ER = aV_o^n \quad (2.1)$$

$$ER = a(V_o - V_c)^n \quad (2.2)$$

$$ER = ae^{nV_o} \quad (2.3)$$

Where ER is the erosion rate, V_o is the impact speed, a is a material constant, n is an exponent, and V_c is the threshold speed. Equation (2.2) is particularly relevant for practical applications, as it accounts for the threshold velocity below which erosion is not expected to occur.

In summary, the impact speed determines the onset of erosion and critically influences the rate and nature of material degradation. Understanding and accurately modeling its effects is essential for predictive modeling and designing erosion-resistant materials. This is a central part of the current thesis.

2.2.2 Droplet Size

Droplet size plays a critical role in water droplet erosion (WDE), as it directly influences the distribution of impact pressure, the exposed area, the duration of the pressure pulse, and the amount of kinetic energy transferred to the material surface [2, 32, 33]. Although smaller droplets occur more frequently, larger droplets generally cause more severe damage because of their higher kinetic energy and longer impact duration [2, 34, 35]. Importantly, not only the droplet's mass but also the radius at the point of contact governs the stress distribution during impact. Some studies have reported that, for the same total water volume, a small droplet can generate localized stress patterns similar to those of a much larger droplet [2, 28, 31, 34]. Nevertheless, the difference in kinetic energy remains substantial due to the variation in mass, making droplet size a decisive factor in WDE progression. According to the pressure wave erosion theory, larger droplets are likely to cause more erosion than smaller ones for the same volume of water [3]. Experimentally,

several studies confirmed the effect of droplet size on WDE, such as the work in [33, 36, 37]. For example, Hattori and Lin [37] concluded that the volume loss per droplet is proportional to droplet diameter raised to the power of 4.7, indicating the droplet diameter's strong influence. This highlights the importance of droplet size in WDE testing. Therefore, the droplet size and distribution should always be reported during WDE tests.

2.2.3 Impact Angle

Impact angle is an essential parameter in erosion studies, though opinions vary regarding its influence on WDE. Some studies suggest its effect is limited to the early stages, particularly during the incubation period [32], while others report a strong influence, with maximum erosion observed when droplets strike perpendicularly at 90° [38]. In this study, the impact angle is fixed at 90° , representing the worst-case erosion scenario. The significance of impact angle is closely linked to its effect on the normal component of impact speed, which strongly governs erosion severity [3, 32, 38]. For instance, Ahmed et al. [38] demonstrated this by testing stainless steel under varying impact angles. They found that 90° produced the most significant erosion damage for the same exposure time, as illustrated in Figure 2.4.

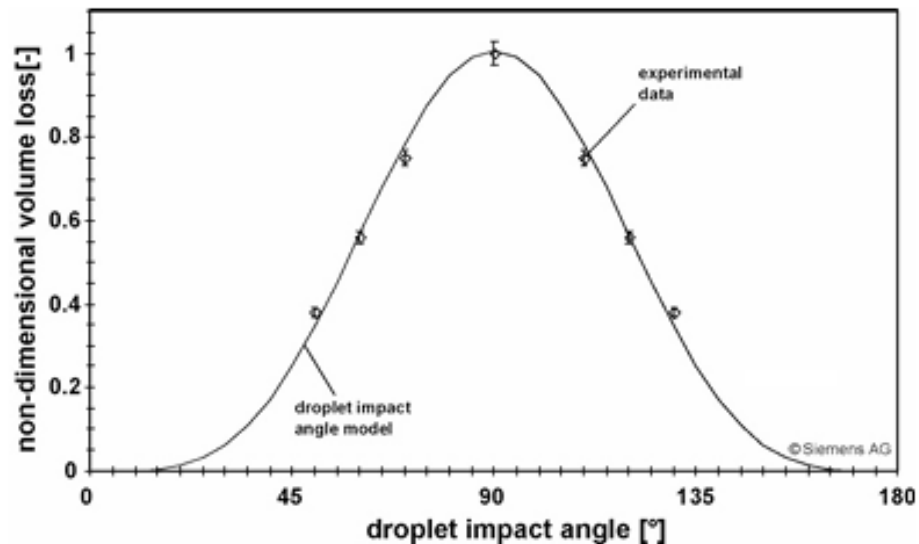


Figure 2.4. Non-dimensional volume loss of stainless steel as a function of impact angle [38]

2.2.4 Mechanical Properties of the Target Material

Mechanical properties of the target material significantly influence its resistance to WDE. The mechanical properties of the target material are crucial to water droplet erosion and all wear phenomena. In particular, strength (the yield and ultimate), hardness, toughness, hardenability, and ultimate resilience are generally the mechanical properties that influence the water erosion performance of materials [2, 35]. However, the erosion process is so complex that selecting a single mechanical property to represent erosion resistance is challenging. As explained earlier, a more adequate description of erosion resistance would be a synergistic effect of these properties. Hardness is the most widely discussed property when considering the erosion performance of a material. Heymann [35] reported that the erosion resistance is proportional to hardness raised to the power of 2.5 [35]. He also introduced the elastic modulus and ultimate tensile strength of the material as essential parameters affecting WDE behavior. He reported that the erosion resistance is directly proportional to $(\sigma_u^2 E)^{2/3}$, where σ_u is the material's ultimate strength and E is the modulus of elasticity [35]. Ahmad et al. [39] considered the hardness-induced elastic resilience as a property that would describe the erosion strength well. However, more studies are still needed to understand better the combinatorial effects between various mechanical properties on WDE behavior.

Additional mechanical properties also contribute to the erosion resistance of materials. Plane strain fracture toughness (K_{IC}) is a critical property that defines a material's ability to resist crack initiation and propagation under constrained conditions [40]. Density (ρ) influences how the material responds to the kinetic energy of impacting droplets—materials with lower density may experience accelerated erosion due to their reduced resistance to momentum transfer [34, 41].

Poisson's ratio (ν) determines the degree of lateral deformation under compressive stress, thereby affecting the stress distribution at the impact site, which in turn influences crack initiation and material fatigue [42]. Lastly, the speed of sound in the material (C_s) affects the propagation and reflection of stress waves generated by droplet impact. These stress waves can interfere constructively or destructively, amplifying or suppressing crack formation, depending on the wave interactions within the material [43].

2.3 Modeling of Water Droplet Erosion

Research on WDE has been ongoing for decades, with most studies focusing on testing metallic materials and analyzing experimental results qualitatively [28, 44]. Significant efforts have also been made to predict this phenomenon using semi-empirical and empirical approaches. In recent years with the advancement of computing tools and data analysis techniques, machine learning (ML) has emerged as a promising approach in materials science and has been used to predict the WDE behavior as well. The following section will briefly introduce these models.

2.3.1 Analytical and Empirical Models

Over the years, many empirical and semi-analytical models have been proposed to understand and predict WDE behavior. One of the early efforts in this direction was made in 1974 by Springer and Baxil [12], who suggested a power law relationship between erosion rate and impact velocity. While this model has been commonly referred to in the literature, it has limited success in accurately estimating the incubation period in the WDE process, as it does not account for the early damage stage where no material loss is yet visible. To address these limitations, Heymann [14] introduced a more detailed model that considered both material properties and impact conditions. Although insightful, this model also depends on empirical constants that need to be determined through experiments for each material, which limits its general applicability. Lesser [17] attempted an analytical solution to the liquid-solid impact problem in 1981, mainly concentrating on shock wave propagation in the liquid while neglecting the physical processes occurring in the solid target. In the following years, a few attempts were made to build more physically grounded models using fracture mechanics. Heymann's widely cited model [14] offers equations to estimate the incubation time and maximum erosion rate in the steady-state stage. However, its predictive accuracy was limited to specific materials, such as titanium and certain types of steel, under restricted operating conditions. Anderson et al. [24] studied WDE on various materials, including metal alloys like stainless steel 316 and aluminum 6061, as well as non-metallic materials, to understand their erosion behavior under controlled conditions. However, their study was restricted to a single impact speed and angle for all the tested materials, which may not accurately reflect the wide range of conditions found in real-world applications. Recently, Ibrahim and Medraj [26] developed a semi-analytical approach that addresses the threshold criteria for initiating WDE damage in metals. This model predicts the threshold speed for WDE based on

the mechanical properties of the materials. Recently, Slot et al. [45] proposed a fatigue-based analytical model using a wide range of experimental data for both ferrous (stainless steel AISI 316) and nonferrous (aluminum 6061-T6) materials. They claimed excellent agreement for impact speeds ranging from 140 to 400 m/s. However, the model's focus on specific materials may limit its broader applicability to other alloys and operating conditions. Smoothed Particle Hydrodynamics (SPH) and Finite Element Method (FEM) simulations have been applied to investigate the effects of droplet size, velocity, and impact angle on the erosion process [46]. These studies employed simplified material and damage models for a Ti-6Al-4V titanium alloy target surface. However, they employed a relatively coarse finite element mesh to balance simulation accuracy with computational resource constraints.

Empirical and semi-empirical models were developed to predict erosion behavior based on material properties and erosion conditions [12, 14, 26, 32]. However, these models often depend on several adjustable parameters that do not have a clear physical interpretation, which restricts their accuracy and limits their applicability across different materials and conditions. These drawbacks underscore the need for more reliable and generalized prediction methods, paving the way for the application of machine learning models in erosion prediction.

2.4 Machine Learning Approaches in Material Science

To overcome the limitations of traditional empirical and semi-analytical models in predicting water droplet erosion (WDE), machine learning (ML) has emerged as a powerful alternative. ML models can capture complex, non-linear relationships within high-dimensional datasets without relying on explicit physical equations, allowing them to generalize across a wide range of conditions and materials [20, 47]. The general ML workflow includes data acquisition, preprocessing, model training, testing, and deployment [19], as illustrated in the flowchart shown in Figure 2.5.

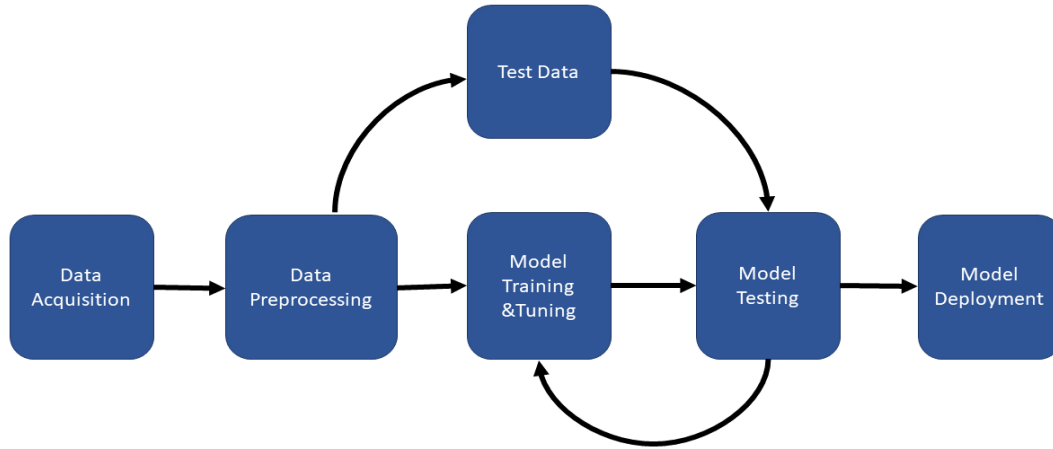


Figure 2.5. Machine learning process

Recent advancements in ML have shown significant promise in materials science applications such as wear prediction, surface roughness estimation, fatigue modeling, and cavitation damage prediction of stainless steels [19–21]. ML is a powerful tool in materials science that can uncover hidden trends in complex datasets without requiring fully defined physical relationships [47]. ML has been successfully applied to predict mechanical strength [48], corrosion resistance [49], and fatigue life [50], demonstrating its versatility in tribological applications. In erosion-related problems, such as water droplet erosion (WDE), solid particle erosion, and cavitation damage, ML models have shown promise by capturing non-linear relationships that traditional models struggle to address [19–21]. For instance, support vector regression outperformed empirical methods in predicting erosion in pipe elbows [51], while Artificial Neural Networks (ANNs) were used to accurately predict cavitation damage [21], as well as wear loss in aluminum-magnesium alloys [52].

In addition, ML can be used with Multiple Linear Regression (MLR) [53] to capture fundamental relationships between variables and the response. Still, its linear assumptions and outlier sensitivity limit its ability to model WDE’s non-linear dynamics. Tree-based models, including Decision Trees (DT) [54], Random Forests (RF) [54], and Gradient Boosting Regressors (GBR) [55], excel in noisy, small datasets by capturing non-linear and threshold behaviors; DTs model critical transitions like threshold velocities but risk overfitting, RFs and GBRs capture subtle trends

despite higher computational cost and reduced interpretability ANNs could be effective in modeling complex, non-linear interactions, such as those between impact conditions and material properties, and may outperform LR in predicting wear loss for alloys. However, they require large datasets to avoid overfitting, which is difficult to achieve due to the constraints, cost, and time demands of WDE experiments.

2.4.1 Data Acquisition and Preprocessing

Robust ML models rely heavily on high-quality data and thorough preprocessing. Data were sourced from experimental results and literature, covering key features influencing WDE, such as impact velocity, droplet size, material hardness, and fracture toughness [2, 3, 14, 26]. The dataset underwent several preprocessing steps, treating outliers and normalizing variables to ensure numerical stability [21, 47, 48]. Feature importance and correlation analysis were conducted to evaluate the influence of input variables on the output. Finally, the preprocessed dataset was split into training and testing subsets to facilitate reliable model evaluation and generalization.

2.4.2 Exploratory Data Analysis (EDA) and Data Transformation

Correlation analysis, such as Spearman's rank correlation matrix [56], plays a critical role in modeling by quantifying monotonic relationships between variables, with coefficients ranging from -1 to +1, often visualized through heatmaps. While effective at revealing directional trends, these correlations frequently miss complex, non-monotonic interactions essential for capturing multifaceted dynamics [56].

Data transformation techniques [57, 58] may have the potential to improve the accuracy of machine learning models by preprocessing input features, such as impact conditions and material properties. Methods such as Box-Cox and Yeo-Johnson transformations stabilize variance and optimize data distributions, significantly enhancing model precision, particularly for predicting WDE's incubation period, while preserving physical insights into erosion processes. Together, these transformations and correlation analysis enable ML models to better capture the complex dynamics of WDE, balancing predictive power with interpretability.

2.4.3 Data Augmentation

Data augmentation is a ML technique used to create new, synthetic data from existing datasets, helping improve the accuracy and generalizability of model predictions [59–63]. It is particularly beneficial when the available training data is limited or not diverse enough. By increasing the size and variety of the dataset, augmentation reduces the chances of overfitting, allowing the model to perform better on unseen data. Several studies have demonstrated the effectiveness of this approach across various domains [60–62]. Therefore, synthetic data generation techniques are expected to enhance ML model performance in WDE by addressing small or imbalanced datasets. Collectively, these techniques could improve predictive accuracy and robustness in WDE modeling.

For regression problems, common augmentation strategies include:

- **Noise-based perturbation:** Introducing small, controlled variations in input feature to simulate measurement uncertainty and expand the feature space.
- **Interpolation and extrapolation methods:** These methods generate synthetic data points between or slightly beyond observed values to increase the density of training data in underrepresented regions.
- **SMOTE for regression (SMOTER):** An extension of Synthetic Minority Over-sampling Technique (SMOTE) designed to generate synthetic continuous targets by interpolating both input features and output variables.
- **Generative approaches:** Advanced techniques such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) can be adapted to regression tasks to simulate realistic feature–target pairs that follow the underlying physical distribution.

In broader machine learning applications, these augmentation strategies have been shown to improve regression model generalization in domains with sparse or costly data collection. For WDE modeling, they are especially relevant for mitigating small datasets, enriching rare combinations of material properties and impact conditions, and improving the robustness of incubation-period predictions.

2.4.4 Dimensional Analysis

Dimensional analysis is utilized to ensure dimensional homogeneity, and by applying Buckingham's π theorem, nondimensional expressions are derived to establish relationships between dependent and independent variables or parameters [64, 65]. Dimensional analysis, based on Buckingham's π theorem, ensures dimensional homogeneity and derives nondimensional expressions that establish robust relationships among variables [64, 65]. This powerful method, as demonstrated in theoretical investigations on ball bearing parameters using Buckingham's π theorem [66], effectively reduces the number of independent variables, simplifies problem-solving, and generalizes results, particularly when complete mathematical models are unavailable. It can provide a framework for generalizable models, such as those in solid particle erosion that link material properties to damage mechanisms, enhancing interpretability through physical parameters. This approach helps capture non-linear effects and abrupt changes in erosion behavior, refining models to account for complex phenomena

2.4.5 PCA-Based Models

Principal Component Analysis (PCA) [68] reduces highly dimensional data into principal components, improving prediction stability on small datasets [68, 69]. In materials science, PCA has been used to focus on key patterns, such as interactions between features, minimizing overfitting [69]. For material science predictions, although PCA can stabilize predictions by reducing redundant features such as correlated mechanical properties, transformed components may obscure direct physical relationships, requiring careful selection [69].

2.4.6 Advanced Tabular Data Methods

Transformer-based models designed for tabular data have advanced erosion-related predictions by leveraging attention mechanisms to capture complex relationships within datasets [70–75]. The FT-Transformer [70] employs self-attention to model intricate feature interactions, demonstrating strong potential for WDE by effectively capturing relationships between variables such as droplet size and material properties. Similarly, SAINT integrates self-attention with inter-sample attention to account for cross-sample relationships, excelling in handling complex datasets. However, its performance on smaller WDE datasets is limited by the risk of overfitting, which poses challenges

for its application in data-scarce scenarios [71]. Additionally, TabNet combines decision-tree-like feature selection with deep learning, providing both interpretability and computational efficiency [72, 73], and has shown promise in leveraging augmented data to enhance performance, particularly on smaller datasets where overfitting is a concern. In contrast, Tabular Prior-data Fitted Networks (TabPFN) represent a promising methodology for predicting water droplet erosion (WDE), particularly under conditions of limited data availability [74, 75]. TabPFN achieves robust generalization even on small datasets by employing pre-trained transformers for zero-shot learning. Empirical evidence indicates that TabPFN consistently outperforms conventional models in materials science applications, underscoring its suitability for estimating WDE incubation periods with minimal experimental input [74, 75].

These ML approaches could offer diverse strategies for WDE prediction, balancing interpretability (e.g., LR, MLR) with the ability to capture complex dynamics (e.g., ANNs, transformers). Their effectiveness depends on the dataset size, the relevance of input features, and the relation between inputs and outputs. To ensure reliability, model performance is quantitatively evaluated using standard metrics [48,49] such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), which collectively assess prediction accuracy, residual variance, and generalization capacity. Ultimately, the limitations of traditional water droplet erosion (WDE) models—constrained by material-specific constants and poor generalizability across varying conditions—underscore the need for more robust predictive approaches. While Section 1.2 highlights the limitations of existing methods, the objectives and contributions of this study build on these insights by proposing a potential ML-based framework for predicting both the WDE incubation period and maximum erosion rate with improved accuracy and interpretability.

2.5 Critical Gaps and Need for the Study

WDE affects the performance and life of components in various industries. Components such as turbine blades, compressor vanes, and aircraft surfaces are often subjected to high-speed liquid droplet impacts, which over time lead to material damage and failure. A key challenge in understanding WDE is predicting the incubation period and the maximum erosion rate. Both of these parameters are essential for designing erosion-resistant materials and components.

Traditional approaches to WDE studies often rely on experimental testing, which is resource-intensive and time-consuming. Moreover, many existing models in the literature are tailored to specific materials or conditions, making it challenging to apply them broadly. To address these gaps, the present thesis explores the application of ML for predicting both the incubation period and maximum erosion rate under different material and testing conditions. This work leverages a relatively extensive dataset generated by previous researchers at Concordia University for five industrially important alloys: 12% Cr stainless steel, 17-4 PH stainless steel, Ti-6Al-4V, Al-2024-T4, Al-7075-T6, and TiAl.

The thesis undertakes a study that includes building ML models using diverse input features such as mechanical properties (e.g., elastic modulus, hardness, yield strength, and plane strain fracture toughness), and impact conditions. Various ML model types are examined and compared, from simple regression-based to more advanced deep learning and transformer-based models. To improve performance and handle data scarcity, data transformation and augmentation techniques are also explored. Alongside predictive accuracy, the interpretability of the models is assessed to understand which factors most influence erosion behavior.

The broader motivation behind this work is to provide a practical and flexible alternative to empirical and analytical methods for prediction. A well-trained ML model, based on reliable experimental data and relevant physical parameters, can be a powerful tool for engineers and designers working in erosion-prone environments. Such a tool can reduce the need for repeated testing, support material selection, and accelerate the development of erosion-resistant components—ultimately contributing to increased efficiency, safety, and durability in critical industrial applications.

This research advances water droplet erosion (WDE) prediction through a methodologically progressive framework addressing three critical challenges. The first study [128] established a robust approach for incubation period prediction by integrating machine learning with statistical data transformations, demonstrating the feasibility of combining experimental data with computational techniques to enhance predictive accuracy beyond conventional empirical models. The investigation systematically evaluated multiple algorithmic architectures while employing feature importance analysis to identify dominant physical parameters governing erosion initiation.

The second study addressed the inherent challenge of maximum erosion rate prediction under data scarcity conditions. By implementing advanced data augmentation methodologies, the research demonstrated that synthetic data generation could effectively overcome limitations in experimental datasets, enabling reliable erosion rate modeling across diverse material systems. This approach proved particularly valuable for scenarios where comprehensive experimental data acquisition remains prohibitively resource intensive.

The third investigation extended this work by conducting a comparative analysis of model architectures, evaluating the trade-offs between predictive performance, generalization capabilities, and physical interpretability. The research contrasted contemporary deep learning methodologies with physics-informed approaches, revealing fundamental insights into model selection criteria for engineering applications. This analysis highlighted the critical balance required between computational sophistication and domain knowledge integration, establishing guiding principles for implementing machine learning in materials degradation prediction.

These studies show that machine learning is a powerful tool for understanding and predicting WDE. By using smart data processing methods, trying different models, and considering both accuracy and physical understanding, prediction systems can be built that are strong and easy to apply in real-world scenarios. These findings can hopefully contribute to a better understanding of WDE.

Chapter 3: Application of Machine Learning for Predicting the Incubation Period of Water Droplet Erosion in Metals

Khaled AlHammad, Mamoun Medraj, Moussa Tembely

Note. This chapter is adapted from: AlHammad, K., Medraj, M., & Tembely, M. (2025). Application of machine learning for predicting the incubation period of water droplet erosion in metals. Discover Applied Sciences, 7(7). <https://doi.org/10.1007/s42452-025-07268-8>

Abstract

Water droplet erosion (WDE) is a critical degradation phenomenon that significantly affects component lifespan and performance in power generation, aerospace, and wind energy industries. The incubation period—the initial phase before visible material loss occurs—is particularly crucial for maintenance planning and material selection, yet remains challenging to predict accurately due to the complex interplay of material properties and impact conditions. Traditional empirical models have shown limited predictive capability due to their reliance on numerous adjustable parameters with insufficient physical interpretation. This study aimed to develop and validate a machine learning (ML) approach for accurately predicting the WDE incubation period across different metallic materials and impact conditions. The performance of various ML algorithms is evaluated while investigating the effect of data transformation techniques on prediction accuracy. A range of ML models—linear regression (LR), decision tree regressor (DT), random forest regressor (RF), gradient boosting regressor (GBR), and artificial neural networks (ANN)—were trained and validated using experimental data from five different alloys under various impact conditions. Data transformation methods significantly enhanced model performance, with the LR model using Box-Cox transformation achieving the highest accuracy ($R^2 > 90\%$, low MAE), followed by the ANN model with Yeo-Johnson transformation ($R^2 > 85\%$). Feature importance analysis through SHAP values revealed that impact velocity and surface hardness were the most

influential factors affecting incubation period, providing valuable physical insights into the erosion mechanism. Hyperparameter optimization techniques showed minimal improvement in model performance, suggesting that the transformations effectively captured the underlying relationships in the data. This research represents the first comprehensive application of ML techniques to WDE incubation period prediction, establishing a methodological framework that integrates experimental data, statistical analysis, and advanced ML algorithms. Unlike previous approaches, our methodology (1) systematically evaluates multiple ML algorithms and transformation techniques for WDE prediction, (2) provides quantitative assessment of feature importance that aligns with physical understanding of erosion mechanisms, (3) demonstrates superior predictive accuracy compared to traditional empirical models, and (4) offers a generalizable approach applicable across different metallic materials and impact conditions. This work bridges the gap between data-driven modeling and physical understanding of WDE, providing a valuable tool for engineers to optimize material selection and maintenance strategies in erosion-prone applications.

Keywords. Water droplet erosion, Machine learning, Incubation period, Prediction models, Material degradation, Data transformation.

Article Highlights

- Data-driven models predict water droplet erosion of metals, enabling engineers to design more durable components.
- Impact speed and surface hardness are key factors in erosion resistance.
- Statistical data transformation substantially improves prediction accuracy.

3.1 Introduction

Material degradation due to erosion is a critical issue in materials science and tribology that plays a significant role in various applications such as aerospace and power generation [3, 76]. Despite extensive research, challenges persist in predicting the incubation period and erosion rates, as well as in fully understanding the underlying mechanisms [3, 14, 77]. Previous studies have focused on the mechanics of erosion, the development of empirical models for erosion prediction, and experimental methods for studying erosion resistance [14, 77, 78, 79, 80].

Analytical approaches to the liquid-solid impact problem have primarily focused on the mechanics of shock wave propagation within the impacting liquid, often neglecting the physical phenomena occurring within the solid target [17]. Empirical and semi-empirical models have been developed to estimate erosion behavior based on material properties and erosion conditions, such as those proposed by Springer [12] and Heymann [14]. However, these models often rely on numerous adjustable parameters that lack physical meaning, posing challenges for accuracy and generalization. For instance, Heymann's widely cited model [14] provides equations to estimate the incubation period and maximum erosion rate at the steady-state stage, but its predictive capability is limited to specific materials and operating conditions. This limitation highlights the need for more robust and generalized predictive models. Other models, such as the Springer model, predict incubation periods in homogeneous materials [12]. This model derives an algebraic expression for the incubation period based on fatigue theory. However, its applicability diminishes under altered impact conditions and becomes inapplicable when the damage regime and dominant mechanisms change. Similarly, Mednikov et al. [81] investigated and quantified damage onset in 20Kh13 steel blades during the droplet erosion incubation period. While their findings offer valuable insights into erosion wear, the study's focus on a single steel type and limited exploration of damage mechanisms restrict its broader applicability. These limitations highlight the need for more comprehensive predictive models for the incubation period in WDE.

In recent efforts to improve accuracy, Ibrahim and Medraj [26] developed a semi-analytical approach that addresses the threshold criteria for initiating WDE damage in metals. This model predicts the threshold speed for WDE based on the mechanical properties of materials. Similarly, Slot et al. [45] proposed a fatigue-based analytical model that agreed well with experimental data

from ASTM laboratory tests on stainless steel AISI 316 and aluminum 6061-T6, conducted at impact speeds ranging from 140 to 400 m/s. However, the model's focus on specific materials may limit its broader applicability to other alloys and operating conditions.

Predicting the WDE incubation period remains a challenging task due to various influencing factors. Mora et al. [46] employed Smoothed Particle Hydrodynamics (SPH) and Finite Element Method (FEM) simulations to investigate the incubation regime of WDE, focusing on the effects of droplet size, speed, and impact angle. However, their findings, based on Ti-6Al-4V alloy and specific damage models, may not be generalized to other materials and practical conditions. This shows the need for other predictive approaches that are flexible, general, and capable of addressing the limitations of empirical and analytical models.

To overcome these challenges, machine learning (ML) techniques offer a promising alternative. In contrast to empirical models that rely heavily on adjustable parameters with limited physical interpretation, ML models can capture complex, non-linear relationships from extensive datasets and generalize effectively across a wide range of materials and conditions. By incorporating ML, this research aims to develop more accurate and comprehensive predictions for the incubation period in WDE.

Machine learning has proven effective in materials science, particularly for solving complex problems and predicting experimental trends. Researchers have applied ML to predict wear loss and surface roughness [82], fatigue damage [83], fatigue crack opening [84], cavitation damage [21], and composite material properties [85]. For instance, Gao et al. [21] employed ML techniques to predict cavitation damage in stainless steels 316L and 420, achieving improved prediction accuracy. Given the parallels between cavitation and erosion damage mechanisms, ML is expected to provide substantial improvements in predicting WDE, particularly in critical regimes such as the incubation period and maximum erosion rate.

While extensive research has focused on WDE of metallic materials, studies on polymers and ceramic coatings have revealed distinct erosion behaviors and mechanisms. For ceramic coatings, Oka and Miyata [86] demonstrated that erosion behavior is characterized by the incubation period and subsequent damage depth rate, with significant variations among ceramic materials. Their

research showed that zirconia-based ceramics exhibited the longest incubation periods and lowest damage depth, making them particularly promising for erosion-resistant applications. In contrast, some ceramic bulk materials displayed short incubation periods and significant damage rates. For polymeric materials, Wang et al. [87] investigated the WDE behavior of aeronautical PMMA, revealing that different manufacturing processes significantly affect erosion resistance. Their study identified four primary damage mechanisms in polymers: water-hammer pressure-induced deformation, stress wave propagation, lateral flow shear, and hydraulic penetration. The erosion process in polymers typically progresses through distinct stages, beginning with surface roughening during the incubation period, followed by crack formation and propagation, and ultimately material detachment. Ibrahim and Medraj [3] further noted that polymeric composites used in wind turbine blades exhibit complex erosion patterns influenced by fiber orientation and matrix properties, with thermosetting polymers generally showing better erosion resistance than thermoplastics. These studies collectively highlight that material selection for WDE-prone applications must consider the specific erosion mechanisms and incubation behaviors unique to polymers and ceramic coatings, which differ substantially from those observed in metals.

The present study seeks to bridge the gap between traditional empirical models and ML techniques by focusing on the prediction of WDE incubation period under different operating conditions. Consistent experimental data collected from our controlled erosion rig is used to train and validate the ML models. By integrating statistical data analysis and ML approaches, this work aims to enhance prediction accuracy and facilitate the selection of materials based on their predicted incubation periods. The methodology includes data collection and preprocessing, statistical analysis, hyperparameter tuning, data transformation techniques, and the application of ML models using regression-based algorithms.

ML applications in tribology have grown significantly in recent years, the specific application to WDE incubation period prediction represents a novel contribution to the field. Most recent studies focus on general wear, friction coefficient, or tool wear, but not specifically on WDE incubation periods. For example, very recently in 2025, Jatavallabhula et al. [88] demonstrated the effectiveness of ML for predicting tribological properties at elevated temperatures using Elastic Net, KNN, and Random Forest regression models. They predicted wear and coefficient of friction for blended coatings and found that Random Forest regression performed best with R^2 values of

0.95-1.00. On the other hand, in 2024, Harouz et al. [89] employed Random Forest, Gradient Boosting Regressor, and XGBoost algorithms to predict wear rate of WC/Co-based cermets with varying content of Ni, Cr, TiC, TaC, and NbC. Perhaps the most relevant recent work is that of Espinoza-Jara et al. [90] who used Artificial Neural Networks (ANN) to predict the erosion-corrosion behavior of X65 steel. They compared ANN with conventional linear regression based on multifactorial analysis and demonstrated better statistical accuracy with ANN but noted limitations regarding training dataset size. Similar issue with dataset size has been encountered in the current work due to the nature of WDE experiments that required a long time to perform. Each incubation period measurement takes weeks to complete.

While ML has been applied to various tribological properties such as wear, friction, film thickness, this work advances the state-of-the-art by targeting WDE incubation period prediction. This presents unique challenges compared to general wear prediction due to the complex physical mechanisms involved in the WDE process, particularly the liquid-solid interactions. In the current work, multiple ML models (LR, DT, RF, GBR, ANN) have been systematically evaluated with various statistical transformations, providing a more comprehensive comparison than many recent studies that typically focus on 2-3 algorithms. Besides, the detailed SHAP analysis provides specific insights into the physical factors governing WDE incubation, which has not been previously explored using ML techniques.

The remainder of this paper is structured as follows: Section 3.2 outlines the methodology, detailing the experimental dataset, the properties of the selected metals that were used as input variables, and the machine learning algorithms employed in this study. Section 3.3 presents the results and discussion of our machine learning approaches, including analysis of feature importance using SHAP values, baseline and optimized model performance, and the effects of various data transformations. Finally, Section 3.4 concludes the paper with a summary of key contributions and suggestions for future research directions. Throughout this structure, we demonstrate how machine learning techniques can effectively predict WDE incubation periods with higher accuracy than conventional approaches, while also providing valuable physical insights into the erosion mechanisms.

3.2 Methodology

3.2.1 Data Collection

Data were collected from experiments conducted by the TMG group at Concordia University, using a rotating disc rig for water droplet erosion testing [26]. Five alloys were tested, namely, 12% chromium (Cr) stainless steel, 17-4 PH stainless steel, Ti-6Al-4V alloy, aluminum 2024-T4, and 7075-T6 alloys. Two different average droplet sizes—460 μm and 600 μm — and a range of velocities from 140 to 350 m/s were used. The selection of these materials was primarily based on their diverse elastic and mechanical properties, which are vital for understanding the influence of material properties on the incubation period and for validating the developed model [3, 14]. Moreover, these alloys hold practical significance for applications where WDE is a major failure threat [3, 26, 77, 79, 91]. For instance, stainless steel is commonly utilized as a steam turbine blade material, Ti-6Al-4V is used as a blade material for compressors in gas turbines, and aluminum alloys are employed in aerospace applications for aircraft bodies [7, 26, 32]. All these applications are susceptible to WDE. Table 3.1 provides the materials and the range of impact velocities for the droplet sizes used. The incubation period of the five materials was measured at different impact speeds.

Table 3.1. Impact speed ranges for the materials tested with different droplet sizes

Materials	$d = 463 \mu\text{m}$		$d = 603 \mu\text{m}$	
	Impact Speed (m/s)	N_o (min)	Impact Speed (m/s)	N_o (min)
17-4 PH Stainless Steel	350 - 220	2- 141	300 - 215	4.25 - 90
Ti-6Al-4V	350 - 210	1 - 139	300 - 200	2.5 - 73
Al7075-T6	300 - 155	1 - 160	250 - 150	2.5 - 90
12%Cr Stainless Steel	350 - 205	1.5 - 227	300 - 200	3 - 102
Al2024-T4	300 - 150	1 - 98	250 - 145	2 - 52

Table 3.2 shows the mechanical properties of the tested materials, such as elastic modulus (E), plane strain fracture toughness (K_{IC}), yield strength (σ_y), Vickers hardness (H_v), density (ρ), Poisson's ratio (ν) and speed of sound (C_s) in the metal. These properties provide important insights into materials' behavior, durability, and performance under different conditions and loads [3, 20].

Table 3.2. Mechanical properties of tested materials

Material types	Material Properties						
	E (GPa) [92]	H_v (MPa) [26,92,93]	σ_y (MPa) [94]	K_{IC} (MPa.m ^{1/2}) [20,92,93,95]	ρ (kg/m ³) [92]	ν [50,92]	C (m/s) [92,93,95]
Al2024-T4	73.1	1422	543	30	2780	0.33	5088
Al7075-T6	71.7	1834	702	35	2810	0.33	5026
Ti-6Al-4V	114	3316	1105	65	4430	0.32	5072
12%Cr stainless steel	210	2942	980	40	7810	0.29	5100
17-4 PH stainless steel	195	3531	1177	90	7750	0.28	5006

To understand how different materials respond to WDE, it is essential to consider their mechanical properties, starting with elastic behavior. E and ν govern the elastic response of materials during initial water droplet impact. They directly influence stress wave propagation and the material's ability to absorb and dissipate impact energy without permanent deformation. The elastic response is particularly critical during the incubation period when damage accumulates primarily through elastic-plastic transitions. K_{IC} characterizes a material's resistance to crack propagation, which is a fundamental mechanism in the later stages of the incubation period. Materials with higher fracture toughness can withstand more impact cycles before developing the microcracks that eventually lead to material removal. σ_y represents the threshold stress at which a material begins to deform plastically. It is directly relevant to WDE as it determines when repeated impacts transition from causing elastic deformation to permanent damage accumulation. Surface hardness measures a material's resistance to localized plastic deformation. Our SHAP analysis (detailed below) confirmed this parameter's high importance, as it directly relates to a material's ability to resist the localized plastic deformation caused by droplet impacts. Stress wave propagation and/or reflection are governed by ρ and C_s . These parameters determine the acoustic impedance of the material, which governs the magnitude and nature of stress waves generated during water droplet impact. The pressure at impact is proportional to ρC_s , making these parameters fundamental for predicting erosion behavior.

To effectively utilize these parameters for predicting N_0 , a robust experimental dataset was essential. The dataset employed in this study was collected from 56 consistent experiments. Each

experiment required over 10 data points to accurately determine N_o . These experiments were conducted over the years using the same experimental rig to minimize bias and ensure consistency. This carefully curated, high-quality yet limited dataset forms the foundation for model training in this study. However, recent studies in related fields have shown that accurate predictions can still be achieved with relatively small datasets. For example, Mhetre et al. [96] successfully predicted the diameter of TiO₂ nanofibers using various regression models with an R^2 -score exceeding 90%, despite using only 30 data points. Given these constraints, all machine learning models in this study were carefully fine-tuned and validated to avoid overfitting, as will be discussed in the following sections.

3.2.2 Data Preprocessing and Statistical Analysis

The collected data on the N_o were preprocessed, and statistical analysis was performed, as shown in Table 3.3.

Table 3.3. Statistics of the incubation period

	Mean	Median	Standard Deviation	Range	Kurtosis	Skewness
$N_o(\text{min})$	40.50	15.00	49.10	226.00	2.50	1.60

The results indicate that the incubation period (N_o) does not follow a normal distribution. From Table 3.3, the skewness value for the incubation period is 1.60, which is greater than 1. This indicates that the data distribution is skewed, as demonstrated by the right-skewed distribution and the deviation of data points from the line of equality in the quantile-quantile plot (Q-Q plot) shown in Figure 3.1. A Q-Q plot is a graphical method to assess if a data set follows a certain theoretical distribution, such as a normal distribution [97]. It compares the quantiles of the data set against the quantiles of a theoretical distribution. If the data points fall along a straight line (line of equality), it suggests that the data are approximately normally distributed.

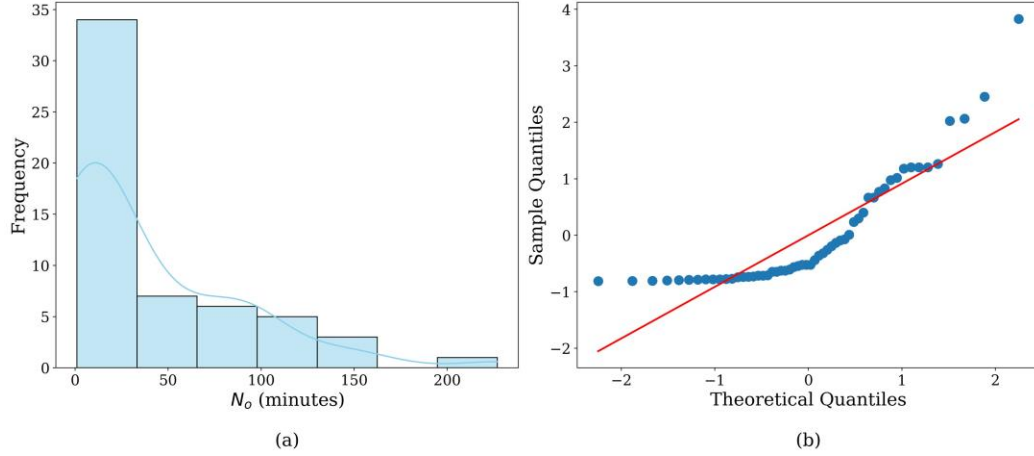


Figure 3.1. a) Data distributions for the incubation period, b) Q-Q plot for the incubation period

However, significant deviations from this line indicate that the data deviate from normality. Such a skewed distribution may present challenges in terms of ML model stability and convergence [97 – 100]. Therefore, an additional step is needed to resolve this issue. One potential solution is to perform data transformation to better handle the data distribution and enhance the stability and convergence necessary for applying ML.

3.2.3 Feature Importance and Correlation Matrix Analysis

Feature importance and correlation matrix analysis are crucial for understanding the relationships between input features and the target variable [101, 102]. Impact conditions (droplet size and impact speed), as well as material properties such as hardness, yield strength, fracture toughness, modulus of elasticity, modulus of resilience, and density all play significant roles in determining the material's response to WDE [3,26,34]. By analyzing the feature importance and correlations, it is possible to identify which features have the most significant impact on the incubation period and how they relate to each other. A feature importance plot, highlights the relative significance of various material properties, impact speed, and water droplet size, helping to understand which features exert the most substantial influence on the incubation (N_o) prediction model. The present model will employ Spearman's rank correlation and SHAP (SHapley Additive exPlanations) for feature importance analysis [103], as detailed below in the results section.

3.2.4 Incubation Period Prediction by Machine Learning

Different machine learning methods have been employed to predict the incubation period of the five materials under various impact conditions. These techniques comprise tree-based, linear regression-based, k-nearest neighbors, and neural network-based algorithms as shown in Figure 3.2.

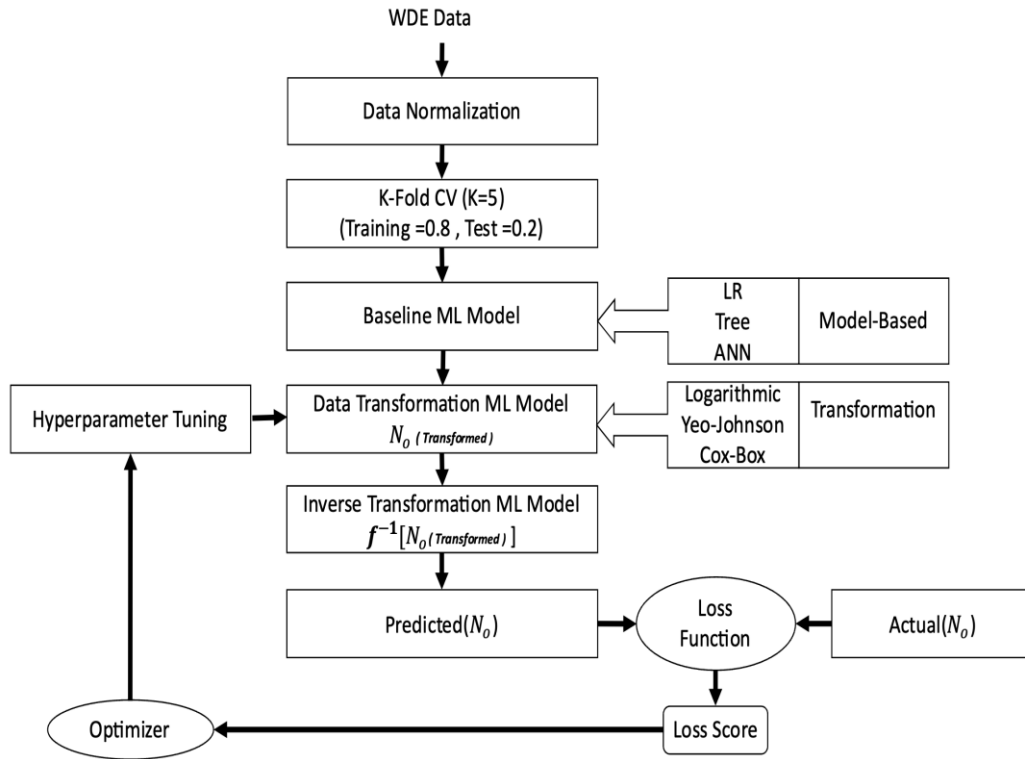


Figure 3.2. Workflow for predicting WDE incubation period

Figure 3.2 depicts a systematic approach for applying and improving ML models using TMG experimental data. The dataset is preprocessed through normalization in the first step, where the feature input values are scaled between 0 and 1. In the next step, a K-fold Cross-Validation (CV) method [104, 105] is implemented by splitting the data into five segments or subsets ($K = 5$), where 80% of the data is used for training and 20% for testing. During each iteration, one-fold serves as the validation set, while the remaining $K-1$ folds are used for the training set. Once all iterations

are complete, the average performance metrics across the folds provide a more generalized and reliable estimate of the model's performance, minimizing bias from any single data split. The procedure ensures that a model has been trained and tested on different datasets, yielding a more reliable estimate of its performance. It begins by creating a baseline model based on Linear Regression (LR), tree-based models, and Artificial Neural Networks (ANN). The linear regression-based models include linear (LR), Lasso, Ridge, and Elastic Net (EN) regressions. In contrast, tree-based models utilize Extreme Gradient Boosting Regressor (XGBoost), Gradient Boosting Regressor (GBR), Adaptive Boosting Regressor (AdaBoost), Light Gradient Boosting Machine (LightGBM), Decision Trees Regressor (DT), Extra Trees Regressor (ET), Random Forest Regressor (RF), and K-Nearest Neighbors Regressor (KNN). Additionally, the neural network regressors include the shallow neural network predictor [106, 107]. Transformations such as logarithmic, Yeo-Johnson, and Box-Cox are applied to adjust the data distribution, enhancing model performance—particularly for methods sensitive to skewness or outliers. The model is trained on the transformed dataset, and predictions are subsequently inverse-transformed to restore the original scale. This enables accurate comparison with the original data, facilitating the assessment of the proposed models' performance [106, 108]. The loss function value serves as a guide for the optimizer to iteratively update the model's hyperparameters until convergence [108]. Through this iterative cycle of training, evaluation, and optimization, the model's performance is refined, resulting in a robust predictive model.

3.2.5 Algorithm Selection Rationale

The selection of machine learning algorithms for this study was guided by several key considerations related to the specific challenges of predicting WDE incubation period employing a complementary algorithm strategy. We deliberately selected algorithms from different families of machine learning approaches to provide a comprehensive evaluation of modeling paradigms.

Linear Regression (LR) was included as a baseline model due to its interpretability and ability to capture fundamental physical relationships. Many tribological phenomena follow power-law relationships that can be linearized through appropriate transformations, making LR particularly valuable when combined with the transformation methods explored in this study. The coefficients in LR provide direct insights into the contribution of each feature to the incubation period,

enhancing our understanding of the underlying physical mechanisms. Whereas Decision Tree (DT) regressor was selected for its ability to model nonlinear relationships without assuming any specific functional form, which is crucial for complex tribological phenomena such as WDE. DT's splitting nature is well-suited to capture threshold behavior often observed in erosion mechanisms, where material response changes dramatically at certain values of impact velocity or hardness. Random Forest (RF) regressor was chosen as an ensemble method that reduces overfitting while maintaining the ability to capture complex non-linear relationships. RF provides robust feature importance metrics that help identify the most influential material properties and test conditions affecting WDE incubation. Additionally, RF performs well with the relatively small datasets typical in WDE testing, where experimental data collection is expensive and time-consuming. Gradient Boosting Regressor (GBR) was included for its sequential learning approach that focuses on correcting errors of previous models, making it particularly effective for capturing subtle patterns in WDE data that simpler models might miss. Recent literature [90] in tribology has shown ANN to be particularly effective for wear and erosion prediction tasks similar to WDE incubation period prediction. On the other hand, ANN were selected for their ability to identify complex, non-linear patterns in data and model intricate interactions between material properties and test conditions. The hidden layers in ANNs can represent complex relationships that might not be captured by other algorithms, providing a complementary modeling approach.

This algorithm selection spans a range of complexity levels, from the simplest (LR) to the most complex (ANN), allowing for an assessment of whether increased model complexity yields proportional improvements in predictive performance. The selection also balances highly interpretable models (LR, DT) with high-performance black-box models (RF, GBR, ANN), addressing both the need for physical insights and predictive accuracy in WDE incubation period prediction.

Nevertheless, these algorithms have some limitations too. Despite its interpretability, LR is limited by its assumption of linear relationships between features and the target variable. This may oversimplify the complex, non-linear physical processes involved in WDE, particularly at threshold conditions where erosion behavior changes dramatically. Additionally, LR is sensitive to outliers and may not capture interaction effects between material properties without explicit feature engineering. While effective at capturing threshold behaviors, DT is prone to overfitting,

especially with limited training data such as the case with WDE. Its hierarchical splitting structure can also create artificial discontinuities in predictions that may not reflect the continuous nature of physical erosion processes. Furthermore, small variations in the training data can lead to significantly different tree structures, resulting in prediction instability. Though RF mitigates some limitations of individual decision trees, it sacrifices model interpretability, making it challenging to extract clear physical insights from the ensemble. The averaging effect of multiple trees can also smooth out important threshold behaviors in erosion phenomena. Additionally, RF requires careful tuning of ensemble parameters to balance bias and variance, which can be challenging with limited WDE experimental data. GBR's sequential nature makes it computationally more intensive than other algorithms, potentially limiting its applicability in the industry. It is also highly sensitive to noisy data and outliers, which are common in experimental WDE measurements. The complex interaction of multiple weak learners further reduces model interpretability compared to simpler approaches. On the other hand, ANNs require substantial training data to generalize effectively, which is a significant limitation given the typically small datasets in WDE research. Their "black-box" nature makes it difficult to extract physical insights about erosion mechanisms. ANNs are also sensitive to hyperparameter selection and initialization conditions, potentially leading to inconsistent results without extensive tuning. Additionally, they risk memorizing rather than generalizing from training data when applied to small datasets. These limitations underscore the importance of our complementary algorithm approach, where the strengths of one model can compensate for the weaknesses of others, providing a more robust overall prediction framework for WDE incubation period.

3.2.6 Model Performance Evaluation

Model performance evaluation is crucial for assessing the effectiveness and accuracy of the predictive models. A common approach involves studying performance metrics such as (RMSE) Root Mean Squared Error, (MAE) Mean Absolute Error, R^2 (coefficient of determination), and the loss function (L). Table 3.4 summarizes each of these metrics in detail.

Table 3.4. Summary of machine learning model performance evaluation metrics

Metric	Description	Formula	Reference
Root Mean Squared Error (RMSE)	Measures the average magnitude of errors between actual values N_{oi} and the predicted values $\widehat{N}_{oi}$	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (N_{oi} - \widehat{N}_{oi})^2}$	[109]
Mean Absolute Error (MAE)	Calculates the average absolute difference between the actual and predicted values.	$MAE = \frac{1}{n} \sum_{i=1}^n N_{oi} - \widehat{N}_{oi} $	[109]
Coefficient of Determination (R^2)	R^2 is a statistical metric assessing a regression model's goodness-of-fit, indicating how well the model's predictions align with observed data points. R^2 values range from 0 to 1. $\bar{N}_o$ is the mean of the actual values N_{oi}	$R^2 = 1 - \frac{\sum_{i=1}^n (N_{oi} - \widehat{N}_{oi})^2}{\sum_{i=1}^n (N_{oi} - \bar{N}_o)^2}$	[109]
Loss Function (L)	Measures the mean squared error. It calculates the average of the squared differences between the actual values N_{oi} and the predicted values $\widehat{N}_{oi}$.	$L = \frac{1}{n} \sum_{i=1}^n (N_{oi} - \widehat{N}_{oi})^2$	[110]

Evaluating a model using these metrics provides valuable insights into its accuracy, precision, and ability to generalize to unseen data [109, 110]. This evaluation process is crucial for improving predictive model performance and ensuring reliability across multiple folds. In this study, a 5-fold cross-validation approach is employed. Consistent R^2 score values across folds suggest a stable model, while significant variations may indicate overfitting [104, 105].

3.2.7 Data Transformation

Data transformation is a technique used to modify data distributions, enhancing the accuracy of predictive models [97–100]. These transformations include methods such as logarithmic, Box-Cox, quantile, Yeo-Johnson, square root, and inverse transformations, all aimed at reducing skewness in the original data. This process can improve the stability of models and facilitate their

convergence during training [98–101]. Table 3.5 outlines the transformation techniques employed in this study.

Table 3.5. Summary of the different transformations employed on N_o in the present work

Transformation technique	Equations	Reference
Box-Cox	$N_{o(Transfomed)} = \begin{cases} \frac{(N_o^\lambda - 1)}{\lambda} & \text{if } \lambda \neq 0 \\ \log(N_o) & \text{if } \lambda = 0 \end{cases}$ λ is Cox – Box parameter $-5 \leq \lambda \leq 5$	[98]
Logarithmic	$N_{o(Transfomed)} = \text{Log}(N_o)$	[98]
Quantile	$Q_{N_o}(p) = F_{N_o}^{-1}(p)$ $Q_{N_o}(p)$: The quantile function $F_{N_o}(p)$: The cumulative distribution function (CDF) of the normal distribution $F_{N_o}(p) = \frac{1}{2} \left(1 + \text{erf} \left(\frac{p - \mu}{\sqrt{2}\sigma} \right) \right)$ erf: is the error function	[97,99]
Yeo-Johnson	$N_{o(Transfomed)} = \begin{cases} ((N_o + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \ y \geq 0 \\ \ln(N_o + 1) & \text{if } \lambda = 0 \ y \geq 0 \\ -(N_o + 1)^{(2-\lambda)} - 1)/(2 - \lambda) & \text{if } \lambda \neq 2 \ y < 0 \\ -\ln(-N_o + 1) & \text{if } \lambda = 2 \ y < 0 \end{cases}$	[99,100]
Square Root	$N_{o(Transfomed)} = \sqrt{N_o}$	[100]
Inverse	$N_{o(Transfomed)} = \frac{1}{N_o}$	[100]

Finally, after defining performance metrics and applying data transformations, hyperparameter tuning (e.g., regularization coefficient, learning rate, maximum depth of the tree, type of optimizer, etc.) was performed. This process involves iterative adjustment to the ML algorithm settings to optimize predictions with reduced errors [111, 112]. By fine-tuning the hyperparameters of the baseline model, the prediction results could be improved compared to the untuned (baseline) model results.

3.3 Results and Discussion

3.3.1 Correlation Matrix and Feature Impact

The correlation matrix illustrates the connections between the input features and the target variable [56]. Figure 3.3 shows the correlation matrix for 9 input features V_o , d , E , K_{IC} , σ_y , H_v , ρ , ν , C_s , and the output (N_o). In this study, Spearman's rank correlation, particularly suited for non-normally distributed and skewed data, was adopted to evaluate the relationships between variables [56]. To visualize these relationships, a Spearman correlation matrix heatmap [56] was utilized, providing a clear depiction of the strength and direction of correlations between various input features ($V_o, d, E, K_{IC}, \sigma_y, H_v, \rho, \nu$, and C_s) and the target variable (N_o). The Spearman correlation coefficient (denoted as r_s) range from -1 to 1, indicating the degree to which changes in one variable are associated with changes in another. As shown in Figure 3.3, the impact speed (V_o) exhibits a strong negative correlation with the incubation period ($r_s = -0.74$), suggesting that the incubation period significantly decreases as impact speed increases. Conversely, material properties such as yield strength (σ_y), hardness (H_v), and plane strain fracture toughness (K_{IC}) demonstrate weak positive correlations ($r_s = 0.14$). This indicates that while these properties have some influence on the incubation period, their effects are relatively minor compared to the more substantial impact of V_o .

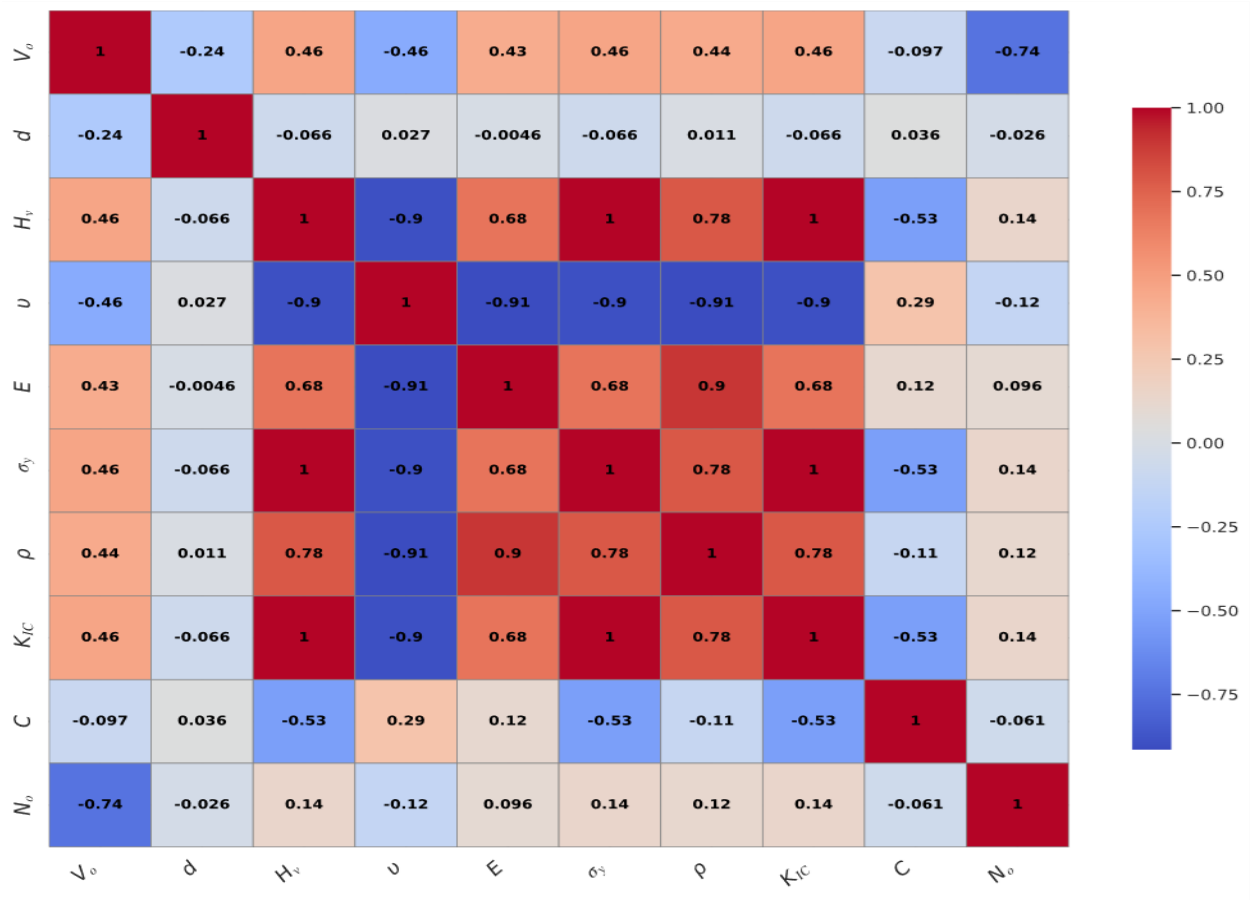


Figure 3.3. Spearman's rank correlation matrix

Furthermore, speed of sound (C_s) demonstrates only a minimal correlation with the incubation period (N_o), suggesting that this variable exerts a negligible direct influence on N_o . These observations imply that fluctuations in C are unlikely to significantly affect N_o , making it a less critical feature in the model. In contrast, when considered alongside the previously noted strong negative correlation between V_o and N_o , it becomes evident that certain features have a substantially more pronounced impact on the incubation period.

While the Spearman correlation matrix provides valuable insights into the relationships between variables, it also highlights the challenges in accurately predicting WDE. Spearman's rank correlation coefficient is effective for identifying qualitative relationships but offers limited predictive capability. Specifically, it can only capture monotonic relationships-whether linear or non-linear-between input and output. Although Spearman's coefficient is an excellent tool for describing qualitative trends, a transition to more quantitative approaches, such as feature

importance analysis, becomes necessary for developing a more accurate predictive model. In machine learning (ML), feature impact is a key concept that helps identify the input variables (features) exerting the highest influence on the output variable (target) [101].

To gain a detailed understanding of variable interactions, alternative techniques such as SHAP (SHapley Additive exPlanations) can be employed [103, 113]. SHAP values provide a quantitative assessment of each feature's impact on model's predictions, offering a numerical measure of how input variables influence outcomes. This method can uncover complex relationships that Spearman's correlation might not capture. In this context, the relevance of various features in predicting the incubation period was examined. The study evaluated the importance and impact (or relevance) of nine input features in determining the incubation period. Figure 3.4 presents a SHAP summary plot using mean absolute SHAP values to illustrate each feature's importance and how it influences the incubation period. In this analysis, impact speed emerges as the most influential factor, consistent with the literature [3, 14], exhibiting a mean absolute SHAP value approaching 32—significantly higher than any other feature. This finding aligns with physical expectations: higher velocities typically induce more severe material deformation, resulting in shorter incubation periods before damage or erosion.

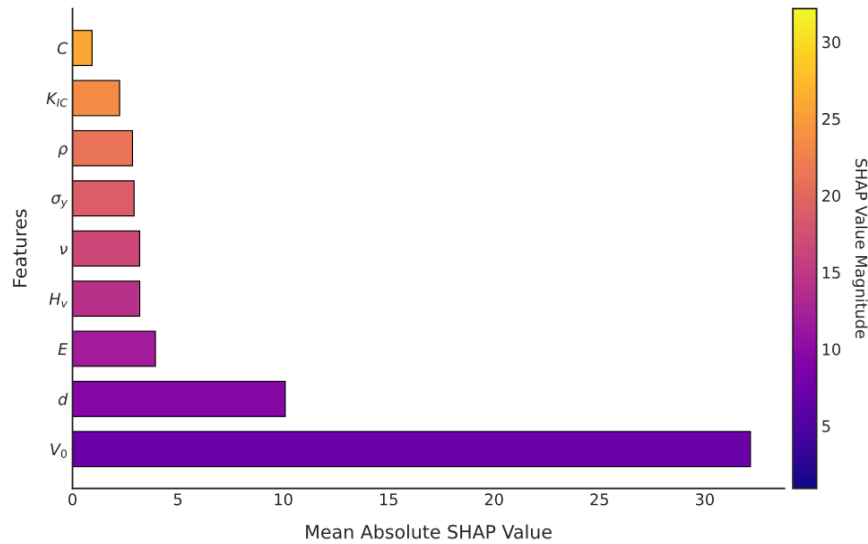


Figure 3.4. Features impact on N_0 using mean absolute SHAP values

In addition to impact speed (V_0), which is by far the most influential factor, the SHAP plot highlights droplet diameter (d) as the second most important feature in predicting N_o . Other features such as Poisson's ratio (ν), elastic modulus (E), and Vickers hardness (H_v) show moderate importance, indicating that material elasticity and resistance to deformation do play roles, though less dominantly. Properties such as density (ρ), yield strength (σ_y), fracture toughness (K_{IC}), and speed of sound (C_s) contribute minimally. Since the speed of sound, C_s , is mathematically dependent on both E , ν , and density ρ , its relatively low SHAP value in our plot might be due to redundancy or correlation with those features.

Overall, the SHAP analysis underscores the primary importance of droplet speed in the erosion process, highlighting that material properties—particularly hardness and yield strength—are essential for understanding a material's resistance to impact-induced damage.

3.3.2 Engineering Implications of SHAP Analysis

The high SHAP values for impact speed and hardness have significant engineering implications for designing erosion resistant components and systems. For impact speed, which showed the highest SHAP values in the current analysis, the engineering implications include energy transfer dynamics, component lifetime prediction, and protective measures prioritization. Also, the high SHAP value quantitatively confirms that velocity is the dominant parameter controlling the erosion process through kinetic energy transfer to the material surface. This suggests that even small reductions in impact velocity can significantly extend component lifetime, with potentially exponential benefits due to the non-linear relationship identified. Engineers can use the quantified SHAP values to develop more accurate lifetime prediction models for components exposed to water droplet erosion. For example, in turbine blade design, the current results indicate that a reduction in operational velocity could meaningfully extend the incubation period, as suggested by the dominant influence of velocity observed in the SHAP analysis. This indicates that velocity reduction (through deflectors, shields, or operational constraints) would yield the most benefit compared to other potential interventions. This provides clear guidance for engineers when designing protective systems where weight and/or space are limited. On the other hand, the high SHAP value for surface hardness has an implication on materials selection criteria too. It provides quantitative justification for prioritizing hardness in material selection for WDE-prone

applications. Our analysis suggests that increasing surface hardness could extend the incubation period, offering a potential strategy for improving erosion resistance. Therefore, engineers can focus on surface hardening treatments (such as carburizing or nitriding) as effective strategies to extend the incubation period. The SHAP values provide quantitative estimates of the expected improvements, allowing for cost–benefit analysis of different treatment options. Another suggestion that can be made based on these results is the use of composite materials. For engineered materials and composites, our results suggest that designing for increased surface hardness (even at the expense of other properties) may be a feasible strategy for WDE resistance in applications where operational water droplet impact speed cannot be reduced.

3.3.3 Baseline Models

The baseline models refer to the initial unoptimized ML models, which demonstrated moderate accuracy in predicting the incubation period. The overall performance of these initial models was assessed using metrics such as mean absolute error (MAE), root mean square error (RMSE), and R^2 metrics. Figure 3.5 shows the baseline prediction results for the incubation period using various algorithms including XGBoost, GBR, AdaBoost, LightGBM, DT, ET, RF, KNN, LR, Lasso, Ridge, EN, and ANN.

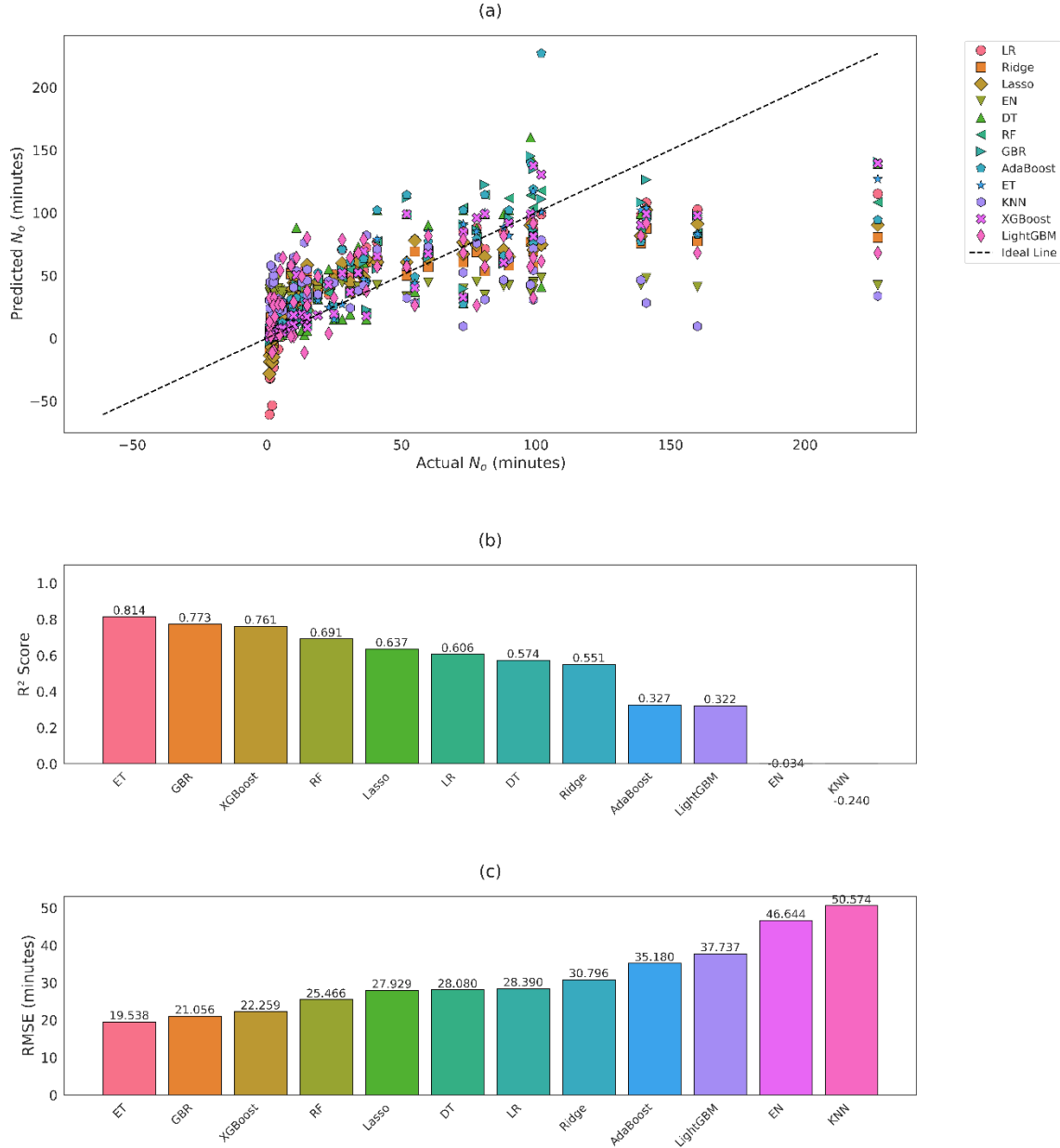


Figure 3.5. Baseline model predictions of the incubation period using various ML algorithms, b) R^2 values for each model, c) RMSE values for each model

As shown in Figure 3.5, the ET model demonstrated the best performance, achieving an R^2 value of 0.81, and an RMSE of 19.5. In contrast, KNN performed the poorest, with an R^2 value close to -0.24, and an RMSE of 50.57. To enhance the predictive capabilities of the models, other techniques, such as hyperparameter optimization and data transformation, will be explored.

3.3.4 ML Model Improvements

In this work we tried to improve the predictive capacity of the ML model through hyperparameter tuning and data transformation.

3.3.4.1 Model Hyperparameter Tuning

Hyperparameter optimization is crucial for improving ML performance [111, 114]. This section explores the significance of fine-tuning the hyperparameters of the different ML models [114, 115]. Hyperparameter tuning aims to find the optimal combination of configurations to enhance accuracy, reduce overfitting, and improve generalization [115–117]. Techniques such as grid search cross-validation and feature selection were employed to find the best hyperparameters for the algorithms [115, 116]. Table 3.6 summarizes the final tuned hyperparameters for all tested models, including k-nearest neighbors, linear, and tree-based algorithms. The listed hyperparameters, such as regularization strength (including L1 and L2 penalties), tree depth, number of estimators, learning rate, and neighbor selection, directly control model performance and learning behavior. These values were fine-tuned to balance model complexity and improve predictive accuracy in estimating the incubation period of WDE. Figure 3.6 illustrates the prediction outcomes and metric values for each algorithm. As shown, there was no notable improvement in model performance. Compared with the baseline models, the R^2 values for XGBoost and KNN increased from 0.76 to 0.81 and from -0.24 to 0.18, respectively. While Hyperparameter optimization did not improve the R^2 values for both DT and LR algorithms (remaining at around 0.60 for both the baseline and tuned models), hyperparameter tuning slightly enhanced the performance in some models but did not yield the desired overall improvements. Therefore, alternative approaches, such as data transformation, may be beneficial for further improving model performance. The next section will explore various data transformation techniques aimed at increasing data stability and optimizing distribution to enhance prediction accuracy.

Table 3.6. Tuned hyperparameters for predictive regression algorithms

Algorithm	Key Hyperparameters
Ridge Regression	alpha=0.001 (L2 Regularization strength)
Lasso Regression	alpha=0.001 (L1 Regularization strength)
ElasticNet	alpha=10, l1_ratio=0.1, max_iter=1000
Decision Tree Regressor	max_depth=5, min_samples_split=2, random_state=42
Random Forest Regressor	n_estimators=100, max_depth=5, min_samples_split=2
Gradient Boosting Regressor	n_estimators=200, learning_rate=0.2, max_depth=3
AdaBoost Regressor	n_estimators=100, learning_rate=0.1, random_state=42
Extra Trees Regressor	n_estimators=200, max_depth=10, min_samples_split=2
K-Nearest Neighbors	n_neighbors=3, algorithm='auto', weights='distance'
XGBoost Regressor	n_estimators=200, learning_rate=0.2, max_depth=3
LightGBM Regressor	n_estimators=50, learning_rate=0.01, num_leaves=31

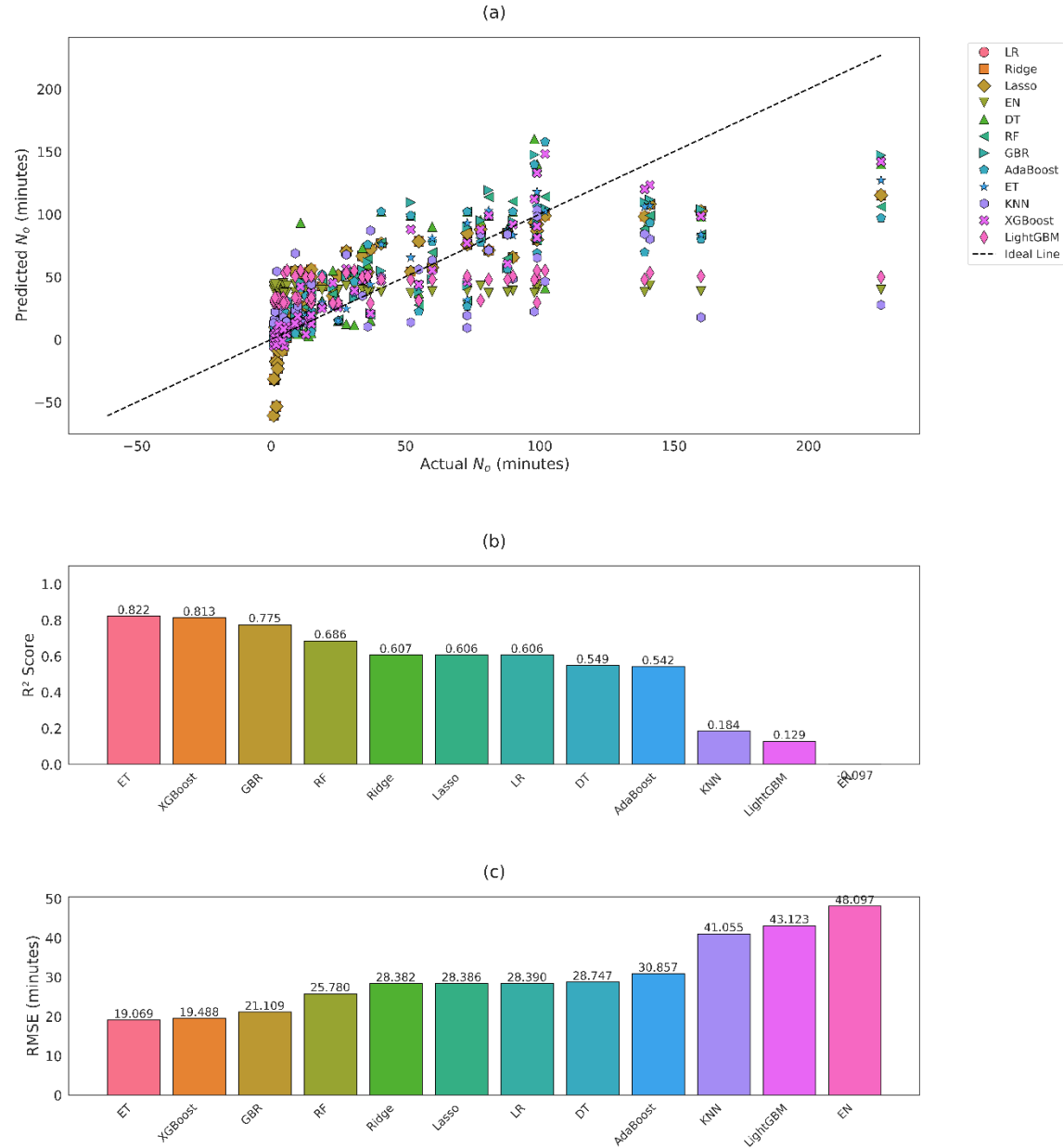


Figure 3.6. a) Model predictions after hyperparameters tuning using various machine learning algorithms, b) R^2 values for each tuned model, c) RMSE values for each tuned model

While hyperparameter tuning plays a critical role in optimizing model performance, its application in the context of WDE prediction presents certain limitations. The above-mentioned observation of 'no substantial improvement' from hyperparameter tuning deserves further explanation, as it represents an important finding rather than a limitation of this study. Several factors likely

contributed to this result such as data volume constraints, feature quality ceiling, and statistical data transformation effects.

Materials testing, especially for WDE, is often expensive and time-consuming, resulting in relatively small datasets compared to other machine learning domains. Due to the relatively small WDE data set used in this work, it is more likely that the ML model performance is data-limited rather than hyperparameter tuning-limited. Besides, the available features, while carefully selected based on domain knowledge, may not fully capture all aspects of the complex WDE phenomena. Our work for so many years on this problem indicates that there are still undiscovered materials attributes that contribute significantly to the occurrence of WDE. This creates a natural ceiling on model performance regardless of tuning efforts. The high R^2 values achieved with the base models (>90% for LR with Box-Cox, >85% for ANN with Yeo-Johnson) suggest that we may be approaching this ceiling.

Nevertheless, the data transformations applied (logarithmic, Box-Cox, Yeo-Johnson) may have already captured most of the potential model improvements, diminishing the additional benefit of hyperparameter tuning. This is particularly relevant for our study, where transformations were systematically evaluated and optimized before hyperparameter tuning.

These findings have important implications for future research in ML applications to materials science. Rather than extensive hyperparameter tuning, efforts might be better directed toward data augmentation strategies, such as physics-informed data generation, or hybrid modeling approaches that combine data-driven methods with physical models to overcome the limitations of pure machine learning approaches with limited data. Some of these aspects will be covered in a near future paper on this topic.

3.3.4.2 Model Improvement through Data Transformations

Applying transformation methods to the incubation period data enhanced the model's stability and convergence. These transformations addressed issues of skewness using techniques such as logarithmic transformation, Yeo-Johnson transformation, quantile transformation, Box-Cox transformation, square root, cubic root transformation, and inverse transformation [118-119]. Data transformation is crucial in stabilizing variance and improving data usability, which is essential

for many statistical techniques. The logarithmic transformation, for instance, reduces right skewness by compressing the higher values. Techniques such as Yeo-Johnson and Box-Cox transformations can adjust the data distribution to effectively address skewness by estimating the optimal transformation parameters. The quantile transformation explicitly maps the data to follow a normal distribution, while statistical rank transformations help mitigate the effect of outliers. After applying each transformation, the Q-Q plot is re-examined to evaluate the success of the transformation. Ideally, the transformed data Q-Q plot will show points closely aligned with a straight line, indicating a distribution closer to normality. Figure 3.7 illustrates the transformed data distributions on the left side, while the right side shows Q-Q plots. The logarithmic, Yeo-Johnson, Box-Cox, and quantile transformations are effective techniques for reducing skewness and making data more normally or uniformly distributed, as evidenced by the reduction in deviation from the straight line on the Q-Q plot. Implementing the previously mentioned data transformations improved model performance. Figures 3.8 to 3.10 illustrate the top-performing models resulting from Logarithmic, Box-Cox, Quantile, and Yeo-Johnson transformations, respectively.

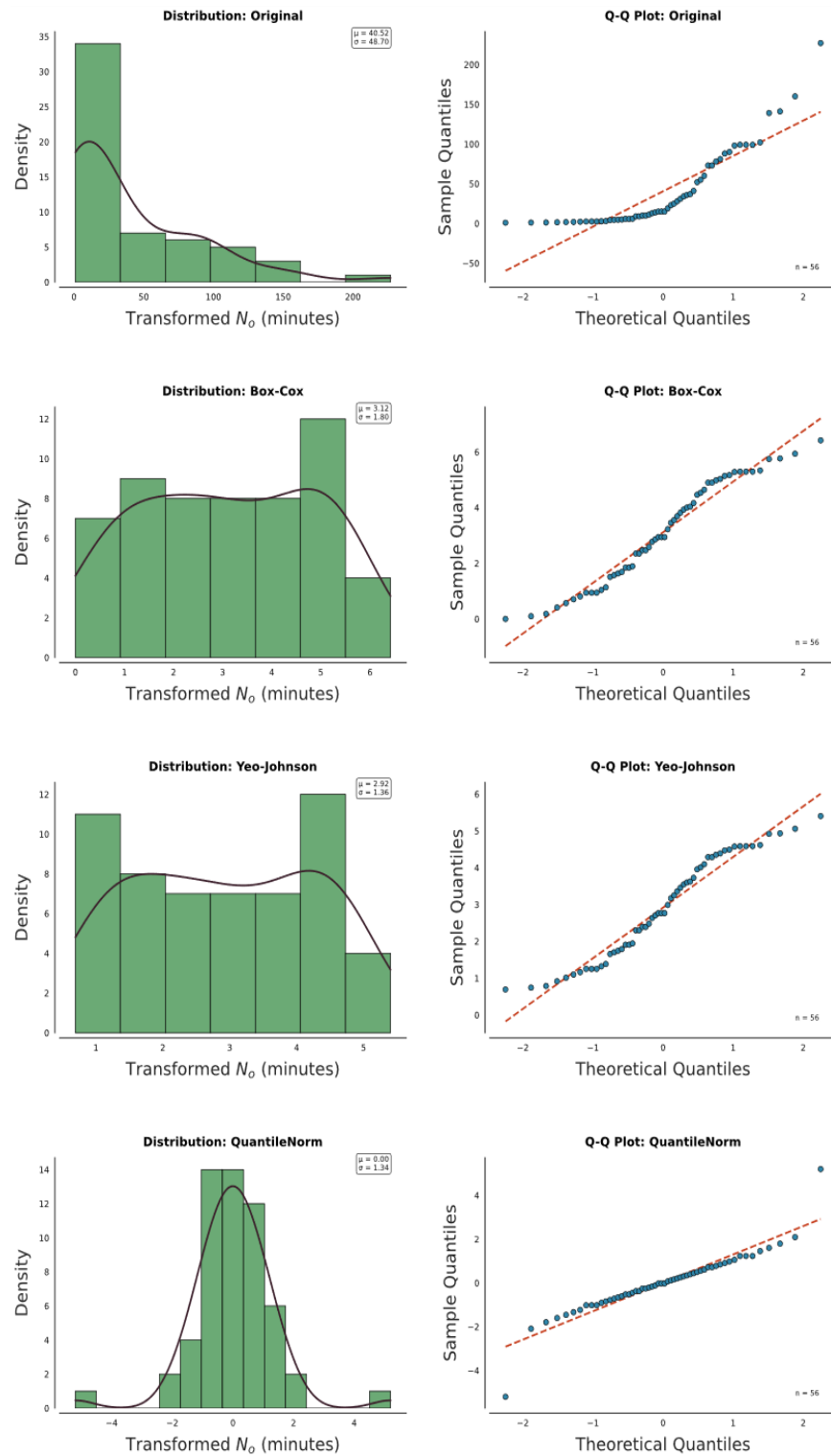


Figure 3.7 Transformed data distributions (Left) and Q-Q plot (Right) for the incubation period

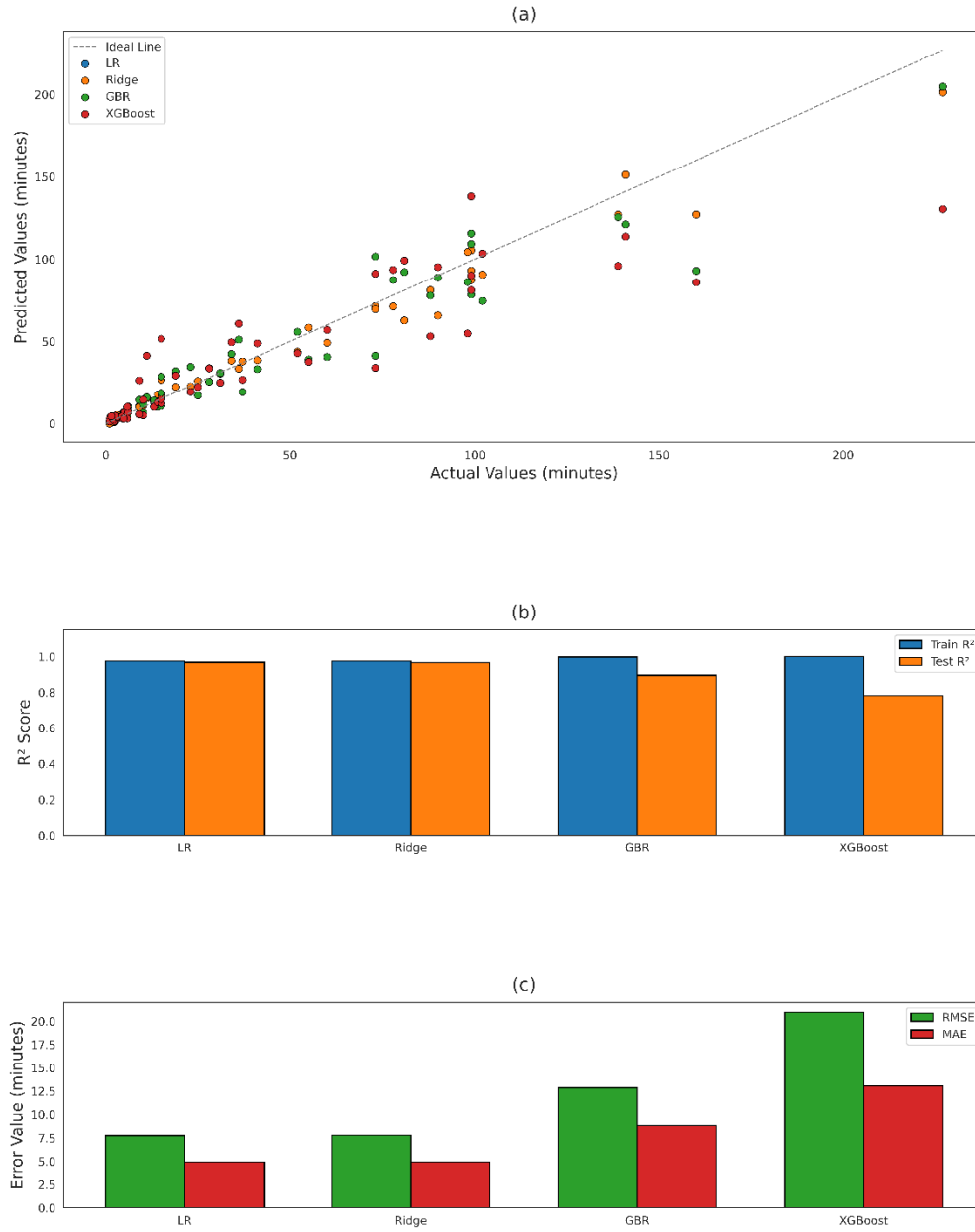


Figure 3.8. a) Model predictions with logarithmic transformations using various machine learning algorithms, b) R^2 values for each transformed model, c) RMSE and MAE values for each transformed model

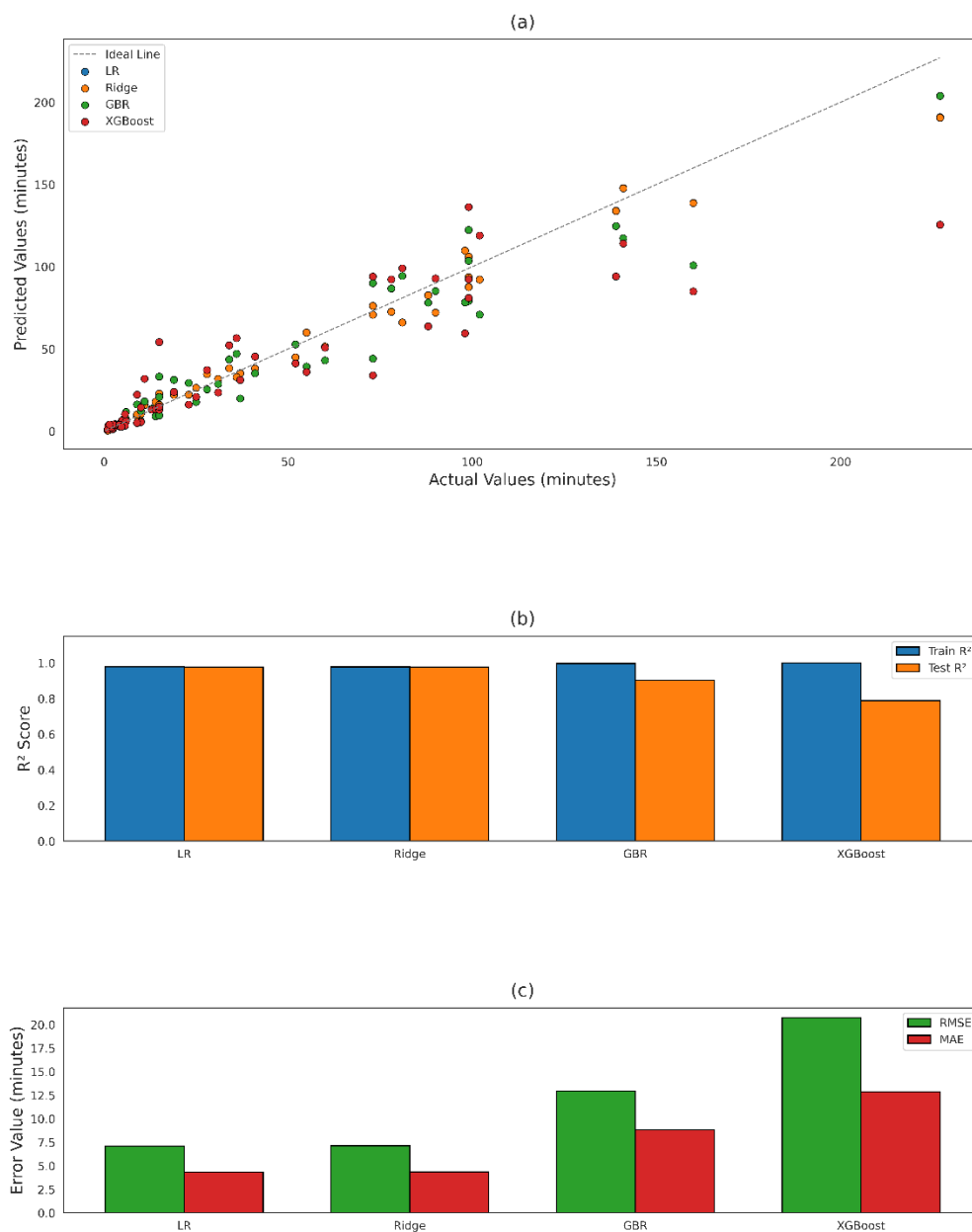


Figure 3.9. a) Model predictions with Box-Cox transformation using various machine learning algorithms, b) R^2 values for each transformed model, c) RMSE and MAE values for each transformed model

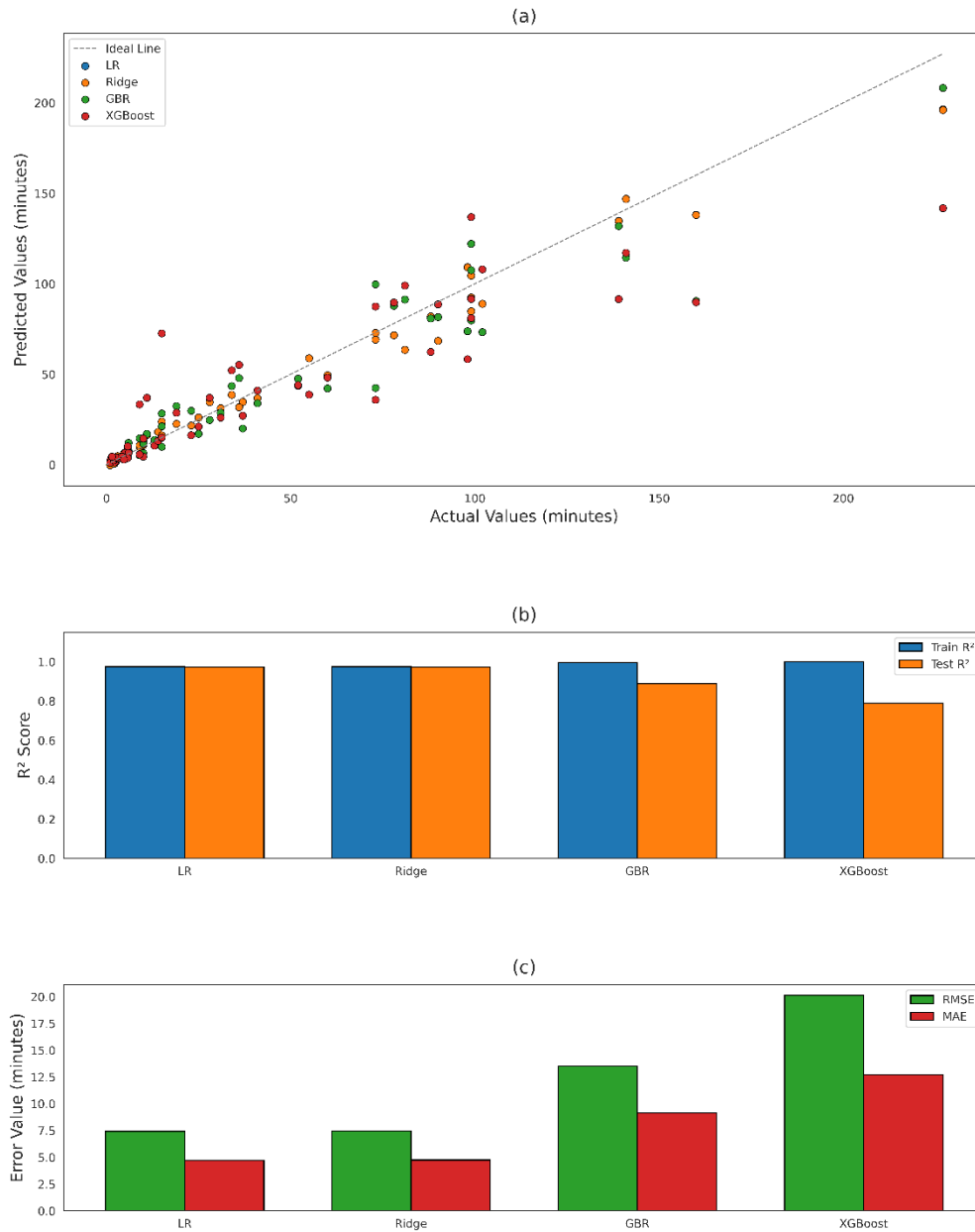


Figure 3.10. a) Model predictions with Yeo-Johnson transformation using various machine learning algorithms, b) R^2 values for each transformed model, c) RMSE and MAE values for each transformed model

Table 3.7 compares the best incubation period predictions using the proposed transformations with those of the original model, evaluated by RMSE, MAE, and R^2 . All performance metrics are derived from the average results of a five-fold cross-validation across ML models with the applied data transformations.

Table 3.7. Best ML transformed models' evaluation metrics from 5-fold cross-validation

Transformation	Algorithms	R^2 (Training)	R^2 (Testing)	ΔR^2	MAE (min)	RMSE (min)
Logarithmic	LR	0.98	0.97	0.01	4.80	7.62
Box-Cox	LR	0.98	0.98	0.00	4.32	7.10
Box-Cox	Ridge	0.98	0.98	0.00	4.35	7.11
Yeo-Johnson	LR	0.97	0.96	0.01	4.72	7.41
Quantile (Normal)	Ridge	0.87	0.74	0.13	13.25	21.98
Square Root	GBR	0.99	0.87	0.13	10.74	16.04
Inverse	XGBoost	0.97	0.62	0.35	18.94	24.99

After evaluating several prediction models using various transformation techniques on the data, the LR model with a logarithmic transformation performed well, achieving a value of 0.97, an RMSE of 7.62 minutes, and an MAE of 4.80 minutes. However, the LR model with a Box-Cox transformation yielded the best performance, with a value of 0.98 (as shown in Figure 3.11(b)), an RMSE of 7.10 minutes, and an MAE of 4.32 minutes. The LR ML model with a Yeo-Johnson transformation also performed well, achieving an RMSE of 7.41 minutes, an MAE of 4.72 minutes, and a value of 0.96. The square root transformation applied to the GBR model also showed

improvement, with a value of 0.87, an RMSE of 16.04 minutes, and an MAE of 10.74 minutes. Figures 3.11 compare the predicted versus actual outcomes for the baseline model with the Yeo-Johnson and Cox-Box transformations, confirming earlier findings about the performance of different transformations in predicting the incubation period.

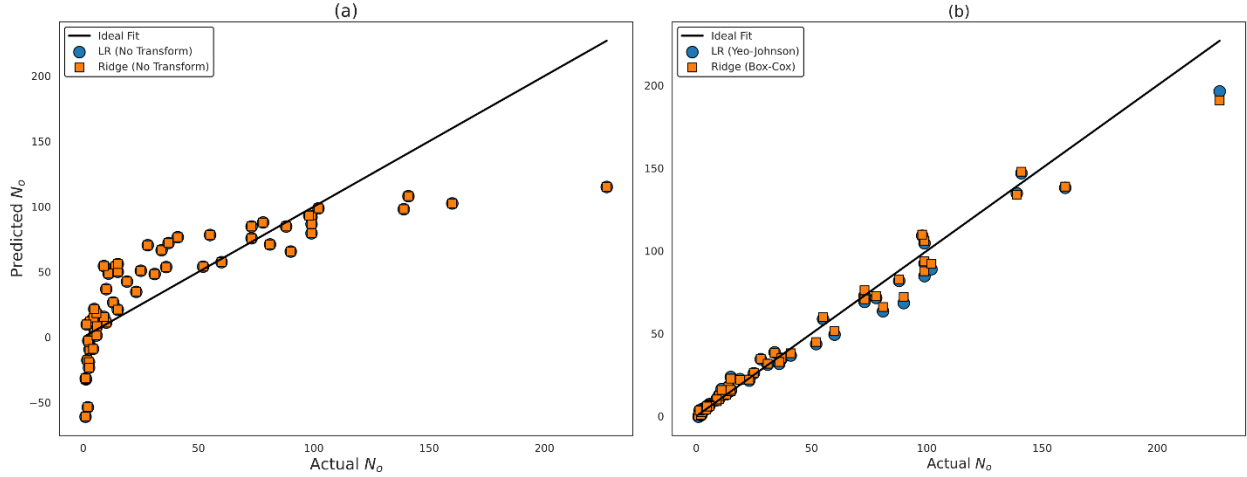


Figure 3.11. Comparison of actual vs. predicted incubation period: (a) without data transformation (LR and Ridge), (b) with data transformation using Ridge with Box-Cox and LR with Yeo-Johnson

To avoid overfitting, each algorithm was carefully configured, and regularization methods (such as L1 and L2 penalties) were applied where appropriate to discourage overly complex models. In addition to analyzing learning curves, the performance metrics—including R^2 , MAE—show consistent scores between the training and test sets, indicating that both overfitting and underfitting were effectively mitigated. To better illustrate these points, the cross-plot of N_0 using the LR model with Box-Cox transformation, shown in Figure 3.12, demonstrates strong agreement between predicted and actual values for both the training ($R^2 = 0.964$) and test set ($R^2 = 0.958$), with points closely aligned to the ideal fit line. These results further support the robustness and generalizability capability of the model.

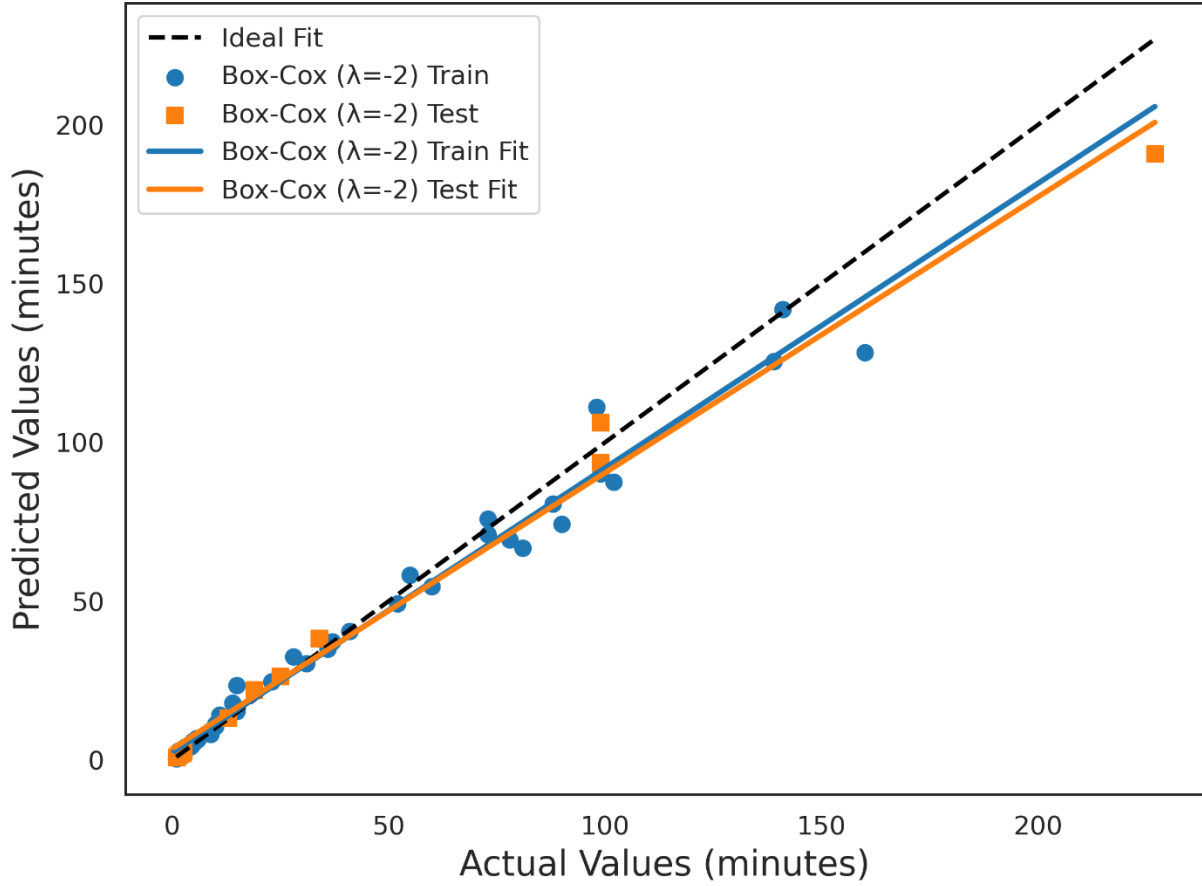


Figure 3.12. Predicted vs. actual values of N_0 for the LR model with Box-Cox transformation applied; the cross-plot for both training and test sets indicates accurate predictions and effective mitigation of overfitting

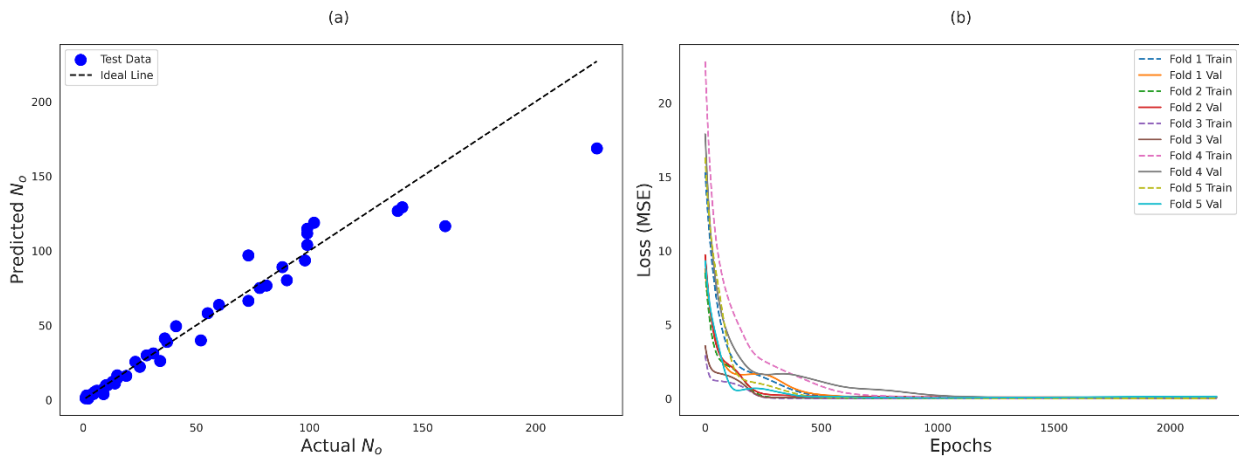
3.3.5 Artificial Neural Network

To enhance prediction accuracy, data transformations combined with hyperparameter tuning were applied to Artificial Neural Networks (ANNs). Specifically, a shallow neural network (SNN) was used, featuring two hidden layers with 9 and 5 neurons, respectively. By conducting a grid search, the network was fine-tuned using different optimization techniques. The Adam optimizer [120] with a sigmoid activation function was employed for the Yeo-Johnson and Box-Cox transformations, while the RMSprop optimizer [121] (also with a sigmoid activation function) was used for logarithmic transformation. In both cases, a learning rate of 0.003 was identified as optimal. The model performance was evaluated using RMSE, MAE, and R^2 to ensure accurate and reliable predictions. Table 3.8 presents the improvements achieved by applying hyperparameter tuning and data transformations in the SNN.

Table 3.8. ML model performance metrics using SNN

Transformation	Hyperparameters		
	Solver	Activation Function	Metrics
Yeo-Johnson	Adam	Sigmoid	$R^2 = 0.94$ RMSE=10.55 min
Box-Cox	Adam	Sigmoid	$R^2 = 0.93$ RMSE=12.05 min
Logarithmic	RMSprop	Sigmoid	$R^2 = 0.92$ RMSE=12.23min

Figure 3.13 illustrates the ANN's prediction results after applying data transformations and optimizing hyperparameters. In particular, the configuration with two hidden layers (9 and 5 neurons, respectively), a learning rate of 0.003, and the Yeo-Johnson transformation achieved the best results, yielding an RMSE of 10.55 min and an R^2 of 0.94. Also shown in Figure 3.13 are the predicted versus actual incubation period values and the corresponding loss function. The plot of the loss function over multiple epochs across five cross-validation folds, for both the training and validation datasets, provides insights into model performance. A well-tuned model typically exhibits a steadily decreasing loss that converges at relatively low values, indicating effective learning and a reduced likelihood of overfitting [120].

Figure 3.13. Yeo-Johnson transformation: a) Predictions of N_o using ANN and b) Loss function

Finally, to demonstrate the capability of ML approaches in predicting the incubation period for WDE, Figure 3.14 compares predictions from two of the best-performing models—a Linear Regression (LR) model with Box-Cox transformation and a Neural Network model with Yeo-Johnson transformation—against actual incubation period data for PH 17-4 stainless steel. Predictions from both models exhibit a high degree of accuracy, closely aligning with the actual data across the tested range. This highlights the potential of ML-based models to reliably predict incubation periods in metals.

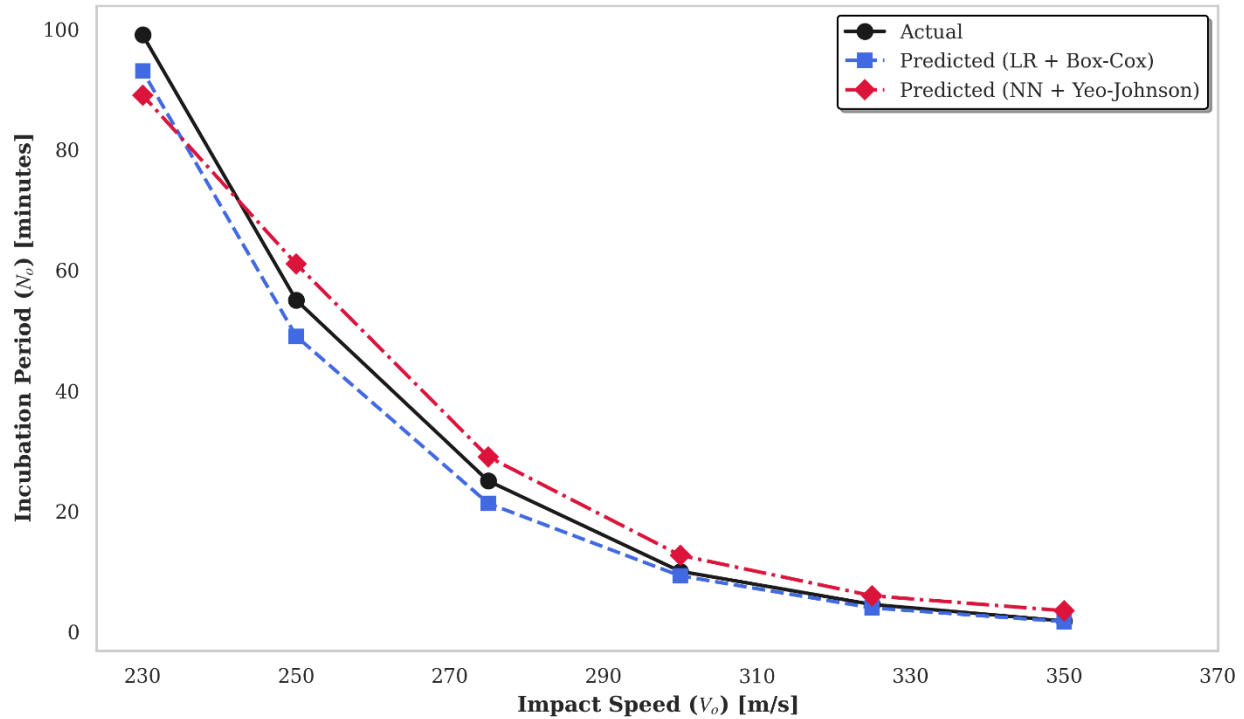


Figure 3.14 Actual vs predicted incubation period for PH 17-4 SS at $d=463 \mu\text{m}$, comparing ANN model and LR with Box-Cox transformation

In summary, the data transformations employed in this research have effectively enhanced ML model performance. Among the various transformations used, the logarithmic, Yeo-Johnson, and Box-Cox techniques stand out as the most successful. By adjusting skewed data distributions and reducing both variance and skewness, these methods facilitate more efficient convergence in the machine learning models.

The observed performance differences between transformation methods across different ML models can be explained from both statistical and physical perspectives. From a statistical viewpoint, Box-Cox transformation yielded the best results with LR ($R^2 > 90\%$) because it specifically optimizes for normality, which directly addresses the linear regression assumption of normally distributed residuals. Box-Cox is particularly effective at stabilizing variance across the data range, which is crucial for linear models that assume constant variance (homoscedasticity).

In contrast, Yeo-Johnson transformation performed best with ANN ($R^2 > 85\%$) due to its ability to handle both positive and negative values without requiring preprocessing or shifting. This flexibility aligns well with the inherent non-linearity of ANNs, allowing the model to maintain relationships between variables in a physically meaningful way while still benefiting from normalized inputs.

From a physical perspective, the WDE incubation period often exhibits power-law relationships with parameters such as impact velocity and material hardness. Box-Cox transformation, with its adjustable power parameter, can effectively capture these relationships, making it particularly suitable for regression models. Meanwhile, Yeo-Johnson's preservation of zero points and consistent treatment across the entire data range regardless of sign better supports the complex, non-linear interactions that ML model can handle.

A direct comparison between our ML approach and existing models for WDE incubation period prediction presents inherent challenges that warrant discussion. Despite extensive literature review, analytical models specifically for WDE incubation period could not be found in the literature, highlighting a significant gap that our work addresses. Existing empirical models, such as those proposed by [12, 14], are highly material-specific and testing rig dependent, with limited generalizability across different alloys and test conditions. These empirical models typically rely on power-law relationships with adjustable parameters that require recalibration for each new material or test configuration, limiting their predictive capability for novel materials. In contrast, our machine learning approach offers several advantages: (1) it captures complex, non-linear relationships between multiple material properties and test conditions without requiring predetermined functional forms; (2) it demonstrates superior generalizability across different materials as evidenced by the cross-validation results; (3) it provides quantitative insights into

feature importance through SHAP analysis, which empirical models typically cannot offer; and (4) it achieves higher prediction accuracy ($R^2 > 90\%$ for LR with Box-Cox transformation) compared to traditional empirical approaches (typically R^2 around 80% as reported in [12, 14]). This performance improvement underscores the potential of ML-based approaches to advance WDE prediction beyond the capabilities of conventional empirical modeling.

The reliability of our machine learning approach for WDE incubation period prediction is supported by several methodological strengths. First, our rigorous K-fold cross-validation methodology demonstrated consistent performance across multiple data partitions, with the LR model with Box-Cox transformation maintaining R^2 values above 90% across all folds. This consistency indicates strong generalizability to new data within the parameter space covered by our experiments. Second, the convergence of results across multiple algorithms provides algorithmic robustness, suggesting that our findings reflect genuine patterns in the WDE phenomenon rather than artifacts of a particular modeling approach. Third, the SHAP analysis revealed that the most influential factors in our models (velocity and hardness) align with established theoretical understanding of erosion mechanisms, providing physical validation of our data-driven approach.

3.4 Conclusion

This work presents a machine learning (ML) approach to predict the incubation period of water droplet erosion (WDE) for different metals and impact conditions, focusing on power generation and aerospace applications. The methodology outlined in this work presents a systematic approach to predict the incubation period of different alloys under water droplet erosion conditions, integrating experimental data collection, statistical analysis, and machine learning techniques. The study demonstrates that ML models can predict the incubation period accurately, especially those incorporating data transformations and hyperparameter optimization. The Linear Regression (LR) model with logarithmic or Box-Cox transformations achieved a coefficient of determination (R^2) of over 90% and low mean absolute error (MAE), while the Artificial Neural Network (ANN) model with Yeo-Johnson transformation exhibited an R^2 score of over 85% with similarly low MAE. These results underline the efficiency of ML in capturing the complex interactions between material properties and erosion conditions to make accurate predictions.

The present work highlights the importance of input feature impact analysis, which identifies the most significant parameters affecting the incubation period. This analysis shows that impact speed is the dominant factor influencing the incubation period in WDE, demonstrating a strong negative correlation in which higher velocities result in shorter incubation periods. Although other factors, such as yield strength and droplet diameter, show more complex relationships, the results emphasize the necessity of focusing on these critical variables to improve predictive model accuracy. These findings can inform material selection and maintenance strategies, ultimately leading to enhanced performance and durability in engineering applications.

The application of multiple data transformation techniques significantly improved the model's stability and predictive capabilities for incubation period data, with the Box-Cox transformation providing an especially accurate prediction model. Overall, these transformations effectively mitigated skewness, resulting in improved model accuracy as evidenced by superior performance metrics.

Furthermore, applying the Yeo-Johnson transformation to the ANN further improved prediction accuracy and model reliability. This underscores the importance of selecting suitable transformations to optimize ML models and boost their predictive capabilities.

Finally, this study provides a robust tool for predicting WDE, offering notable advantages over conventional and empirical models. By transforming skewed data to a more normal or uniform distribution, the stability and accuracy of ML models are notably improved. Future research could further refine these models and explore additional factors affecting WDE, thereby advancing the field of WDE prediction in materials science.

Acknowledgments: Hashemite University is acknowledged for sponsoring the first author's PhD studies. MT and MM gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC). Finally, the paper benefited from the use of large language models (LLMs) for assistance with language editing and polishing. All scientific content, data collection, analysis, and conclusions are the sole work of the authors.

Author contributions: All authors contributed to the study conception and design. **Khaled AlHammad:** Data curation, Formal analysis, Investigation, Methodology, Writing –original draft. **Mamoun Medraj** and **Moussa Tembely:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. All authors have read and approved the final manuscript.

Funding: Not applicable.

Data Availability: The datasets analyzed during the current study are publicly available in the GitHub repository at the following link: <https://github.com/khaledalhammad/WDE-DATA>

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests

Having established a strong foundation for predicting the incubation period of water droplet erosion (WDE) in Chapter 3, the next logical step is to extend the predictive framework to later stages of the erosion process—particularly the maximum erosion rate. Chapter 4 takes up this challenge by focusing on the second paper, which addresses the issue of limited experimental data using advanced data augmentation techniques. Methods such as SMOTER, Gaussian Process Sampling, and CTGAN are explored to synthetically expand the dataset and enhance model training. By integrating these techniques with machine learning algorithms such as k-nearest neighbors (KNN) and Artificial Neural Networks (ANN), the study aims to assess whether meaningful and reliable predictions of maximum erosion rate can be achieved despite data scarcity. Chapter 4 thus builds upon the methodologies developed in Chapter 3 and investigates how data augmentation can strengthen model performance and generalization in the context of WDE progression.

Chapter 4: Data-Augmented Learning for Estimating the Maximum Water Droplet Erosion Rate in Metallic Alloys

Khaled AlHammad, Mamoun Medraj, Moussa Tembely

Note. This chapter is based on a manuscript authored by K. AlHammad, M. Medraj, and M. Tembely, currently submitted for publication to the Journal of Engineering for Gas Turbines and Power

Abstract

Water droplet erosion (WDE) is a significant concern in aerospace, power generation, and other engineering applications, leading to material degradation over time. Accurate prediction of the maximum erosion rate (R_e) is crucial for optimizing material selection and maintenance strategies. Traditional empirical models often suffer from poor predictive accuracy due to their dependence on multiple parameters with a limited physical basis. This study develops a machine learning (ML) approach to predict R_e using a small dataset for training and testing, and employs 5-fold cross-validation to ensure robust model evaluation and prevent overfitting. In parallel, additional measurements were held out as an external validation dataset to assess the model's robustness and generalization. The experimental data were derived from tests on six metallic alloys widely used across aerospace and energy sectors. Two droplet sizes were used, with average diameters of 463 μm and 603 μm , and impact speeds (V_o) ranging from 150 m/s to 350 m/s. Initial predictions with base models—linear regression (LR), Lasso, Ridge, decision tree regressor (DT), random forest regressor (RF), adaptive boosting regressor (AdaBoost), k-nearest neighbors regressor (KNN), and artificial neural networks (ANN)—showed poor accuracy across all cross-validation folds. Despite applying hyperparameter tuning and various data transformations (logarithmic, Yeo-Johnson, and Box-Cox), these techniques yielded only marginal improvements in predictive accuracy. Hence, various data augmentation methods were implemented to generate additional synthetic WDE data, including Multivariate Interpolation, Noise Injection, Mixup, Synthetic Minority Over-sampling

Technique for Regression (SMOTER), Gaussian Process Sampling, and Conditional Tabular Generative Adversarial Network (CTGAN). Results demonstrate that ML models trained on augmented data significantly outperformed base models. Using SMOTER augmentation, the KNN achieved an R^2 score exceeding 95% with low RMSE on both training and testing data, maintaining consistent performance across all cross-validation folds. When evaluated on the unseen validation dataset, this model retained its high performance with an R^2 score over 90%, demonstrating excellent generalization capability. The ANN model enhanced with SMOTER augmentation demonstrated comparable results, achieving an R^2 score exceeding 95% on cross-validated training/testing, and on the external validation dataset. This independently validates the effectiveness of the data augmentation strategy, paving the way for faster material screening, better maintenance planning, and reduced experimental testing in applications facing the WDE challenge.

Keywords: Water droplet erosion, material degradation, machine learning, data augmentation, maximum erosion rate, prediction models

4.1 Introduction

Water Droplet Erosion (WDE) represents a critical challenge in industries such as aerospace and power generation, where surfaces are subjected to high-velocity water impacts [6, 122]. The impingement of water droplets generates localized high-pressure shock waves leading to gradual material degradation over time, as highlighted by Heymann [2]. This process underscores the necessity of addressing WDE to ensure material durability and operational reliability.

The WDE process unfolds through several distinct phases, each defined by a specific erosion rate. The process begins with the incubation phase, characterized by progressive surface roughening without substantial material loss. This is followed by the acceleration phase, then by the steady-state erosion phase, during which the maximum erosion rate (R_e) is maintained over an extended duration. The erosion rate then reduces during the deceleration phase, ultimately reaching the stable terminal phase [2, 3].

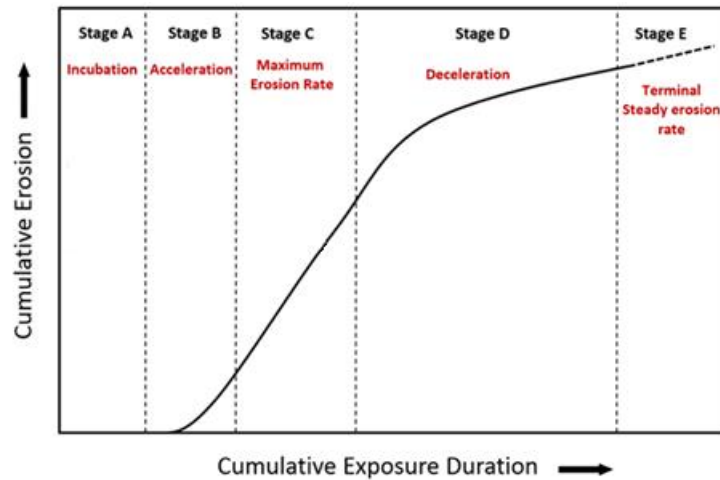


Figure 4.1. Typical stages in water droplet erosion damage curves [3]

A schematic representation of these stages is depicted in Figure 4.1, illustrating the temporal evolution of material damage and erosion rate. Among these stages, the incubation period and steady-state maximum erosion rate are the most critical [2, 3]. The incubation period reflects the material's resistance to repeated water droplet impacts. It marks the onset time of measurable erosion during WDE exposure, whereas the steady-state maximum erosion rate quantifies the long-term damage extent. Materials exhibiting a longer incubation period and a lower steady-state

erosion rate are considered more resistant to WDE. Accurate predictions of incubation time and maximum erosion rate are therefore essential for assessing and mitigating WDE effects. Consequently, most prior studies have focused on experimental investigations of these two stages due to their critical role in evaluating material performance under WDE conditions [3].

Decades of WDE research have involved extensive testing of metallic materials and qualitative analysis of experimental outcomes [17, 24]. Concurrently, predictive methodologies have been developed using analytical, semi-empirical, and empirical approaches. Lesser [17], for instance, employed an analytical model to estimate the impact pressure generated by liquid-droplet impact on both rigid and elastic solid targets. This approach modeled shock-wave propagation in the liquid while overlooking the transient response of the solid target. Various predictive approaches, including empirical and semi-empirical [12, 14] as well as semi-analytical models [26], attempt to predict erosion behavior from material properties and operating parameters. However, these models depend heavily on numerous adjustable parameters that lack clear physical significance, which limit their precision and general applicability. Heymann's widely referenced model [14], for example, offers equations to calculate the incubation time and the maximum erosion rate during the steady-state phase. Yet, its predictive capability is confined to specific materials such as titanium and certain steels under restricted conditions. These limitations underscore the need for more robust and generalizable predictive frameworks.

Experimental investigations, such as those by Anderson et al. [24], have explored WDE across various materials, including metals such as stainless steel 316 and aluminum 6061, alongside non-metallic materials, under controlled conditions. Nevertheless, their analysis was limited to single impact speed and angle, potentially failing to capture the variability of real-world scenarios. Recently, the threshold speed for initiating WDE damage in metals has been determined using a semi-analytical method that leverages impact conditions and material mechanical properties [26]. A fatigue-based analytical model [45], validated with experimental data for ferrous (stainless steel AISI 316) and nonferrous (aluminum 6061-T6) materials, showed strong agreement for impact velocities between 140 and 400 m/s. However, its scope is restricted to similar material classes. Mora et al. [46] utilized Smoothed Particle Hydrodynamics (SPH) and Finite Element Method (FEM) simulations to examine the influence of droplet size, velocity, and impact angle on erosion in Ti-6Al-4V. To reduce computational cost, they employed a relatively coarse mesh and

simplified constitutive/damage laws, which cannot fully resolve transient pressure and localized damage evolution; together with limited experimental validation, this constrains the study's generalizability [46].

Numerous interacting, nonlinear factors complicate the prediction of WDE. Conventional analytical and empirical models can incorporate only a subset of these variables and often rely on adjustable parameters with minimal physical basis. Machine learning (ML) techniques offer a promising alternative, capable of discerning intricate nonlinear patterns from datasets and generalizing across diverse materials and conditions. Applications of machine learning (ML) in materials science include predictions of wear loss, surface roughness [82], fatigue crack growth and fatigue life [83, 84], cavitation damage [21], and mechanical properties of composite materials [16]. In particular, artificial neural network (ANN) models have effectively estimated cavitation damage in stainless steels such as AISI 316L and AISI 420 [21]. Given the parallels between cavitation and erosion, both driven by complex, nonlinear damage mechanisms, ML techniques can substantially improve the accuracy of WDE predictions, particularly for the incubation period and maximum erosion rate.

Developing comprehensive WDE datasets requires controlled erosion rigs, precise measurements, and rigorous testing across diverse conditions and materials, which are costly and time-consuming. Data augmentation techniques offer a viable solution by synthetically expanding existing datasets, improving ML model training, enhancing generalization ability, and increasing predictive accuracy. Several data augmentation strategies are relevant to regression tasks in materials science and potentially applicable to WDE. These include Multivariate Interpolation, Noise Injection, Mixup, Synthetic Minority Oversampling Technique for Regression (SMOTER), Gaussian Sampling Processes (GP), and Conditional Tabular Generative Adversarial Networks (CTGANs). Each method generates synthetic data differently, aiming to increase the training dataset size and improve ML model performance when applied to complex physical phenomena such as WDE.

ML model performance generally improves with larger datasets, particularly for imbalanced and/or nonlinear problems [123]. To address dataset limitations, researchers have employed various data augmentation techniques. For example, Liu et al. [124] and Zhu et al. [125] utilized CTGANs and ensemble models to augment small datasets, thereby enhancing the performance of

deep convolutional neural networks. Multivariate Interpolation has proven valuable for augmenting sparse datasets resulting from resource-intensive experiments in materials science applications. Horňas et al. [126] demonstrated this approach when predicting the fatigue life of Ti-6Al-4V alloys fabricated by selective laser melting (SLM), highlighting its effectiveness in modeling complex nonlinear relationships between defect parameters (size, distance from the free surface, compactness, and sphericity), and fatigue life. Complementarily, Maharana et al. [127] demonstrated how Noise Injection reduces overfitting and improves model generalization in neural networks, especially for numerical data in regression tasks.

This study focuses on predicting the maximum erosion rate under diverse operating conditions, extending our previous work [128] that successfully applied machine learning to predict the incubation period in WDE with high accuracy, achieving an R^2 value exceeding 90%. Using experimental data from a controlled erosion rig to train, test, and externally validate the models, this research combines data augmentation and machine learning (ML) methodologies to overcome data limitations, improve prediction accuracy, and support material selection based on predicted maximum erosion rates. The comprehensive approach includes data collection and preprocessing, statistical analysis, hyperparameter optimization, multiple data augmentation strategies, and application of regression-based ML algorithms to achieve robust and accurate WDE prediction models.

4.2 Methodology

This study proposes a comprehensive methodology for predicting the maximum erosion rate (R_e) under water droplet erosion (WDE) that (i) curates high-quality WDE datasets and performs feature impact and relevance analysis, (ii) employs data augmentation techniques to mitigate scarcity and improve generalization, and (iii) evaluates various ML models for accurate R_e prediction. The methodology consists of the following steps:

1. Data Collection and Pre-processing: Collection of relevant datasets and preparation for analysis
2. Feature Impact and Correlation Analysis: Identification of key influential factors and quantification of their relationships.

3. Maximum Erosion Rate Prediction by Baseline ML: Application of ML models to predict the maximum erosion rate.
4. Data Augmentation for R_e Prediction: Generate synthetic samples to enrich the training dataset and improve the accuracy and reliability of maximum erosion-rate predictions, while ensuring no data leakage occurs. After data augmentation, perform the prediction of R_e using the enhanced dataset.
5. Model Selection, Tuning and Evaluation: Assessing the accuracy and reliability of the ML models, and improving performance through hyperparameters optimization.

The following sections provide a brief overview of each step.

4.2.1 Maximum Erosion Rate Data Acquisition and Preparation

The data for the present study were collected from experiments conducted using a rotating disk rig. A total of 67 experimental datapoints were collected and divided into 59 points for training and testing, and 8 points set aside for validation. The tests were performed on six (6) different alloys: 17-4 PH stainless steel, Ti-6Al-4V, Titanium Aluminide (TiAl), aluminum 2024-T4, aluminum 7075-T6, and 12% Cr stainless steel. Two droplet sizes were used, with average diameters of 463 μm and 603 μm , and the impact speed ranged from 150 m/s to 350 m/s.

A significant barrier to applying ML in WDE is the limited availability of experimental datasets. Collecting WDE data is inherently challenging due to the extensive experimental preparation required. For instance, plotting a single curve and obtaining one data point for the maximum erosion rate requires at least 10 experimental data points, as shown in Figure 4.2, which constitutes a highly time-intensive process. Consequently, acquiring 67 data points for the maximum erosion rate (R_e) necessitates at least 670 experimental runs, increasing the time and resource demands.

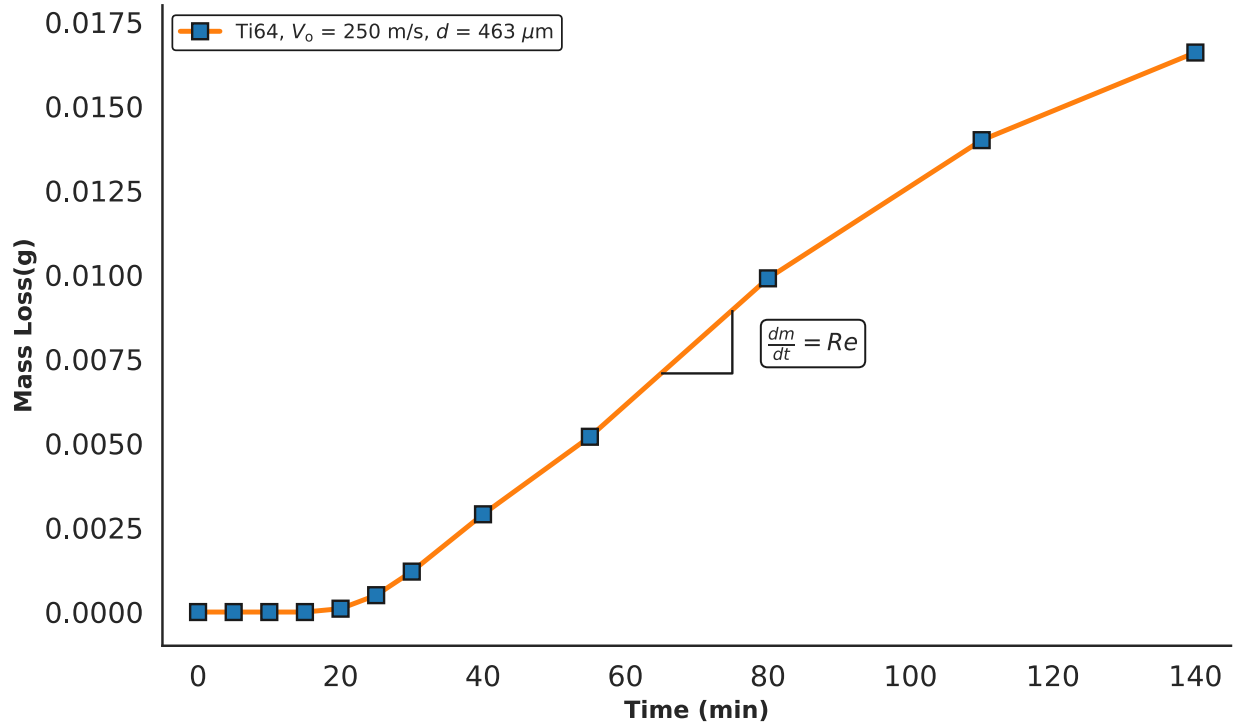


Figure 4.2. Water droplet erosion mass loss rate curve for Ti64 at $V_o = 250$ m/s, and $d = 463 \mu\text{m}$

To further characterize the dataset used in this study, Figure 4.3 provides a comprehensive overview of the ranges of impact velocities and corresponding maximum erosion rate (R_e) values obtained from the WDE experiments conducted on six different materials using two droplet diameters ($463 \mu\text{m}$ and $603 \mu\text{m}$). This figure illustrates the diversity and distribution of the collected data, which are essential for developing reliable predictive models.



Figure 4.3. a) Impact speed ranges for the tested materials at two different droplet sizes, b) Maximum erosion rates for different materials and droplet sizes

Specifically, Figure 4.3(a) illustrates that the impact velocities required to initiate measurable erosion vary with material type and droplet size. Smaller droplets (463 μm) generally require higher impact speeds to cause erosion, whereas larger droplets (603 μm) induce erosion at comparatively lower velocities. This behavior can be attributed to the greater kinetic energy carried by larger droplets, even at moderate speeds, which accelerates the erosion process.

Meanwhile, Figure 4.3(b) presents the maximum erosion rate (R_e) observed under various test conditions. The data reveal distinct material-dependent erosion responses: aluminum alloys, such as Al2024-T4 and Al7075-T6, exhibit significantly higher erosion rates, whereas stainless steels, like 17-4 PH SS and 12%Cr SS, demonstrate enhanced resistance. The TiAl alloy shows intermediate erosion behavior, with rates falling between those of the aluminum alloys and stainless steels. These material-specific differences highlight the necessity of integrating material properties, droplet size, and impact velocity into data-driven models.

These observations provide crucial insights into the experimental domain, forming a robust foundation for the machine learning methodologies employed later in this study to predict erosion behavior with greater efficiency and accuracy.

These alloys are extensively used in industries where water droplet erosion (WDE) significantly contributes to material degradation [2, 7, 27, 129]. Notable examples include 12% Cr stainless steel and 17-4 PH stainless steel, which are frequently employed in steam turbine blades and aerospace components. Similarly, Ti-6Al-4V is commonly used in gas turbine compressor blades,

while aluminum alloys such as 2024-T4 and 7075-T6 are widely applied in aircraft structures. The selection of these materials was guided by their varied mechanical properties, including elastic modulus (E), plane strain fracture toughness (K_{IC}), yield strength (σ_y), Vickers hardness (H_v), density (ρ), Poisson's ratio (ν), and speed of sound (C). Hardness and yield strength, in particular, indicate the material ability to resist deformation and fracture, which relates to its capacity to withstand impact. These properties are crucial in determining how materials behave under high-speed droplet impacts, influencing their erosion resistance and long-term durability under various operating conditions [2, 3, 26].

By analyzing these factors, we can better understand their influence on maximum erosion rates and validate the predictive model developed in this research [2, 3, 24]. Such insights will ultimately improve material performance and reliability in real-world applications. Figure 4.4 outlines the mechanical properties of the tested material [91–94].

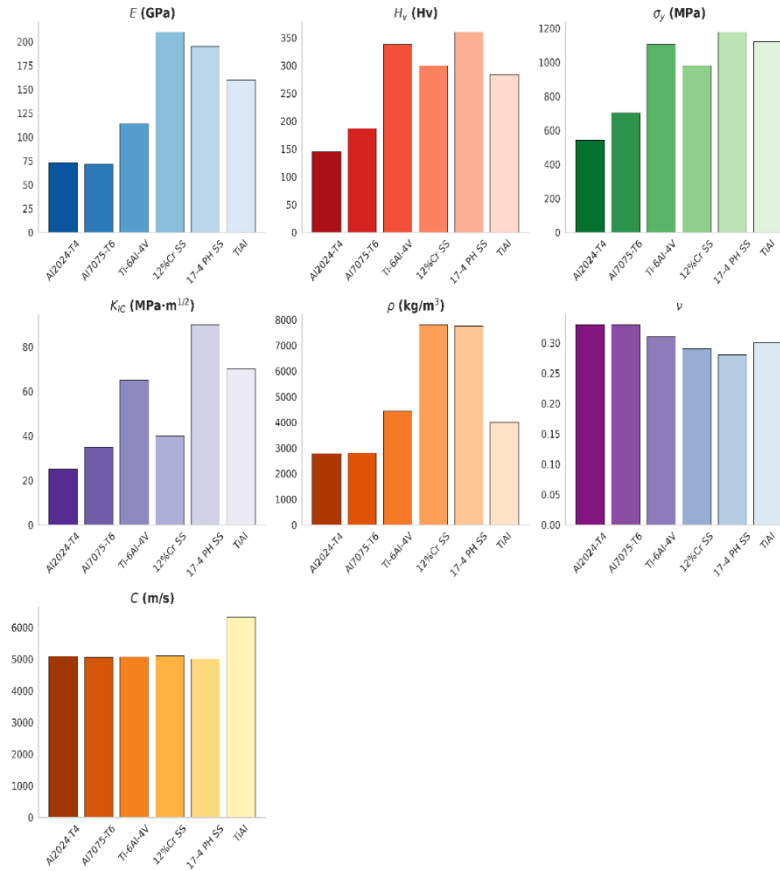


Figure 4.4. Mechanical properties of the tested materials

4.2.2 Analysis of Input Features, Contributions and Correlations

The influence of impact conditions, such as droplet size and velocity, alongside material properties including hardness, yield strength, fracture toughness, elastic modulus, speed of sound, Poisson's ratio, and density, profoundly affects a material's response to water droplet erosion (WDE) [2, 26, 34]. Explaining the relationships between these input parameters and the target variable, maximum erosion rate (R_e), is essential for accurately characterizing WDE behavior [100, 101]. To this end, the study employs correlation analysis and feature importance evaluation [100]. Specifically, a Spearman's rank correlation matrix is utilized to assess monotonic relationships between variables, with outcomes visualized in a heatmap, where correlation coefficients range from -1 to +1, reflecting the strength and direction of each association [130]. While useful for identifying linear and monotonic trends, this method does not capture complex, non-linear, or predictive interactions. However, it cannot capture complex or predictive interactions as a qualitative tool.

To address this, the study employs Feature Impact Analysis using SHAP (SHapley Additive exPlanations), which quantitatively measures each feature's contribution to the output [113, 131]. SHAP provides deeper insight into feature importance and interactions, which are often overlooked by Spearman's method.

4.2.3 Maximum Erosion Rate Prediction by Machine Learning

Machine learning (ML) models were developed to predict the maximum erosion rate of WDE. Various ML methods have been employed to predict the maximum erosion rate of the six selected materials under different impact conditions. These methods include k-nearest neighbors, Tree-based, linear regression-based, and Neural Network-based algorithms, as illustrated in Figure 4.5.

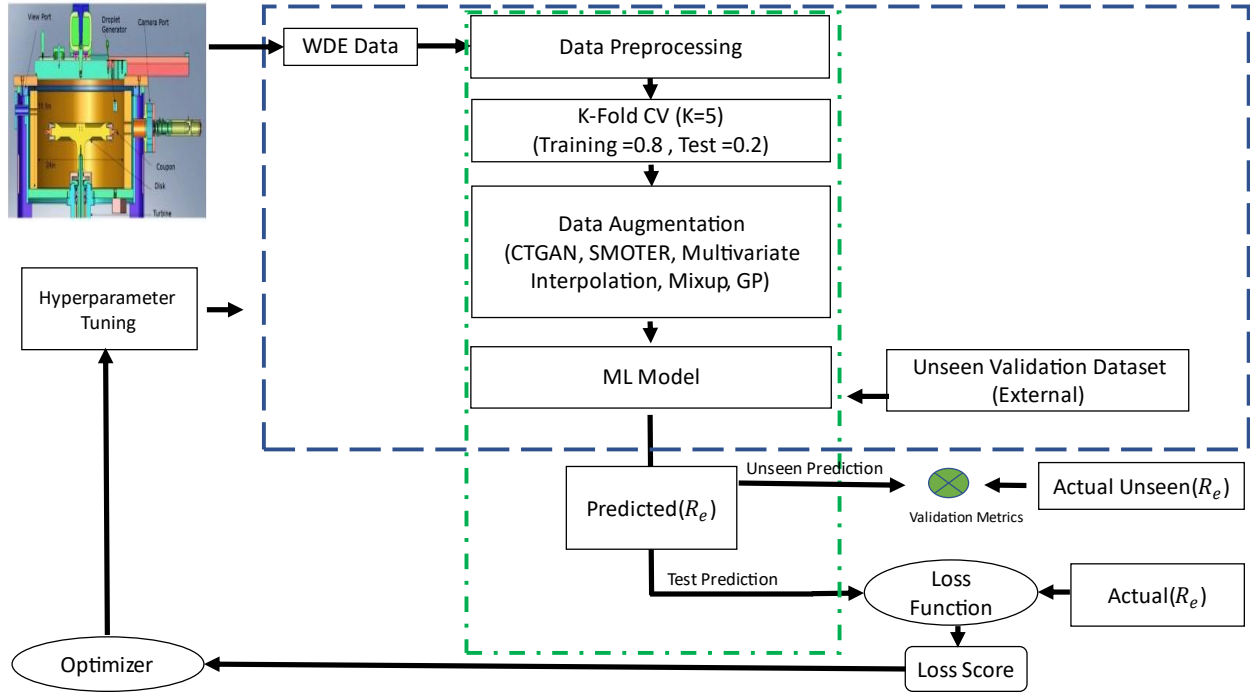


Figure 4.5. Workflow for predicting WDE maximum erosion rate

The flowchart in Figure 4.5 outlines a systematic approach for developing and refining machine learning models using the experimental WDE data. The first step involves normalizing the data to ensure all features are on a similar scale. This is crucial because many machine learning algorithms perform better when no single feature dominates due to its higher values [132]. To ensure model robustness and generalizability, a 5-fold cross-validation (K-fold CV) strategy is adopted [105, 133, 134]. In this scheme, the dataset is partitioned into five equal subsets, where each subset serves as the test set once, while the remaining four subsets are used for training. This yields an 80:20 train-test split in each iteration, enabling thorough model evaluation across different data partitions.

To address the challenge of limited data in predicting the maximum erosion rate (R_e) for water droplet erosion (WDE), a suite of machine learning models is trained, including linear regression, k-nearest neighbors, tree-based algorithms, and artificial neural networks (ANNs). These models are implemented using two parallel approaches: a baseline model and a transformation-based model. In the baseline model approach, algorithms are applied to normalized features without

further modification. Conversely, the transformation-based approach applies transformations such as logarithmic, Yeo-Johnson, or Box-Cox to the target variable R_e to stabilize variance and approximate normality, potentially enhancing regression performance. Post-training, an inverse transformation restores predicted outcomes to the original R_e scale. However, these transformations yield only marginal improvements in model accuracy.

Given the limited dataset size, which may constrain model performance and exacerbate issues such as data imbalance, data augmentation techniques are employed to enrich the training data. The augmented datasets are used to retrain the machine learning models to enhance predictive accuracy and robustness in characterizing WDE behavior.

Hyperparameter tuning was conducted to optimize the model to predict the maximum erosion rate (R_e) in WDE. During training, model parameters were iteratively adjusted to minimize the loss function, defined as the Root Mean Square Error (RMSE), which quantifies the discrepancy between predicted and actual R_e values. The iterative process continued until convergence, while the coefficient of determination (R^2) was monitored to evaluate predictive performance. High RMSE values indicated challenges in achieving optimal predictive accuracy, highlighting the need for further refinement of both the model and the data augmentation strategies.

Finally, the predictive model's performance was rigorously evaluated using quantitative metrics. The R^2 score of predicted versus actual values for the R_e was employed to assess accuracy and to visually identify any systematic deviations. This comprehensive methodological framework, encompassing advanced data preprocessing, transformation, augmentation, and hyperparameter optimization, provides a robust strategy for achieving high-fidelity predictions of erosion behavior.

Quantitative evaluation plays a critical role in validating model performance. To this end, several performance metrics are employed to measure the model's accuracy, reliability, and generalization capability. These include the Root Mean Squared Error (RMSE) [109], Mean Absolute Error (MAE) [109], Coefficient of Determination (R^2) [109], and the Loss Function (L) [110], defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (R_{e_i} - \widehat{R}_{e_i})^2} \quad (4.1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |R_{e_i} - \widehat{R}_{e_i}| \quad (4.2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (R_{e_i} - \widehat{R}_{e_i})^2}{\sum_{i=1}^n (R_{e_i} - \bar{R}_e)^2} \quad (4.3)$$

$$L = \frac{1}{n} \sum_{i=1}^n (R_{e_i} - \widehat{R}_{e_i})^2 \quad (4.4)$$

Where n is the number of data points, R_{e_i} is the actual value i th data point, $\widehat{R}_{e_i}$ is the predicted value i th data point, and $\bar{R}_e$ is the mean of the actual values R_{e_i} .

Evaluating these metrics helps assess the model's accuracy, precision, and generalization ability on test data.

4.2.4 Data Augmentation Methods in Machine Learning

Data augmentation is a machine learning technique used to generate synthetic data from existing data, enhancing the accuracy of model predictions [61, 62, 135–141]. It is beneficial when the training dataset is limited or insufficient [62, 136–138]. Expanding the dataset's size and diversity reduces the risk of model overfitting, thereby improving accuracy and robustness. Various studies have successfully implemented data augmentation, demonstrating its effectiveness in different applications [62, 138–140]. Techniques such as Interpolation, Generative Adversarial Networks (GANs), Noise Injection (Gaussian noise), Synthetic Minority Over-sampling Technique for Regression (SMOTER), Mixup, and Gaussian Process Sampling are commonly employed to improve the convergence of machine learning models [61, 139, 140].

This section provides an overview of various data augmentation methods for the regression task, their applications, and data analysis.

4.2.4.1 Interpolation-Based Methods

Interpolation-based methods generate synthetic data by estimating values between existing data points. These techniques benefit small datasets with clear trends and effectively preserve relationships between numerical features. Standard interpolation techniques include linear,

polynomial, and spline interpolation [141]. These methods are simple to implement and preserve trends in the data, making them suitable for regression tasks and tabular data with numerical target variables [142]. However, they assume a specific type of relationship and are limited to numerical data.

4.2.4.2 Noise Injection

Noise Injection is a data augmentation technique where random variations, such as Gaussian noise, are added to existing data points to generate synthetic datasets [61]. Introducing slight randomness into the training data helps improve the model's robustness and generalization ability. It is instrumental in preventing overfitting, especially when working with limited data. The application of Noise Injection can enhance model performance; however, determining the optimal noise type and level remains a key challenge.

4.2.4.3 Over-Sampling Techniques

Oversampling techniques, such as the Synthetic Minority Over-sampling Technique for Regression (SMOTER) [136], extend the traditional Synthetic Minority Over-sampling Technique (SMOTE) algorithm to address challenges posed by imbalanced datasets in regression tasks. Unlike the original SMOTE, which is designed for classification problems, SMOTER is specifically developed for regression scenarios where the target variable is continuous. SMOTER balances the distribution of the target variable by generating synthetic data in sparse regions using interpolation methods. While it enhances model performance and reduces prediction bias, it can also introduce noise, requires careful parameter tuning, and may struggle with complex non-linear relationships. Despite these limitations, SMOTER remains a useful technique for handling imbalanced regression datasets and improving model accuracy.

4.2.4.4 Deep Learning-Based Generative Models

Deep learning-based generative models such as Conditional Tabular Generative Adversarial Network (CTGAN) [143] help address imbalanced structured tabular data by creating synthetic data points. CTGAN consists of two neural networks, a generator and a discriminator that train together. The generator produces synthetic data, while the discriminator learns to differentiate between real and synthetic data. Over time, the generator improves, generating data that closely

mimics real samples. The synthetic data retains the statistical characteristics of the original dataset, balancing the data and enhancing the accuracy of predictive models across various scenarios.

Similarly, Mixup for Regression [62] is a deep learning-based generative technique that enhances regression model performance by linearly combining input data points and their associated labels. It works by linearly combining existing data points to create new training data. This technique enhances model generalization, reducing overfitting and improving predictive accuracy and robustness.

4.2.4.5 Hybrid Methods

Hybrid methods combine techniques, such as interpolation and Noise Injection, to generate synthetic data. These methods are flexible and adaptable to different scenarios, making them suitable for complex datasets. However, they are more complicated to implement and tune [111]. Table 4.1 summarizes the regression data augmentation methods discussed in the previous subsections. Each method is evaluated based on its applicability, advantages, and potential drawbacks. It provides insights into its suitability for various tasks, including classification, regression, handling of imbalanced datasets, and generating complex data.

Table 4.1. Summary of different data augmentation methods

Method	Type	Use Cases	Strengths	Limitations
Mixup	Deep learning-based	Classification, regression	Improves generalization, simple	Label ambiguity, hyperparameter tuning
Interpolation-Based Methods	Traditional	Small datasets with trends	Simple, preserves trends	Limited to numerical data
SMOTER	Over-sampling	Imbalanced classification	Balances class distribution	Limited to numerical data
Noise Injection	Traditional	Robustness in regression, image tasks	Simple, adds variability	Can distort data
CTGANs	Deep learning-based	Complex data generation	High-quality synthetic data	Computationally expensive
Hybrid Methods	Combination of techniques	Complex datasets	Flexible, adaptable	More complex to implement

Finally, after applying data transformation, hyperparameter tuning -which involves iterative adjustments- was used to optimize model performance [114, 144]. Hyperparameter tuning searches for optimal configurations that maximize accuracy while minimizing overfitting and improving generalization. This study employs hyperparameter tuning with data augmentation techniques to enhance the baseline model's accuracy and predictions.

4.3. Results and Discussion

4.3.1 Exploratory Data Analysis of Maximum Erosion Rate

To guide the selection of relevant input features for predicting the maximum erosion rate (R_e), a Spearman correlation matrix is computed first. This matrix includes nine input features—impact speed (V_o), droplet size (d), elastic modulus (E), fracture toughness (K_{IC}), yield strength (σ_y), hardness (H_v), density (ρ), Poisson's ratio (ν), and speed of sound (C)—along with the output variable, maximum erosion rate (R_e). Spearman's is chosen because it captures monotonic relationships and is robust to outliers and nonlinearity, traits typical of erosion datasets. The correlation matrix is depicted in Figure 4.6(a).

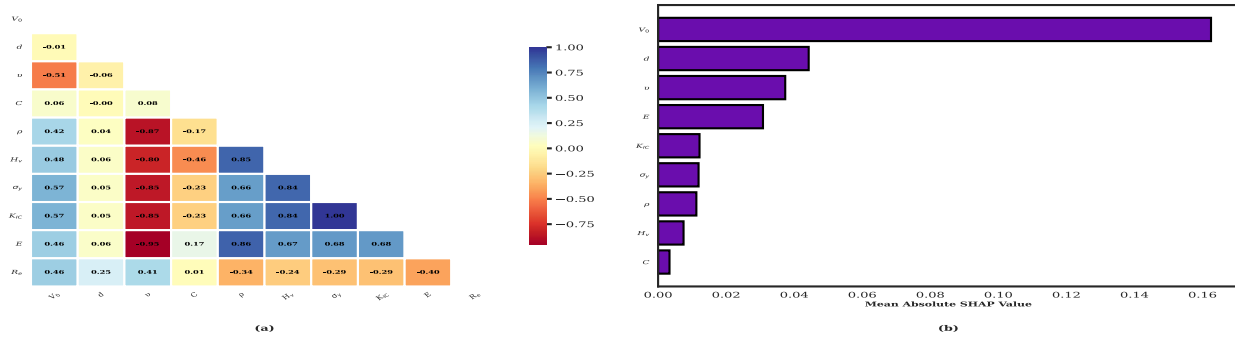


Figure 4.6. a) Correlation matrix of input features and target R_e , b) Features impact on the R_e using an absolute SHAP algorithm

The figure highlights a positive Spearman rank correlation coefficient ($r_s = 0.46$) between impact speed (V_o) and maximum erosion rate (R_e), confirming that R_e increases significantly with higher V_o . This aligns with physical expectations, as higher impact velocities subject the material to more intense stresses, accelerating erosion. In contrast, intrinsic material properties including density

(ρ), yield Strength (σ_y), hardness (H_v), fracture toughness (K_{IC}), and elastic modulus (E) exhibit moderate negative correlations ($r_s = -0.34, -0.29, -0.24, -0.29, -0.40$, respectively) with R_e . These material properties influence erosion resistance; however, their effects are secondary compared to the impact of V_o .

Since correlation analysis captures only monotonic relationships, feature importance is evaluated using mean absolute SHAP values. The feature importance provides insights into the relative importance of different input features in predicting R_e . The contribution of each feature to the target value is visually represented using the mean absolute SHAP algorithm, as shown in Figure 4.6(b). Each feature's mean absolute SHAP value quantifies its contribution to R_e , with higher values indicating a greater impact. To further assess feature importance, SHAP analysis was employed. The results reinforce the correlation findings:

- V_o dominates with the highest mean absolute SHAP value (0.16), confirming its primary influence.
- Droplet Diameter (d) and Poisson's Ratio (ν) rank next, but their contributions are significantly smaller (SHAP ≈ 0.05 and 0.04 , respectively).
- Material properties (E , K_{IC} , σ_y , H_v , ρ , and C) exhibit progressively lower SHAP values, indicating negligible direct effects on R_e .

This dual analysis—correlation and SHAP—consistently identifies V_o as the most critical factor in erosion prediction, while material properties play a less important role. It establishes a priority ranking of input parameters for ML-based erosion modeling.

4.3.2 Performance of Baseline Models with Hyperparameter Optimization

The initial ML models, with tuned hyperparameters, serve as baselines for predicting R_e in the WDE process. The overall effectiveness of these baseline models has been evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. Figure 4.7 presents the prediction results of the baseline model after hyperparameter tuning, comparing individual algorithms including AdaBoost, Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Elastic Net (ENet), Linear Regression (LR), Lasso Regression (Lasso), Ridge Regression, and Artificial Neural Networks (ANN).

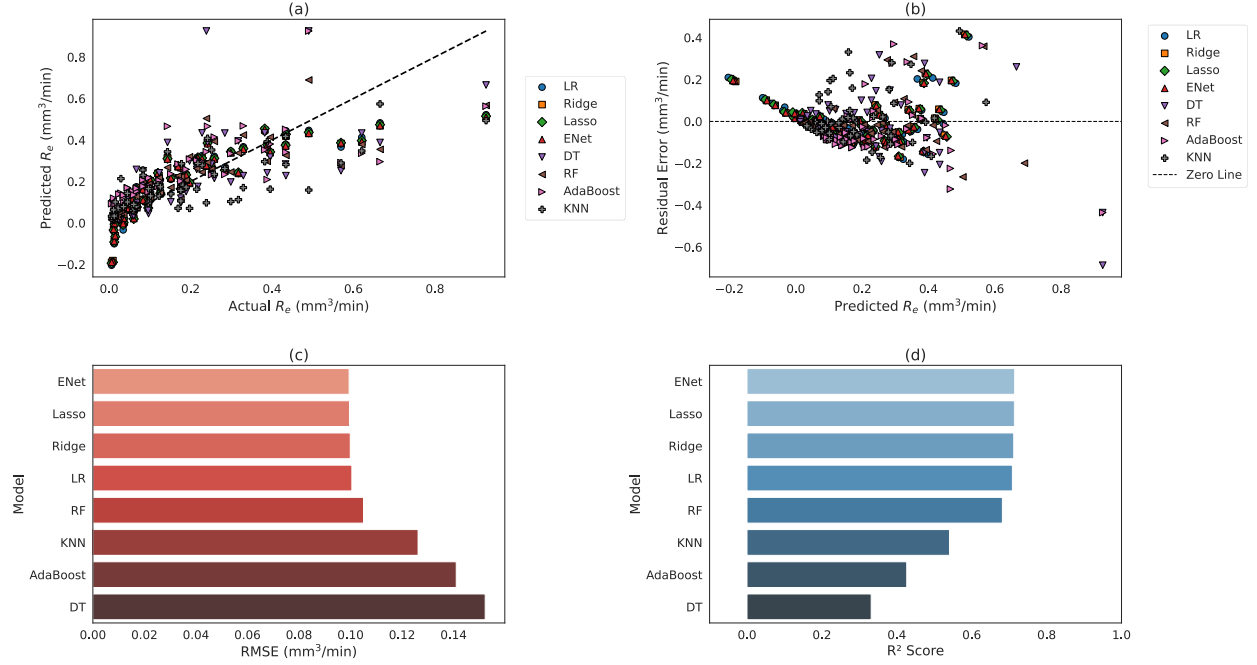


Figure 4.7. Baseline tuned model predictions, (a) Actual vs predicted R_e values, (b) Residual error, (c) RMSE, and (d) R^2 score

The results indicate that hyperparameter tuning produced only a limited gain in predictive accuracy. For instance, the KNN model achieved $R^2 = 0.57$, while the LR, Lasso, and Ridge models remained around $R^2 = 0.70$. Although minor improvements were observed in some cases, the overall impact of hyperparameter tuning was modest in predicting R_e . As a result, this study further explores data transformation techniques—shown effective in our prior work on the incubation period [128] to enhance the predictive performance of these models further.

4.3.3 ML Model Improvements

4.3.3.1 Model Evaluation Using Data Transformation

This section examines the effect of data transformation techniques on various machine learning models. We applied logarithmic, Yeo-Johnson, and Box-Cox transformations to the target variable before training ML models such as Lasso, Ridge, Decision Tree (DT), Random Forest (RF), and K-nearest neighbours (KNN), determining whether these transformations can enhance predictive accuracy, mitigate overfitting, and improve generalization.

Figure 4.8 illustrates one of the best results (Yeo-Johnson Transformation) obtained after applying different transformation methods to the target variable. These transformations did not significantly improve over the baseline models. The R^2 values remained comparable to those obtained without transformation, and RMSE values did not show a notable reduction. Table 4.2 summarizes the ML models performance after data transformations.

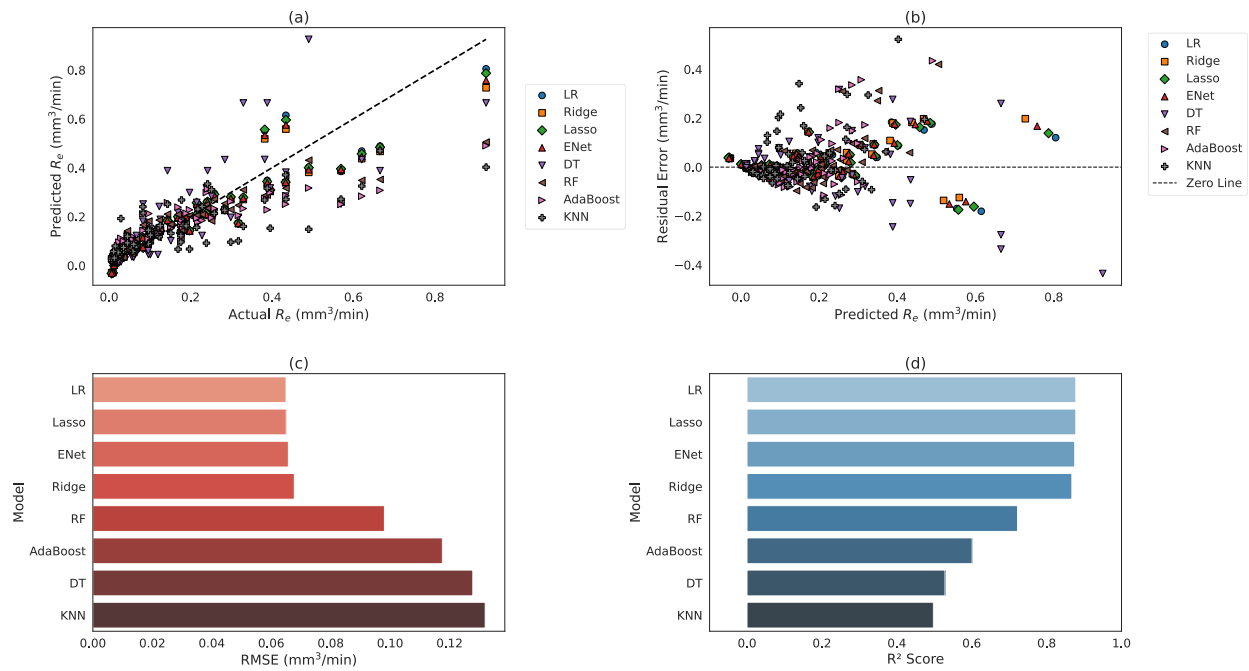


Figure 4.8. ML models with Yeo Johnson transformation: (a) Predictions actual vs predicted R_e , (b) Residual error, (c) RMSE, and (d) R^2 Score

Table 4.2. ML models performance metrics with data transformation

Transformations	Metric	Lasso	Ridge	LR	RF	DT	AdaBoost	KNN	ENet
Yeo- Johnson	R^2	0.87	0.85	0.87	0.75	0.53	0.58	0.49	0.86
	RMSE (mm ³ /min)	0.07	0.09	0.08	0.12	0.14	0.13	0.15	0.09
Box-Cox	R^2	0.86	0.86	0.87	0.71	0.49	0.56	0.48	0.84
	RMSE (mm ³ /min)	0.08	0.08	0.07	0.12	0.14	0.14	0.12	0.10
Logarithmic	R^2	0.87	0.87	0.88	0.69	0.46	0.58	0.44	0.86
	RMSE (mm ³ /min)	0.07	0.07	0.06	0.13	0.14	0.13	0.14	0.09

4.3.3.2 Data Augmentation Improvement with Linear and Non-Linear Models

Data augmentation is a key strategy to enhance ML model performance, especially with limited or imbalanced data. This subsection evaluates the impact of different augmentation techniques on linear, k-nearest neighbors, and tree-based models, divided into two parts: the first focuses on training and testing with cross-validation. In contrast, the second assesses model robustness on unseen validation data. Together, these analyses highlight how augmentation improves predictive accuracy and generalization.

4.3.3.2.1 Training and Testing

This section examines the effect of data augmentation on the performance of various machine learning models, focusing on identifying optimal configurations that enhance predictive accuracy, mitigate overfitting, and improve generalization. The augmented data points are around 1000 points generated by the different suggested methods mentioned previously. Figures 4.9 through 4.13 illustrate the most promising results regarding predictive outcomes and evaluation metrics, derived from training and K-Fold cross-validation procedures. The results demonstrate the models' ability to generalize effectively beyond the training data while minimizing overfitting.

Notably, Figure 4.9(a) illustrates a substantial improvement in predictive accuracy following applying the Multivariate Interpolation technique across different ML models. Specifically, the coefficients of determination (R^2) improve significantly for all the ML techniques.

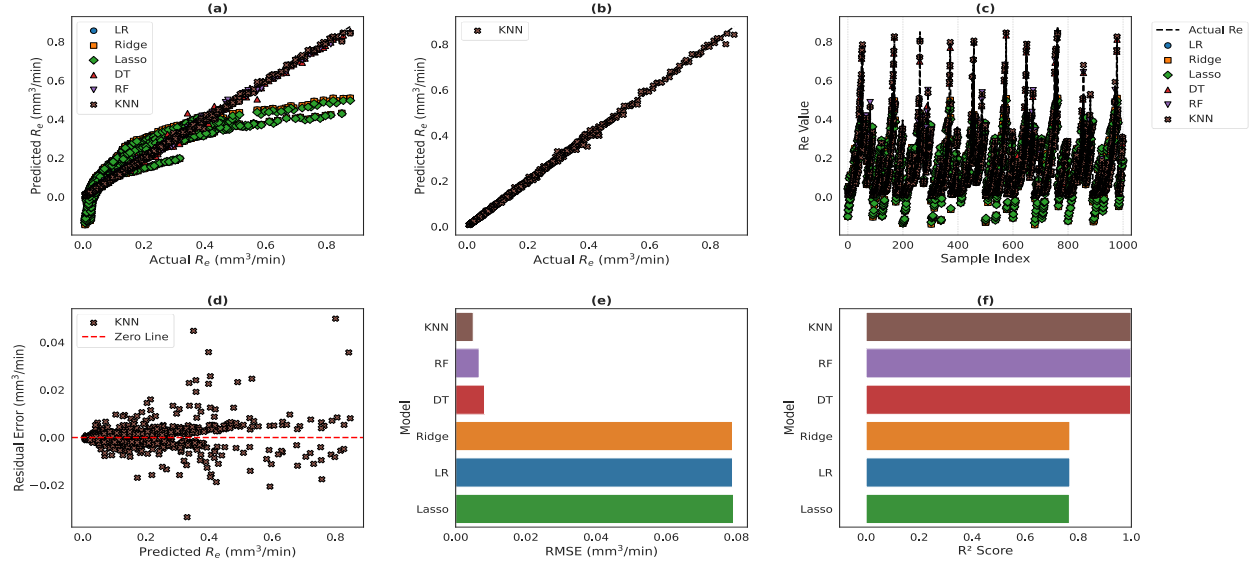


Figure 4.9. (a) Predicted vs. actual R_e values of different models (Multivariate Interpolation), (b) KNN model predictions, (c) R_e Value predictions vs. sample index, (d) Residual error, (e) RMSE, (f) R^2 score

This improvement was accompanied by reduced RMSE values, particularly when using LR, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) regressors. Table 4.3 summarizes the evaluation metrics before and after data augmentations.

Table 4.3. ML Models performance before and after applying data augmentation (Multivariate Interpolation)

Model	Before Augmentation			After Augmentation		
	R^2	RMSE (mm ³ /min)	MAE (mm ³ /min)	R^2	RMSE (mm ³ /min)	MAE (mm ³ /min)
KNN	0.541	0.126	0.089	0.99	0.005	0.003
RF	0.682	0.105	0.069	0.99	0.007	0.004
DT	0.332	0.152	0.096	0.99	0.008	0.005
Ridge	0.712	0.100	0.071	0.77	0.079	0.056
LR	0.708	0.110	0.071	0.76	0.079	0.056
Lasso	0.714	0.100	0.070	0.77	0.079	0.056
ENet	0.715	0.100	0.070	0.77	0.079	0.056
AdaBoost	0.427	0.140	0.107	0.70	0.0893	0.079

Figure 4.9(b) presents a scatter plot of KNN's predicted versus actual maximum erosion rate values, demonstrating exceptional alignment with the 1:1 reference line. This result shows that KNN achieves an R^2 exceeding 0.95 and among the lowest RMSE values across all models. Figure 4.9(d) further validates KNN's robustness by displaying residuals against predicted values, revealing a random distribution centered near zero—indicating minimal systematic error. Figure 4.9(e) and Figure 4.9(f) establish comparable top-tier performance of KNN, DT, and RF, with all three models achieving similarly high R^2 scores and low RMSE values, confirming their superior efficacy in modeling complex WDE relationships.

Figure 4.10 presents a comprehensive evaluation of machine learning models, highlighting the impact of the Mixup data augmentation technique. The results (Figure 4.10a-f) collectively illustrate model performance, error characteristics, and overall efficacy, with a focus on the KNN algorithm.

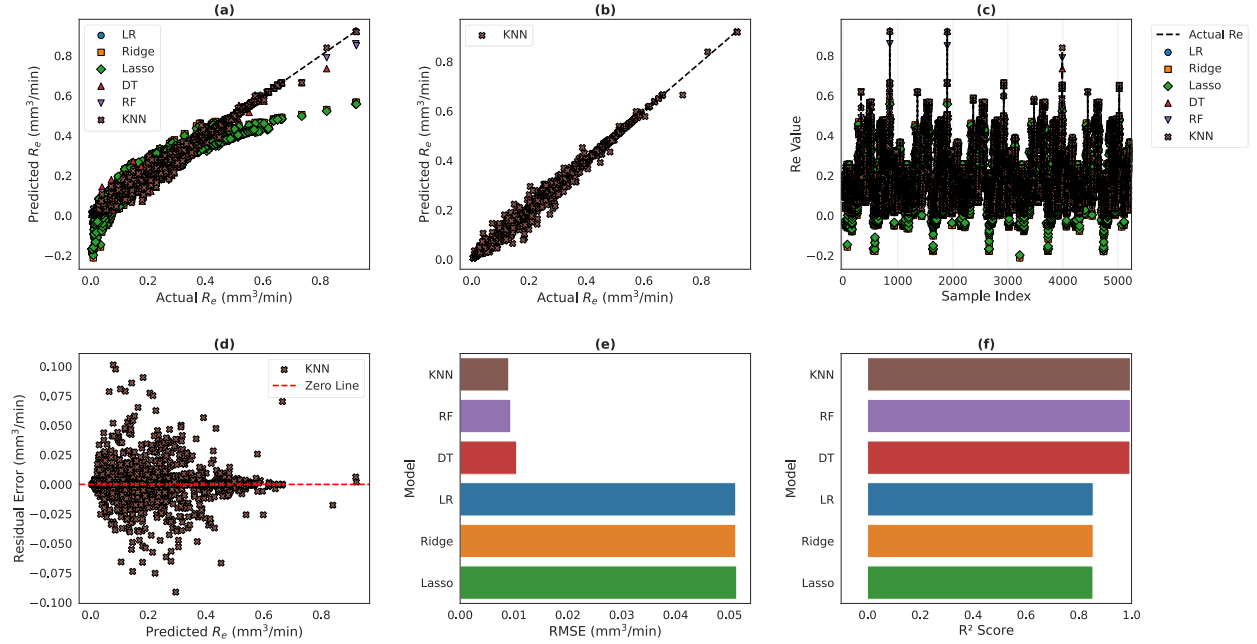


Figure 4.10. (a) Predicted vs. actual R_e values of different models (Mixup), (b) KNN model predictions (c) R_e Value predictions vs. sample index, (d) Residual error, (e) RMSE, (f) R^2

Figure 4.10 (a) provides a comparative analysis of predicted versus actual values across ML models, where proximity to the ideal diagonal indicates predictive accuracy; the KNN model demonstrates superior alignment, reflecting its robust performance. Figure 4.10 (b) shows that most KNN predictions cluster closely along the ideal diagonal, indicating generally high predictive fidelity, though the lowest R_e values display a slightly wider spread. Figure 4.10 (c) depicts predicted values as a function of sample index, illustrating KNN's consistency within R_e ranges in WDE. Figure 4.10 (d) shows residual error, providing insights into error distribution; the absence of discernible patterns indicates unbiased predictions and a well-fitted model. Quantitative metrics in Figures 4.10 (e) and (f) report the Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2), respectively, with KNN achieving the lowest RMSE and highest R^2 , corroborating its superior performance. These results underscore Mixup's efficacy in enhancing predictive accuracy, robustness, and generalization for Decision Tree, Random Forest, and KNN models, achieving R^2 values exceeding 0.95 and substantially reduced RMSE, while effectively mitigating overfitting.

Figure 4.11 evaluates the performance of various machine learning models in predicting the R_e after applying SMOTER (Synthetic Minority Over-sampling Technique for Regression). This highlights how SMOTER improves predictive accuracy, particularly for imbalanced datasets.

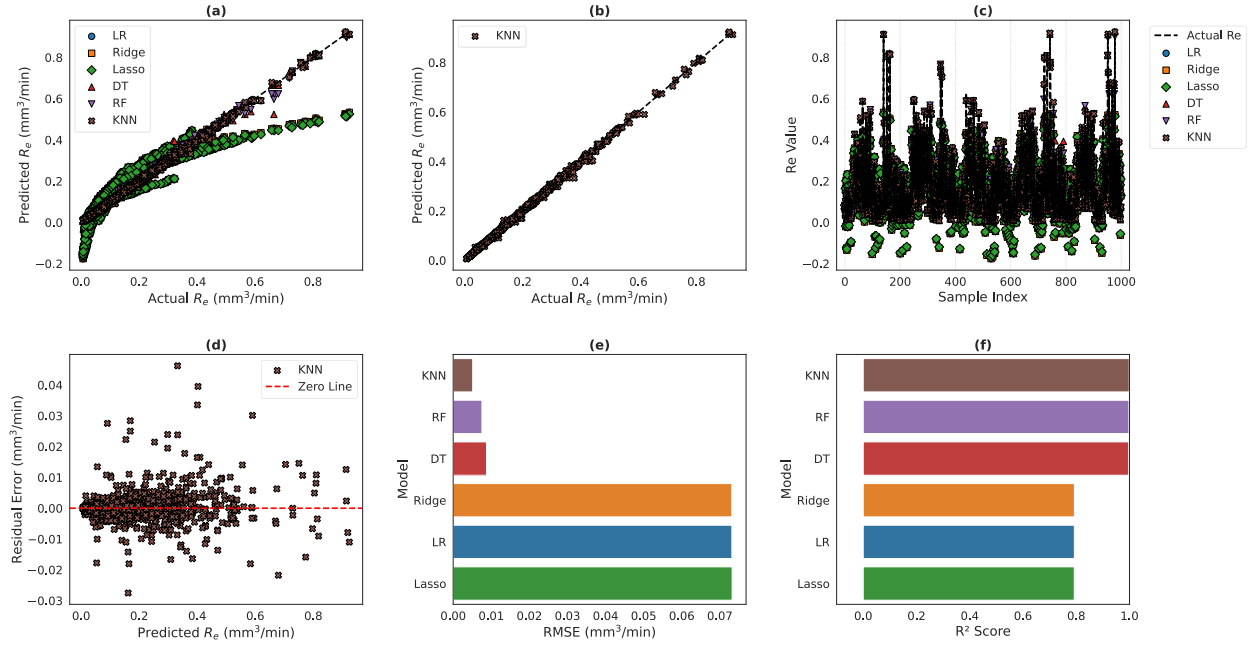


Figure 4.11. (a) Predicted vs. actual R_e values of different models (SMOTER), (b) KNN model predictions, (c) R_e Value predictions vs. sample index, (d) Residual error, (e) RMSE, and (f) R^2

Figure 4.11(b, d) focuses on K-Nearest Neighbors (KNN), the top-performing model. The scatter plot of predicted vs. actual R_e values show a tight clustering along the diagonal. Figure 4.11(c) presents the predicted values versus sample index, demonstrating the KNN model's stable performance within the various R_e ranges encountered in WDE. In addition, Figure 4.11(e) shows KNN achieves a low RMSE of ~ 0.004 mm³/min, demonstrating accurate predictions. While Figure 4.11(f) shows high R^2 values for KNN, DT, and RF, outperforming linear models such as Ridge, LR, and Lasso, highlighting the benefit of nonlinear models and SMOTER augmentation.

SMOTER's effectiveness comes from generating synthetic samples in sparse regions, improving representation of underrepresented data, and enhancing model generalization. This expansion of the feature space increases robustness and predictive performance, as reflected in the superior metrics.

Figure 4.12 shows the effect of Noise Injection as a data augmentation method on Decision Tree, Random Forest, and K-nearest neighbors models for predicting R_e . While R^2 values exceed 0.95, indicating a strong correlation between predicted and actual values, the error rates are higher compared to other augmentation techniques.

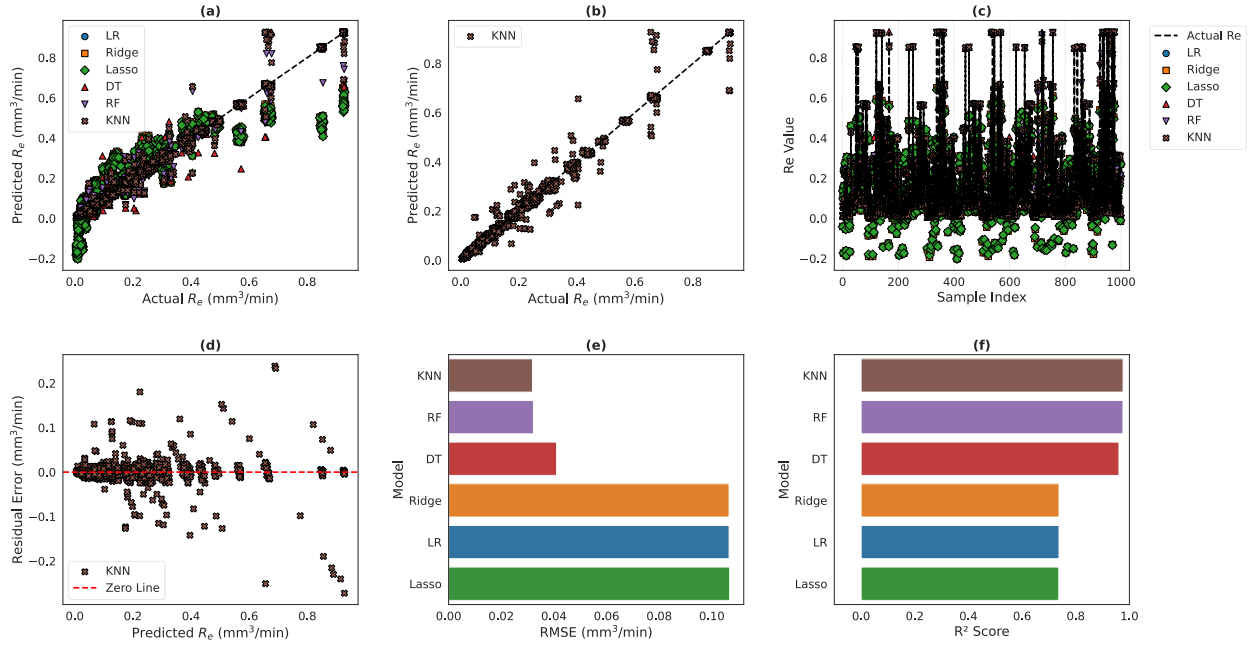


Figure 4.12. (a) Predicted vs. actual R_e values of different models (Noise Injection), (b) KNN model predictions, (c) R_e Value predictions vs. sample index, (d) Residual error, (e) RMSE, and (f) R^2 .

Figure 4.12(b) highlights KNN performance, where the scatter plot shows reasonable clustering along the regression line, though with wider dispersion than seen with Mixup and SMOTER. This is reflected in the RMSE of ~ 0.016 mm³/min, higher than Mixup (0.010 mm³/min) and SMOTER (0.004 mm³/min), indicating lower predictive accuracy. Figure 4.12(c) shows that Noise Injection preserves overall trends but increases deviation, reducing fine-grained precision. Residual analysis (Figure 4.12(d)) reveals a broader error spread, suggesting less consistent performance compared to Mixup and SMOTER. Quantitative metrics (Figures 4.12(e) and 4.12 (f)) confirm that KNN, though dominant among Noise Injection models, underperforms in RMSE and precision.

Figure 4.13 evaluates Conditional Tabular Generative Adversarial Networks (CTGAN) applied to various regression models, assessing its impact on predictive accuracy for potentially imbalanced datasets. Figure 4.13 (a) compares predicted versus actual values across all CTGAN-augmented

models, showing strong alignment with the diagonal for KNN, Random Forest (RF), and Decision Tree (DT), with RF emerging as the top performer.

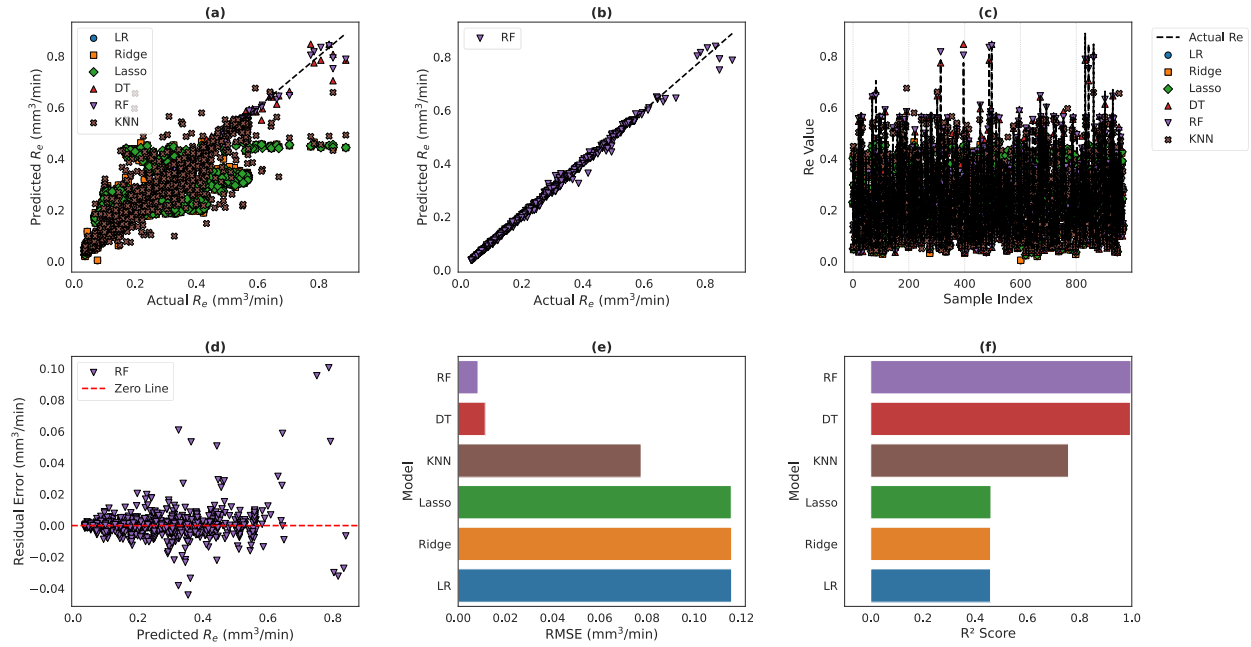


Figure 4.13. (a) Predicted vs. actual R_e values of different models (CTGAN), (b) KNN model predictions, (c) R_e Value predictions vs. sample index, (d) Residual error, (e) RMSE, and (f) R^2

Figure 4.13(b) focuses on RF, where tight clustering along the diagonal demonstrates exceptional predictive power, with $R^2 > 0.95$, accounting for over 95 % of variance in R_e values. Residual analysis (Figure 4.13(d)) shows RF errors randomly distributed around zero with minimal dispersion, indicating unbiased, consistent predictions comparable to SMOTER.

Quantitative metrics (Figures 4.13(e) and 4.13(f)) confirm that RF achieves the lowest RMSE (~ 0.005 mm³/min), outperforming Noise Injection (0.016 mm³/min). Along with KNN and DT, RF attains the highest R^2 , demonstrating the advantages of CTGAN and non-linear models over linear approaches like Ridge, LR, and Lasso.

Table 4.4 compares model performance across six augmentation techniques, revealing significant variations in predictive accuracy and error metrics. The results demonstrate that four augmentation strategies—Mixup, SMOTER, Multivariate Interpolation, and CTGAN—achieve exceptional R^2 values of 0.99, indicating near-perfect model fit. These top-performing methods also exhibit minimal prediction errors, with RMSE values ranging from 0.003 to 0.010 mm³/min.

Table 4.4. ML models' evaluation metrics from different data augmentations

Augmentation	Algorithm	R^2	RMSE (mm ³ /min)
Mixup	KNN	0.99	0.010
SMOTER	KNN	0.99	0.004
Multivariate Interpolation	KNN	0.99	0.003
Noise Injection	DT	0.96	0.016
CTGAN	RF	0.99	0.005
Gaussian Sampling Processes (GP)	RF	0.65	0.087

In contrast, Noise Injection shows diminished performance ($R^2=0.96$, RMSE=0.016 mm³/min), while GP yields substantially inferior results ($R^2=0.65$, RMSE=0.087 mm³/min). This quantitative analysis distinguishes between highly effective and suboptimal augmentation approaches, providing critical guidance for selecting techniques in predictive modeling tasks. The subsequent section builds upon these findings by evaluating the robustness of the top-performing models (Mixup, SMOTER, Multivariate Interpolation, and CTGAN) on unseen validation data to further assess their generalization capabilities.

4.3.3.2.2 Holdout dataset Validation

This section evaluates the robustness of the ML models using an independent validation dataset consisting of 8 data points from the original dataset (around 12 %), data not previously utilized during training or testing. Figure 4.14 compares the unseen maximum erosion rates and the predicted values generated by models trained on augmented datasets employing Mixup, SMOTER, CTGAN, and Noise Injection techniques. This analysis provides critical insight into the models' ability to generalize and maintain predictive accuracy on unseen data.

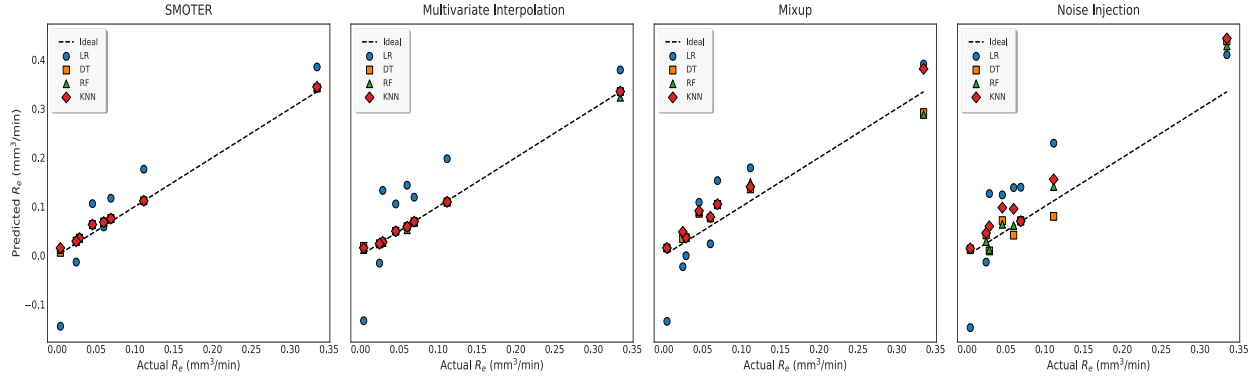


Figure 4.14. Predicted unseen vs. actual values of R_e (SMOTER, Multivariate Interpolation, Mixup, and Noise Injection)

Building upon the validation results presented earlier, the following analysis further explores the comparative performance of these models and the underlying reasons for their varying predictive capabilities. These models were developed using various algorithms, with the best-performing configurations selected based on comprehensive metric evaluations. A higher R^2 value—approaching unity—combined with lower error metrics indicates superior model fit and predictive accuracy on unseen data, thereby reflecting enhanced model robustness. Notably, the Multivariate Interpolation data augmentation method demonstrates a performance trend comparable to SMOTER and CTGAN, as evidenced in Figures 4.9, 4.11, and 4.13. This consistency further underscores the substantial improvements in predictive accuracy achieved through data augmentation across multiple machine learning algorithms for modeling R_e . Table 4.5 shows evaluation metrics (unseen dataset) from different data augmentations.

Table 4.5. ML models' evaluation metrics based on an unseen dataset from different data augmentation techniques

Augmentation	Algorithm	R^2	RMSE (mm ³ /min)
SMOTER	LR	0.54	0.067
	DT	0.99	0.008
	RF	0.99	0.009
	KNN	0.99	0.009
Multivariate Interpolation	LR	0.32	0.082
	DT	0.99	0.005
	RF	0.99	0.006
	KNN	0.99	0.005
Mixup	LR	0.46	0.073
	DT	0.92	0.029
	RF	0.91	0.030
	KNN	0.89	0.033
Noise Injection	LR	0.10	0.094
	DT	0.83	0.041
	RF	0.85	0.038
	KNN	0.76	0.049

The superior performance ($R^2 > 0.90$) was observed in k-nearest neighbors and tree-based models, specifically, the Decision Tree, Random Forest.

Data augmentation has significantly enhanced model performance in Water Droplet Erosion (WDE) prediction tasks. The augmented models consistently yield highly accurate predictions that closely align with observed values across most cases. In contrast, the Linear Regression (LR) model shows comparatively lower accuracy ($R^2 < 0.60$), highlighting the limitations of linear approaches and the necessity for more sophisticated, non-linear methods in effectively modeling WDE phenomena. Consequently, the following section focuses on Artificial Neural Networks (ANN) to apply similar data augmentation strategies and further explore improvements in predictive performance.

4.3.3.3 Model Improvement through Data Augmentation for Artificial Neural Network

This section will examine how data augmentation and hyperparameter tuning improve the performance of ANN models. Data augmentation combined with hyperparameter tuning has been employed to enhance the prediction accuracy of artificial neural networks (ANNs). ANN architecture comprises four hidden layers (256,128,64,32). Hyperparameter optimization was performed using a grid search approach, fine-tuning the network with various optimization techniques. The Adam optimizer and the ReLU (Rectified Linear Unit) activation function were used in the ANN models training on augmented data generated through multivariate interpolation, Mixup, SMOTER, and noise injection techniques. A learning rate of 0.001 was obtained for optimization. Training and testing model performance were evaluated using RMSE, MAE, and R^2 metrics to ensure accurate and reliable predictions. The data was trained and tested using K-Fold CV to ensure a robust model. Figures 4.15 (a-d) illustrate the improvements in ANN model performance following the implementation of Multivariate Interpolation, Mixup, CTGAN, and SMOTE data augmentation techniques.

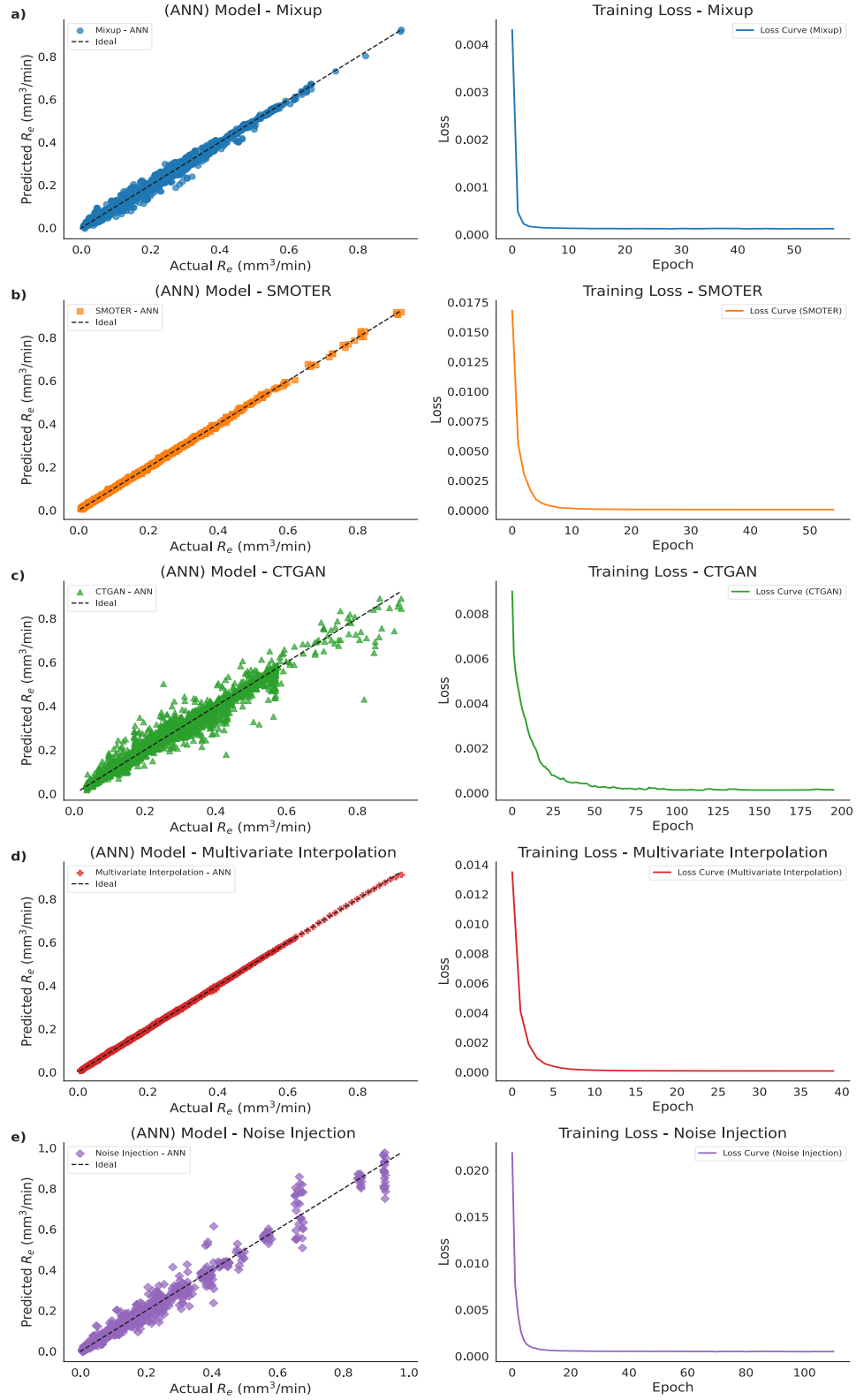


Figure 4.15. ANN model predictions vs. actual R_e values and loss function with data augmentation, a) Mixup, b) SMOTER, c) CTGAN, d) Multivariate Interpolation, and e) Noise Injection

Figure 4.15(a) illustrates the relationship between predicted and actual R_e values following the application of the Mixup method. The scatter plot in blue demonstrates a good correlation, with most points clustering along the diagonal line. Some spread is visible, but predictive accuracy is generally high. The accompanying loss curve, plotted over approximately fifty epochs, exhibits a sharp decline during the initial ten epochs and subsequently stabilizes by the final epoch, demonstrating effective convergence.

Figure 4.15(b) presents the predicted and actual R_e values using the SMOTER method. The orange scatter points align almost perfectly with the diagonal, showing extremely tight clustering and high predictive accuracy across the entire R_e range. The loss curve, plotted over approximately fifty epochs, exhibits a steep decline during the initial ten epochs and subsequently stabilizes.

Figure 4.15(c) depicts the results for the CTGAN method. The green scatter points follow the diagonal with some deviations, particularly at mid-to-higher R_e values, where variability increases compared to Mixup and SMOTER. This indicates reduced predictive consistency. The loss curve, plotted over approximately two hundred iterations, shows a steady decrease during the initial fifty epochs and gradually stabilizes.

Figure 4.15(d) shows the outcomes of the Multivariate Interpolation method. The red scatter points closely follow the diagonal, indicating predictions closely aligned with observed values and demonstrating consistently high accuracy across the entire R_e range. The loss curve, plotted over approximately forty epochs, exhibits a sharp decline during the initial 10 epochs and subsequently flattens, indicating convergence.

Figure 4.15(e) illustrates results for the Noise Injection method. The purple scatter points generally align with the diagonal but with noticeable spread across the range of R_e , indicating moderate predictive accuracy with larger errors than Multivariate Interpolation, Mixup, and SMOTER. The loss curve, plotted over approximately one hundred epochs, exhibits a sharp decrease during the initial twenty epochs and stabilizes thereafter. Although convergence is achieved, the predictive scatter indicates relatively noisier performance.

Table 4.6 summarizes the performance of the R_e predictors in the neural network model for the proposed augmentation methods. The model performance and improvements are evaluated using

RMSE and R^2 metrics. Multivariate Interpolation, SMOTER, and Mixup among the augmentation methods exhibit the best results, followed by the Noise Injection and CTGAN processes. In contrast, Gaussian Process Sampling demonstrates poor predictive capabilities, making it unsuitable for this dataset when predicting the maximum erosion rate in the WDE process.

Table 4.6. The ANN-based ML models' evaluation metrics from different data augmentations

Augmentation	R^2	RMSE (mm ³ /min)
Mixup	0.99	0.011
SMOTER	0.99	0.005
Multivariate Interpolation	0.99	0.003
Noise Injection	0.97	0.035
CTGAN	0.93	0.040
Gaussian Process Sampling	0.61	0.091

The ANN architecture, equipped with tunable hyperparameters, achieved remarkable predictive accuracy when trained on data augmented using the SMOTER and Multivariate Interpolation techniques, reaching RMSE values of 0.005 and 0.003 mm³/min and an R^2 of 0.99. Conversely, the ANN model trained with GP augmentation demonstrated considerably lower performance, with an R^2 value of 0.61 and an RMSE of 0.091 mm³/min.

In summary, the data augmentation methods implemented in this study have substantially enhanced the performance of machine learning models. Among the techniques assessed, SMOTER, Mixup, Multivariate Interpolation, and CTGAN consistently delivered superior results, effectively mitigating data scarcity issues, promoting model stability and convergence, and significantly improving predictive accuracy. Furthermore, the robustness of these models was evaluated on an independent validation dataset comprising previously unseen data. The models trained and tested using Mixup and SMOTER augmentations exhibited superior generalization capabilities, as evidenced by high R^2 values exceeding 0.90 and minimal prediction errors.

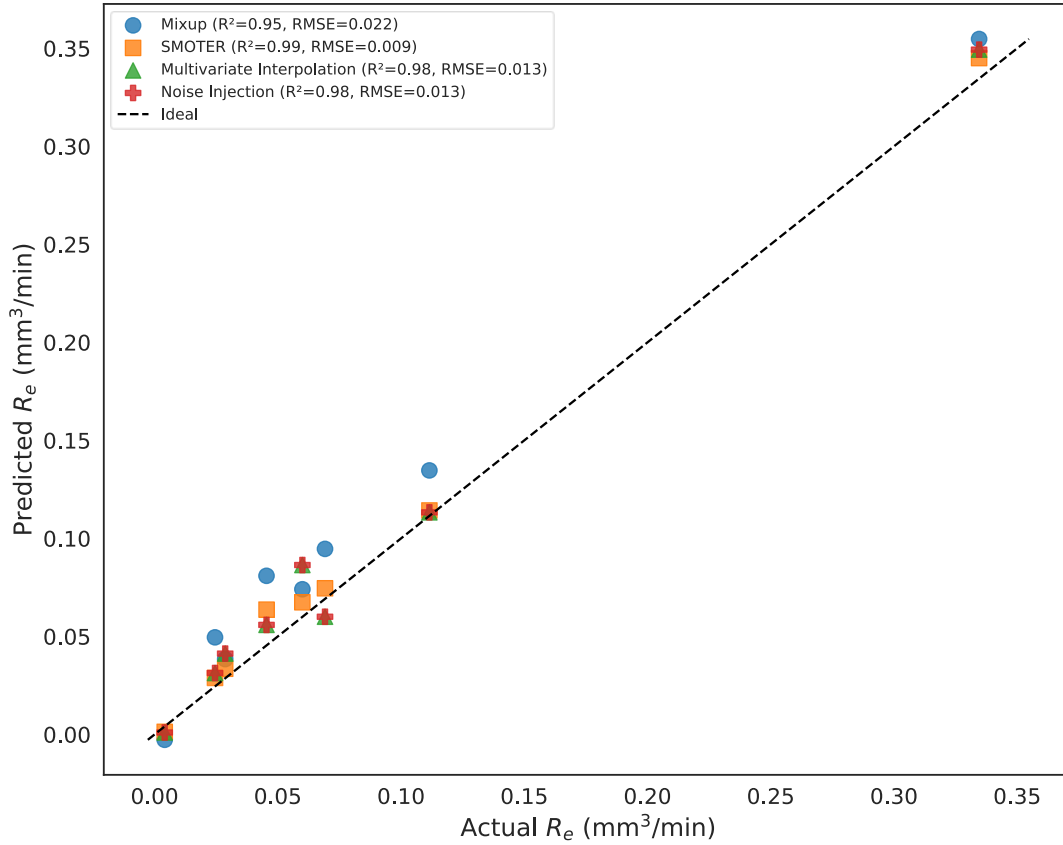


Figure 4.16. ANN Model predicted unseen vs. actual values of maximum erosion rate (Mixup, SMOTER, Multivariate Interpolation, and Noise Injection)

Figure 4.16 illustrates the strong agreement between predicted and actual R_e values for the external dataset, demonstrating the effectiveness of augmentation-enhanced models. All approaches achieved high predictive performance with $R^2 > 0.90$ and low RMSE values. The best results were obtained with the SMOTER method, which achieved the lowest RMSE of $0.009 \text{ mm}^3/\text{min}$, followed closely by Multivariate Interpolation and Noise Injection, while Mixup exhibited slightly higher errors. These findings confirm that data augmentation substantially enhances model generalization and reliability on unseen data.

4.4 Conclusion

This study introduces a data-driven methodology for predicting maximum erosion rate in water droplet erosion (WDE) using machine learning (ML) techniques. Correlation matrices and SHAP analysis were used to identify the importance of features and guide the development of predictive

models. The analysis identified impact velocity (V_o) as the most influential factor affecting R_e , followed by droplet diameter (d) and Poisson's ratio (ν). At the same time, material properties such as yield strength (σ_y), hardness (H_v), and fracture toughness (K_{IC}) showed a moderate impact. In contrast, density (ρ) and modulus of elasticity (E) had minimal effects. Although hyperparameter tuning and data transformation techniques (e.g., Yeo-Johnson, Box-Cox, and logarithmic transformations) were applied, they did not significantly improve model performance, suggesting that transformation-based feature scaling is insufficient for capturing the underlying non-linear relationships that govern the R_e .

In contrast, data augmentation strategies proved highly effective techniques such as CTGAN, SMOTER, and Multivariate Interpolation yielded substantial improvements in predictive accuracy for various models. K-fold cross-validation results showed exceptional R^2 scores and minimal RMSE for KNN, DT, and Random Forest models, while GP consistently underperformed, especially for linear models ($R^2 = 0.61$, RMSE = 0.091). Artificial Neural Networks (ANNs) also benefited from data augmentation, with SMOTER and Mixup producing R^2 values exceeding 0.95 in cross-validation. Validation on unseen data confirmed these trends, with models trained using SMOTER, Mixup, and Noise Injection achieving excellent generalization performance ($R^2 > 0.90$), highlighting their robustness and reliability.

Moreover, the superior performance of tree-based models (Decision Tree and Random Forest) indicates that features relevant to WDE may possess threshold-dependent properties. KNN's strong results imply a smooth and structured feature space where similar instances produce similar erosion responses. On the other hand, linear regression models consistently underperformed, both in cross-validation and on unseen data, reinforcing the necessity for non-linear, data-adaptive models in WDE prediction. Overall, this research demonstrates that integrating advanced machine learning algorithms with strategic data augmentation significantly enhances the predictive modeling of WDE, reducing the reliance on costly and time-consuming experiments and paving the way for more effective material design and erosion mitigation strategies.

Acknowledgments: Hashemite University is acknowledged for sponsoring the first author's PhD studies. MT and MM gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC). Finally, the paper benefited from the use of large language models (LLMs) for assistance with language editing and polishing. All scientific content, data collection, analysis, and conclusions are the sole work of the authors.

Author contributions: All authors contributed to the study conception and design. **Khaled AlHammad:** Data curation, Formal analysis, Investigation, Methodology, Writing –original draft. **Mamoun Medraj** and **Moussa Tembely:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. All authors have read and approved the final manuscript.

Funding: Not applicable.

Data Availability: The datasets analyzed during the current study are publicly available in the GitHub repository at the following link: <https://github.com/khaledalhammad/Maximum-Erosion-Rate>

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests

Following the development of a data augmentation strategy for predicting the maximum erosion rate in Chapter 4, the next step is to explore how different machine learning models perform in terms of accuracy, generalization, and interpretability. While Chapter 4 primarily focused on establishing a robust prediction framework through data augmentation and initial model development under data scarcity, Chapter 5 significantly expands upon this foundation. It shifts attention to a comprehensive evaluation of a diverse range of traditional and deep learning models, explicitly building on the predictive capabilities developed previously. The core distinction lies in Chapter 5's emphasis on rigorously comparing these models not only for their predictive accuracy but also for their ability to generalize to unseen materials and provide physically meaningful insights, which was not the primary focus of the earlier chapters. This includes both state-of-the-art architectures—such as FT-Transformer, SAINT, TabNet, and TabPFN—and more interpretable approaches grounded in domain knowledge, such as dimensionless analysis and principal component-based regression. By systematically comparing these models, Chapter 5 aims to identify which methods best balance predictive power, generalizability, and interpretability for practical applications in WDE prediction, thereby extending the foundational work of Chapters 3 and 4 into a broader comparative analysis.

Chapter 5: Toward Interpretable Data-Driven Models for Predicting Water Droplet Erosion Incubation Period

Khaled AlHammad, Mamoun Medraj, Moussa Tembely, Julio. C Franco-Correa

Note. This chapter is based on a manuscript authored by K. AlHammad, M. Medraj, and M. Tembely, J.C. Franco-Correa, currently submitted for publication to the Computational Materials Science (Journal)

Abstract

Water droplet erosion (WDE) causes progressive material degradation in engineering applications such as steam and gas turbines and aircraft components. Accurate prediction of the incubation period (N_0)—the initial phase before measurable material loss—is crucial for maintenance planning and material selection. Following a dimensionless analysis of WDE, which identified physically relevant variables and yielded an approximate physics-based correlation for the incubation period, machine learning (ML) methods are explored for improved prediction. In this paper, we evaluate multiple ML approaches using a dataset of five alloys, including traditional methods (Multiple Linear Regression, MLR) based on dimensionless analysis, PCA-neural networks (PCA-NN), and advanced tabular architectures (FT-Transformer, TabNet, SAINT, TabPFN). Data preprocessing is tailored to each approach, employing physics-informed dimensionless groups for MLR and standardized scaling for PCA-NN and deep learning models.

On the primary test dataset, TabNet, FT-Transformer, and SAINT achieved the highest accuracy ($R^2 = 0.99$), followed by TabPFN ($R^2 = 0.96$). To validate on previously unseen material, TabPFN and SAINT, leveraging meta-learning and transformer architectures, demonstrated superior generalization, whereas simpler models such as PCA-NN and MLR were less effective when extrapolating to new materials.

This study establishes a robust framework for WDE prediction, balancing predictive accuracy, interpretability, and generalization. The results provide a principled basis for model selection in materials science, particularly for applications requiring reliable predictions across diverse material types.

Keywords: Water Droplet Erosion, Incubation Period, Dimensionless Analysis, Multiple Linear Regression, Principal Component Analysis, Tabular Foundation Model

5.1 Introduction

High-velocity liquid droplets repeatedly striking solid surfaces cause water droplet erosion (WDE), a progressive material degradation process. This phenomenon is particularly relevant in various engineering applications, including steam and gas turbine blades, aircraft components exposed to rain, and offshore structures subjected to sea spray. The erosion process typically progresses through several stages, with the incubation period (N_0) being the initial phase during which no measurable material loss occurs, despite microstructural changes occurring beneath the surface [2, 145].

Accurate prediction of the incubation period is crucial for three key benefits. First, it enables proactive maintenance scheduling before significant material loss occurs. Second, it facilitates material selection during the design phase, helping engineers choose materials with optimal erosion resistance for specific operating conditions. Third, it contributes to a deeper understanding of the physical mechanisms governing the early stages of erosion damage [1, 2, 13].

Predicting N_0 is challenging due to the complex interactions among multiple factors, including impact conditions (velocity, droplet size) and material properties (hardness, elasticity, yield strength). Traditional empirical correlations and simplified analytical models often fail to capture the full complexity of the erosion processes [13, 33, 45]. Despite recent advances in computational and analytical techniques, a systematic comparative analysis of predictive methodologies specifically targeting the WDE incubation period remains notably limited in existing literature, underscoring the importance of addressing this research gap. This limitation is particularly critical, given the incubation period's pivotal role in determining component service life and maintenance intervals [3].

The present study addresses this gap by performing a comprehensive dimensional analysis and by evaluating and comparing multiple machine learning algorithms for predicting the WDE incubation period, with particular emphasis on techniques suitable for small tabular datasets. By identifying the most effective predictive approaches and establishing connections between model performance and the underlying physical mechanisms, this work aims to contribute both theoretically and practically to advancing WDE prediction methodologies.

5.2 Literature Review

WDE research has advanced over the decades, but persistent challenges remain in fully modeling its mechanisms and accurately predicting key parameters such as erosion rate and incubation period. Springer [13] made the earliest significant contribution by establishing a correlation between erosion rate and impact velocity. Despite its widespread citation, this model [13] inadequately addresses the incubation period (N_0). The model cannot capture the complex subsurface damage mechanisms that precede visible erosion. To address these limitations, Heymann [14] developed a more comprehensive model incorporating material properties and impact conditions. His research revealed that the incubation period follows an inverse power-law relationship with impact velocity, with the exponent varying between 5 and 9 depending on the material. While this model [14] provided better physical understanding, it relied heavily on material-specific constants, limiting its broader applicability [2]. This limitation highlights a key challenge: developing universal predictive frameworks for diverse materials and WDE conditions.

Advancements in fracture mechanics experimental techniques furthered the understanding of WDE. Field [145] introduced the Multiple Impact Jet Apparatus (MIJA), enabling controlled repetition of droplet impact events and identifying threshold velocities for damage initiation. His research established correlations between fracture toughness and erosion resistance, particularly during the early stages of WDE. Although Field's work did not yield a closed-form predictive equation, it provided crucial insights into the material parameters governing the incubation period, establishing a foundation for physics-informed WDE modeling [145]. The fracture mechanics perspective has subsequently become instrumental in elucidating the mechanisms of damage initiation and progression.

Dimensionless analysis is a well-known approach in physical problems, particularly in fluid and solid mechanics, serving as a valuable modeling technique by reducing the number of independent variables and revealing scaling relationships across systems. The Buckingham Pi theorem, a cornerstone of this methodology, establishes that any physically meaningful equation can be reformulated using a reduced number of dimensionless parameters, equal to the original number of variables minus the number of fundamental dimensions [146, 147]. Applying dimensionless analysis to parameters such as impact velocity, droplet size, hardness, and fracture toughness has

led to identifying critical dimensionless groups. Notably, dimensionless numbers such as impact pressure to hardness and elastic modulus to yield strength are particularly relevant during the incubation phase, where subsurface damage accumulates without visible loss [2]. The present study will perform a comprehensive analysis by identifying all relevant physical properties. The incorporation of such dimensionless groups could enable the development of predictive models that combine physical interpretability with transferability across different materials and operating conditions.

Recent developments in computational methods and data-driven techniques have expanded machine learning (ML) applications to materials science research. ML techniques uncover hidden patterns in complex data, enabling superior predictive performance compared to traditional empirical approaches [47]. These methods have been successfully applied to predict various material properties, including mechanical strength [48], corrosion resistance [49], and fatigue life [50], demonstrating their versatility in addressing complex materials-related challenges.

Recent studies have demonstrated the potential of ML-based frameworks to outperform traditional models in erosion modeling. For instance, Zhang et al. [148] employed machine learning combined with image recognition techniques to predict the erosion life of steam turbine blade materials, achieving results that closely matched experimental data. Similarly, Zhang and Xu [149] applied Least Squares Boosting (LSBoost), a gradient boosting method, to predict solid particle erosion rates with higher accuracy than conventional empirical models.

Recent research has also focused on erosion modeling in metals, where machine learning (ML) approaches have yielded promising results. Gao et al. [21] utilized an artificial neural network (ANN) to predict cavitation damage in stainless steels, demonstrating that integrating machine learning (ML) with experimental observations can produce reliable erosion assessments. Their results showed strong agreement with empirical data, highlighting the capability of ANNs to capture nonlinear material behavior under erosive conditions.

Principal Components Analysis (PCA) represents another valuable approach for WDE modeling. As a dimensionality reduction technique, PCA transforms correlated input variables into a reduced set of orthogonal components that capture the majority of the variance within the dataset [150,

151]. These components effectively filter out noise and redundancy, allowing regression models to operate on the most informative directions of the data space. The process begins with standardizing input features [152], computing a covariance matrix, deriving eigenvectors and eigenvalues that represent directions of maximum variance, and quantifying the variance explained by each principal component. When PCA is combined with neural networks, particularly those benefiting from the reduced dimensionality and more stable learning dynamics that PCA provides [153, 154], this approach can effectively model the incubation period in small datasets with potential nonlinearities. The application of L2 regularization (weight decay) during training further mitigates overfitting by penalizing large weights and encouraging smoother, more generalizable functions [155].

Recent advancements in deep learning have introduced specialized architectures tailored for tabular data, which further enhance predictive performance in complex, small-data environments. Several state-of-the-art models have been developed specifically for tabular datasets [70–72].

The Feature Tokenizer Transformer (FT-Transformer) [70] adapts the transformer architecture for tabular data by embedding each feature as a token, enabling the model to capture intricate feature interactions via self-attention mechanisms. This capability is particularly relevant in predicting the incubation period, where complex relationships between material properties, impact conditions, and WDE behavior must be learned from limited data.

TabNet [72] represents a significant advancement in modeling tabular data for materials property prediction, particularly for WDE. Unlike traditional neural networks, TabNet employs a sequential attention mechanism that dynamically selects the most relevant features at each decision step, guided by entropy-based regularization. This process generates attention weights that indicate the relative importance of input variables, such as material hardness or impact velocity, in predicting outcomes such as the incubation period. In materials science, where understanding the physical significance of predictive features is as critical as accuracy, TabNet’s interpretability offers a breakthrough. By providing insights into the material properties that drive erosion resistance, TabNet bridges the gap between data-driven predictions and physics-based understanding, surpassing the limitations of black-box models, such as artificial neural networks (ANNs) [21] or empirical correlations [12, 14]. This capability is particularly valuable for applications like WDE,

where complex interactions between material properties and environmental conditions are currently missing and must be elucidated to inform engineering decisions.

The Self-Attention and Inter-sample Attention Transformer (SAINT) [71] represents another advancement through its dual-attention mechanism, which models both intra-sample feature dependencies and inter-sample relationships, enabling the model to learn from patterns across the entire dataset.

Tabular Prior-data Fitted Networks (TabPFN) [156] introduce a novel paradigm by leveraging meta-learning through pretraining on a diverse array of synthetic datasets, enabling effective generalization even with minimal target-specific training data.

Collectively, these investigations highlight the capacity of ML frameworks to reveal complex interactions among material properties, operational conditions, and erosion outcomes. This ability makes them promising tools to complement and enhance physically based erosion models, providing more robust and generalizable predictive performance. The success of these approaches marks a significant advancement in erosion modeling, particularly for metallic alloys, where a combination of microstructural factors and service environment conditions influences damage mechanisms.

Despite growing interest in data-driven approaches, relatively few studies have focused specifically on predicting the incubation period of WDE using machine learning techniques. More critically, a systematic comparative analysis of different ML models tailored to this specific problem remains notably absent from the existing literature. This gap is particularly significant given the importance of the incubation period in determining component service life and maintenance intervals [3].

In the following sections, a detailed explanation of the methodology used to develop and evaluate the machine learning models is provided. This is followed by a presentation of simulation-based results and a discussion of the findings. The paper concludes with implications for WDE prediction and outlines promising directions for future research.

5.3 Methodology

The investigation of water droplet erosion (WDE) incubation period prediction necessitates a systematic approach that integrates physical principles with advanced data analysis techniques. Given the inherent challenges posed by limited experimental data in this domain, our methodology was carefully designed to address these constraints while ensuring robust analysis. This comprehensive approach encompasses dimensionless analysis, strategies for handling data scarcity, rigorous data preprocessing, and the implementation of various machine learning models—each selected based on theoretical foundations and potential applicability to the complex physical phenomena underlying erosion processes.

5.3.1 Addressing Challenges of Limited Dataset in WDE

Modeling WDE incubation period often contends with the inherent challenge of limited experimental data, a common constraint in materials science due to the resource-intensive nature of testing. To mitigate this, our methodology strategically incorporated approaches well-suited for small datasets. We leveraged dimensionless analysis to reduce variable dimensionality and embed physical principles, thereby constraining the model search space with domain knowledge. Principal Component Analysis (PCA) further addresses data sparsity and multicollinearity by transforming features into a reduced set of orthogonal components that capture essential variance. Furthermore, we explore advanced machine learning techniques, including a neural network designed to achieve generalization from limited data through extended training, and Tabular Prior-data Fitted Networks (TabPFN), which utilize meta-learning from diverse synthetic datasets to enable robust predictions without extensive task-specific fine-tuning. The collective aim of these strategies is to enhance model robustness, interpretability, and generalizability despite the limited dataset size.

5.3.2 Dataset Description and Characteristics

The dataset utilized in this study comprises 56 experimental measurements of water droplet erosion incubation periods for five metal alloys (Ti-6Al-4V, 17-4 PH, 12% Cr steel, Al-2024, and Al-7075) under varying impact conditions.

These materials include Ti-6Al-4V, 17-4 PH stainless steel, 12% Cr steel, and two aluminum alloys (Al-2024 and Al-7075), representing a diverse range of engineering materials commonly subjected to water droplet erosion in practical applications [7, 32]. The impact velocity in these experiments spans from 140 to 350 m/s, with droplet diameters of 0.463 and 0.603 mm. The observed incubation periods exhibit substantial variation, from as brief as 1 minute to as extended as 227 minutes, reflecting the complex interplay between material properties and impact conditions [2,3,81]. This wide variation in material types and testing parameters provides a robust foundation for developing predictive models with broad applicability across different engineering scenarios where water droplet erosion is a concern.

Figure 5.1 shows the distributions of input parameters and output variables in the dataset used in this work for each experimental data point, nine parameters were recorded: droplet velocity (V_0) and diameter (d) as impact conditions; and modulus of elasticity (E), Vickers hardness (H_v), yield strength (σ_y), density (ρ), speed of sound (C), Poisson's ratio (ν), and plane strain fracture toughness (K_{IC}) as material properties. The incubation period (N_0), measured in minutes, serves as the output variable to be predicted. This comprehensive set of parameters captures the essential physical factors known to influence erosion behavior, as identified in previous studies [2, 81, 145].

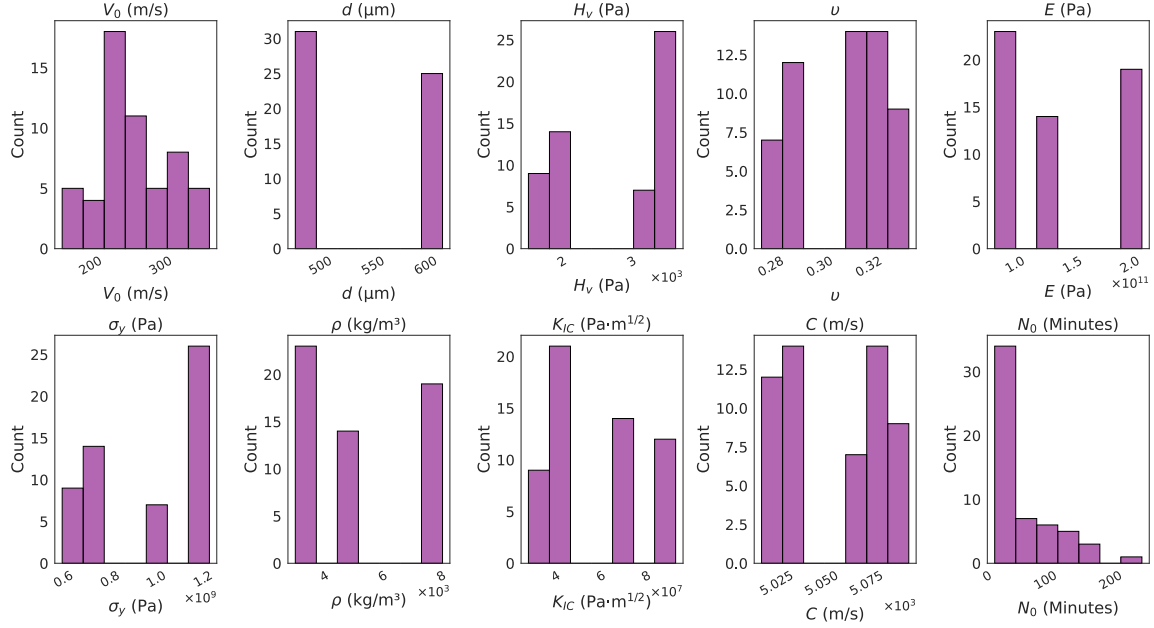


Figure 5.1. Distributions of input parameters and output variables in the dataset

5.3.3 Dataset Preprocessing and Feature Scaling

Before applying machine learning techniques, several preprocessing steps were implemented to enhance model performance and ensure meaningful results. The dataset was randomly divided into training (80%) and testing (20%) sets using a fixed random seed (42), with K-fold cross-validation employed to ensure robust model evaluation and mitigate the risk of overfitting [157]. This approach is critical given the relatively limited size of the dataset, a common constraint in erosion studies due to the time-consuming and resource-intensive nature of experimental testing.

For the PCA-based approach and transformer-based models TabNet, SAINT, FT-Transformer, and TabPFN, z-score standardization was applied to normalize input features. This technique transforms each feature x by subtracting the mean μ of the feature and dividing it by its standard deviation σ , resulting in a standardized value $z = (x - \mu)/\sigma$. This normalization ensures that each feature has a mean of zero and a standard deviation of one, which is crucial for providing stable and efficient learning in neural network-based models [132, 158].

Due to their sensitivity to small datasets, TabNet, SAINT, and FT-Transformer models were further enhanced through data augmentation techniques applied after scaling. These augmentations

aimed to expand the training distribution and reduce overfitting, which is a common issue for deep learning models trained on limited data. In contrast, TabPFN was used without augmentation, as it is specifically optimized for small tabular datasets and demonstrated robust generalization capabilities with only scaled data [74].

These preprocessed features were subsequently fed into their respective predictive modeling pipelines. Detailed descriptions of each model, including their underlying assumptions, architectures, and implementation strategies, are presented in later section (Section 5.3.8).

5.3.4 Dataset Augmentation

Advanced tabular deep learning models—such as FT-Transformer, SAINT, and TabNet typically require large datasets to generalize effectively. Given the limited dataset size (56 samples), data augmentation was essential to enhance their performance. Three augmentation strategies were evaluated:

- SMOTER (Synthetic Minority Over-sampling Technique for Regression) [159], which creates new samples by interpolating feature-target neighborhoods,
- Mixup [160], which linearly combines input-target pairs.
- Gaussian Noise (Noise Injection) Augmentation [161], which creates new samples by perturbing feature values with small Gaussian noise, thereby generating synthetic nearby observations.

SMOTER delivered the most consistent gains in predictive performance and stability without violating physical trends and was thus selected as the primary augmentation technique. All augmentations were applied exclusively to the training folds during cross-validation to prevent data leakage, and more than 1,000 synthetic points were generated to be used in FT-Transformer, SAINT, and TabNet methods.

5.3.5 Dimensionless Analysis for Incubation Period Estimation

Although accurate empirical correlations exist for erosion damage caused by solid particles [66], a reliable predictive model for WDE is still lacking due to the complexity of droplet–solid material interactions. WDE is influenced by many factors—including water properties, solid-material

characteristics, and operating conditions. Therefore, the present study adopts a physics-based strategy for estimating the incubation period N_0 using Buckingham- Π (Pi) dimensional analysis [146, 147].

Initially, 11 physically relevant variables influencing the incubation period are identified (as listed in Table 5.1). Due to the dependency between c_s , E , and ρ_s through ν , we did not retain ν as a fundamental variable for dimensionless analysis. Buckingham Pi analysis is then applied to systematically reduce the dimensional complexity. With 11 variables and 3 fundamental dimensions (mass M , length L , and time T), Buckingham Pi theory indicates that 8 independent dimensionless groups (Π -terms) govern the WDE problem.

The repeating variables chosen (droplet diameter d , impact velocity V_0 , and water density ρ_w) are dimensionally independent and collectively span the base dimensions of mass M , length L , and time T . Each remaining (nonrepeating) variable is rendered dimensionless by combining it with suitable powers of these three variables until the resulting Π -term has units $M^0 L^0 T^0$, i.e., unitless.

For instance, to construct the Π -term associated with fracture toughness K_{IC} (whose dimension is $ML^{-1/2}T^{-2}$), we write

$$\Pi_{K_{IC}} = [K_{IC}]d^a V_0^b \rho_w^c \quad (5.1)$$

Balancing dimensions gives

$$M^{1+c} L^{-1/2+a+b-3c} T^{-2-b} = M^0 L^0 T^0 \quad (5.2)$$

leading to the system $1 + c = 0$; $-\frac{1}{2} + a + b - 3c = 0$; $-2 - b = 0$.

Solving yields

$c = -1$; $b = -2$; $a = -\frac{1}{2}$, therefore,

$$\Pi_{K_{IC}} = \frac{K_{IC}}{\rho_w V_0^2 \sqrt{d}} \quad (5.3)$$

Table 5.1. Physical variables and dimensions involved in the application of the Pi-theorem

Variable	Symbol	Dimension	SI Unit
Incubation period	N_o	T	s
Droplet diameter	d	L	m
Impact velocity	V_o	LT^{-1}	m/s
Density of water	ρ_w	ML^{-3}	kg/m ³
Yield strength	σ_y	$ML^{-1}T^{-2}$	Pa
Vickers hardness	H_v	$ML^{-1}T^{-2}$	Pa
Young's modulus	E	$ML^{-1}T^{-2}$	Pa
Speed of sound (solid)	C_s	LT^{-1}	m/s
Fracture toughness	K_{Ic}	$ML^{-1/2}T^{-2}$	Pa.m ^{1/2}
Speed of sound (water)	C_w	LT^{-1}	m/s
Density of solid	ρ_s	ML^{-3}	kg/m ³

Repeating this procedure for each non-repeating variable yields a complete set of 8 independent, as shown in Table 5.1, dimensionless Π -terms, which govern the incubation period for WDE problem as follows:

$$\Pi_0 = \frac{N_o V_o}{d}, \Pi_1 = \frac{\sigma_y}{\rho_w V_o^2}, \Pi_2 = \frac{H_v}{\rho_w V_o^2}, \Pi_3 = \frac{E}{\rho_w V_o^2}, \Pi_4 = \frac{C_s}{V_o},$$

$$\Pi_5 = \frac{K_{Ic}}{\rho_w V_o^2 \sqrt{d}}, \Pi_6 = \frac{C_w}{V_o}, \Pi_7 = \frac{\rho_s}{\rho_w}$$

Building on Heymann's observation that the incubation period (N_0) follows a power-law relationship with the impact velocity, we assume a similar power-law relationship involving all relevant dimensionless parameters. A comprehensive dimensionless equation describing the incubation period (N_0) can thus be expressed as follows:

$$\Pi_0 = A(\Pi_1)^{\alpha_1}(\Pi_2)^{\alpha_2}(\Pi_3)^{\alpha_3}(\Pi_4)^{\alpha_4}(\Pi_5)^{\alpha_5}(\Pi_6)^{\alpha_6}(\Pi_7)^{\alpha_7} \quad (5.4)$$

and explicitly in terms of physical parameters as:

$$N_0 = A \frac{d}{V_0} \left(\frac{\sigma_y}{\rho_w V_0^2} \right)^{\alpha_1} \left(\frac{H_v}{\rho_w V_0^2} \right)^{\alpha_2} \left(\frac{E}{\rho_w V_0^2} \right)^{\alpha_3} \left(\frac{C_s}{V_0} \right)^{\alpha_4} \left(\frac{K_{IC}}{\rho_w V_0^2 \sqrt{d}} \right)^{\alpha_5} \left(\frac{C_w}{V_0} \right)^{\alpha_6} \left(\frac{\rho_s}{\rho_w} \right)^{\alpha_7} \quad (5.5)$$

where the coefficients A and exponents α_i could be calibrated through experimental datasets, potentially leading to physically interpretable model for estimating WDE incubation period.

Building upon physical insights, we further establish a set of dimensionless groups selected to model the incubation period of WDE. These groups are not merely mathematical constructs; they represent the relative importance of the loading mechanisms and material properties during droplet impact, thereby reducing model complexity and potentially improving predictive capability. For instance, yield strength σ_y is nondimensionalized using the initial *water-hammer pressure* during impact $p_{hw} = \rho_w C_w V_0$ [1,11], rather than the commonly used *dynamic pressure* $\rho_w V_0^2$ where ρ_w is the water density and V_0 is the droplet impact speed. This choice is motivated by the dominance of the initial water-hammer pulse in high-speed droplet impact, which better captures the elastic-plastic response of the material. Other mechanical properties, such as E , H_v , σ_y and K_{IC} were considered, but in the final formulation only E , H_v , the velocity ratios (C_s/V_0) , and (C_w/V_0) , and the solid-liquid density ratio (ρ_s/ρ_w) were retained. These were combined into the acoustic impedance ratio $Z = \rho_s C_s / \rho_w C_w$, which was reasonably adopted to simplify the model (Eq. 5.5) while potentially preserving physical accuracy. In addition, the rig's droplet-impact frequency f can be used as a time scale to nondimensionalize both N_0 and d/V_0 with $N_0 = f N_0$ and $f d/V_0$, and the constant factor A can be absorbed into a scale parameter B . Thus, the dimensionless predictive equation simplifies to:

$$N_0 = B \left(\frac{d}{V_0} \right)^{\beta_0} \left(\frac{\sigma_y}{\rho_w C_w V_0} \right)^{\beta_1} \left(\frac{K_{IC}}{\rho_w V_0^2 \sqrt{d}} \right)^{\beta_2} \left(\frac{E}{\rho_w V_0^2} \right)^{\beta_3} \left(\frac{\rho_s C_s}{\rho_w C_w} \right)^{\beta_4} \quad (5.6)$$

where B and β_i are the fitted parameters of the simplified model. To better capture the influence of the velocity, we introduced an exponent as an additional fitting parameter. Our optimization shows that this improves the model's ability to match the data, as will be demonstrated in the results section.

5.3.6 Multiple Linear Regression with Dimensionless Analysis

This section combines dimensional analysis with multiple linear regression (MLR) to develop predictive, dimensionally consistent models. To predict N_o regression is performed on the logarithms of these Π groups, allowing the model to capture the underlying scaling laws of water droplet erosion while remaining interpretable and generalizable across different materials and testing conditions. To linearize this relationship for regression analysis, a logarithmic transformation was applied to both sides of Eq. (5.4):

$$\log \pi_o = \log A + \alpha_1 \log \Pi_1 + \alpha_2 \log \pi_2 + \alpha_3 \log \pi_3 + \alpha_4 \log \pi_4 + \alpha_5 \log \pi_5 + \alpha_6 \log \pi_6 + \alpha_7 \log \pi_7 \quad (5.7)$$

This transformation enables the application of MLR techniques to determine the exponents, which represent the sensitivity of the incubation period to each dimensionless group [12]. The regression was performed using the ordinary least squares method, which minimizes the sum of squared residuals between predicted and actual values.

5.3.7 Principal Component Analysis for Neural Network

Principal Component Analysis (PCA) [151, 152], in combination with neural network (NN) [153, 154], addresses challenges related to multicollinearity and high dimensionality, while simultaneously capturing complex nonlinear patterns within the dataset. This hybrid methodology is particularly advantageous in contexts with limited data, where conventional deep learning models may encounter difficulties achieving robust generalization.

The implementation commenced by standardizing all input features using z-score normalization, ensuring that each feature contributed equally to the principal component analysis (PCA) process [150]. Subsequently, PCA was applied to transform the standardized features into a set of

uncorrelated principal components. These components, linear combinations of the original variables, are ordered according to the amount of variance they explain in the dataset.

Principal components were extracted by computing the eigenvectors and eigenvalues of the covariance matrix derived from the standardized input data. The eigenvectors identified directions of maximum variance, while the corresponding eigenvalues quantified the proportion of variance captured by each component [151]. The analysis demonstrated that the first five principal components accounted for more than 90% of the total variance. This indicated a substantial reduction in dimensionality, enabling more efficient modeling while preserving the core informational content of the original dataset.

These selected principal components were then utilized as inputs to a neural network [153, 154], providing a reduced input dimensionality that helps the model converge more quickly. This behavior is especially beneficial for predicting the incubation period in WDE, where the available data is limited yet strongly influenced by underlying physical relationships.

A compact multilayer perceptron (MLP) was designed to leverage this phenomenon. The network architecture (Figure 5.2) consisted of an input layer that accepted five principal components (PC_1 , PC_2 , PC_3 , PC_4 , PC_5), followed by two hidden layers with 10 and 5 units, respectively, each employing “LeakyReLU” activation functions with a negative slope of 0.1 and dropout enabled. A final output layer was responsible for predicting the target variable N_o . The network's size was intentionally kept modest to prevent overfitting while retaining enough flexibility to model nonlinear relationships. Training was conducted over an extended number of epochs. The optimization process utilized the Adam optimizer [162] with a learning rate of 0.006 and a weight decay of 0.0001, employing Mean Squared Error (MSE) as the loss function. Dropout regularization was incorporated throughout to support the emergence of the NN model by preventing complex co-adaptations of neurons in favor of simpler, more generalizable ones.

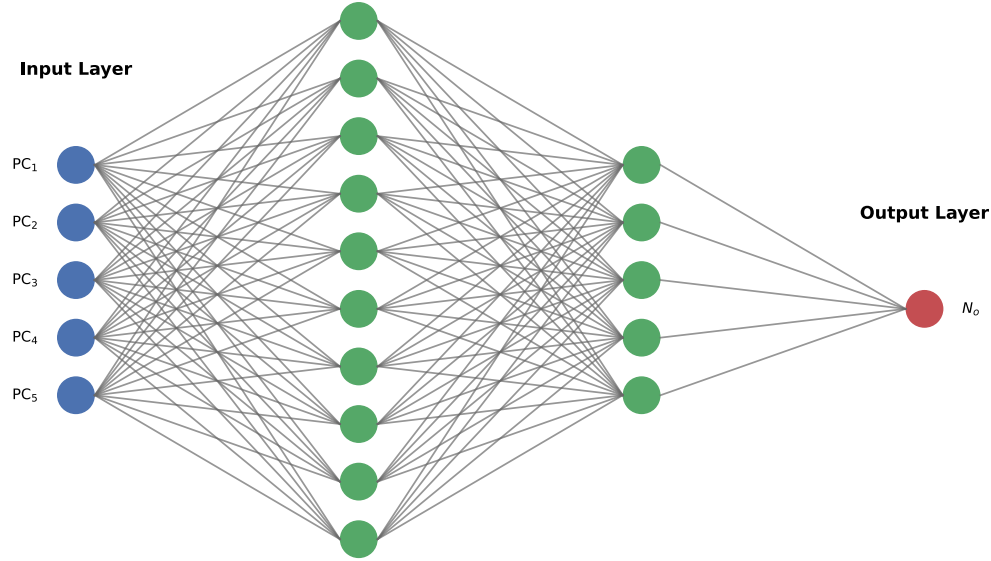


Figure 5.2. PCA-NN architecture for incubation period prediction

To assess model performance robustly despite the small dataset size, a 5-fold cross-validation approach was employed. This technique divides the dataset into five equal parts, iteratively training the model on four parts while testing it on the remaining one. Each data point thus contributes to both the training and validation phases, yielding a comprehensive evaluation of the model's generalizability. For every fold, a new instance of the neural network was trained, and the model's predictive performance was quantified using standard regression metrics: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The final evaluation was based on the average of these metrics across all five folds [109].

5.3.8 Advanced Deep Learning Architectures for Tabular Data

The second category of approaches involves advanced machine learning models designed explicitly for tabular data. These models are efficient in capturing complex feature interactions and

patterns, particularly in scenarios with limited data availability. All models in this category employed z-score standardization for preprocessing and utilized K-fold cross-validation to ensure reliable and generalizable results.

In this work, four representative models from this category were implemented and evaluated: FT-Transformer, TabNet, SAINT, and TabPFN. Each model was selected based on its architectural innovation and proven effectiveness in handling tabular data, especially under data-scarce conditions. The following subsections provide a detailed description of each of these models, highlighting their core architectural features and the rationale for their inclusion in this study.

5.3.8.1 Feature Tokenizer Transformer (FT-Transformer)

The Feature Tokenizer Transformer (FT-Transformer) [70] is implemented using the “rtdl” library (Representation Learning for Tabular Data), a PyTorch-based framework. This architecture adapts the transformer model to tabular data by tokenizing individual features—treating each feature as an input token—while preserving their contextual relationships through self-attention mechanisms.

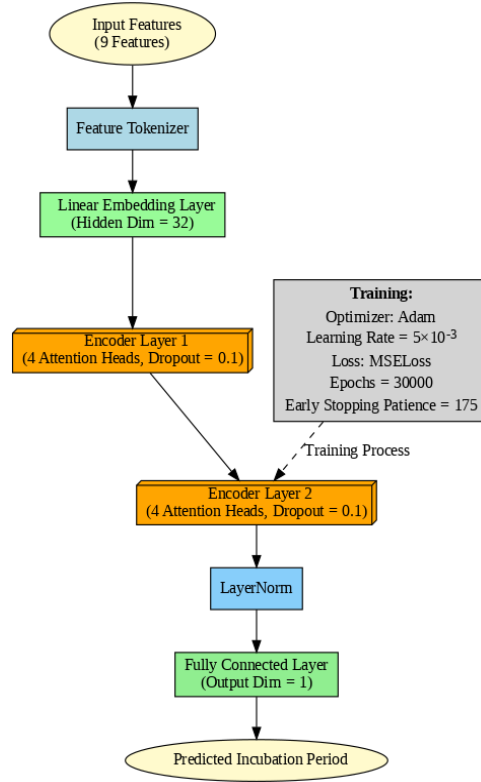


Figure 5.3. Feature tokenization-based transformer for incubation period prediction

The implementation illustrated in Figure 5.3 includes a feature tokenization process for embedding numerical features (i.e., 9), a linear embedding layer with a hidden dimension equal to (32), and two encoder layers, each comprising four attention heads. The model was optimized using the Adam optimizer with a learning rate of 5×10^{-3} and was trained for 30,000 epochs using mini-batch updates to promote convergence.

5.3.8.2 TabNet Architecture for Regression

Arik and Pfister [72] proposed and implemented the TabNet model for regression. This architecture combines decision-tree-like feature selection with deep learning via a sequential attention mechanism. Hyperparameter optimization was conducted to achieve the best model accuracy. The process is explained in Figure 5.4.

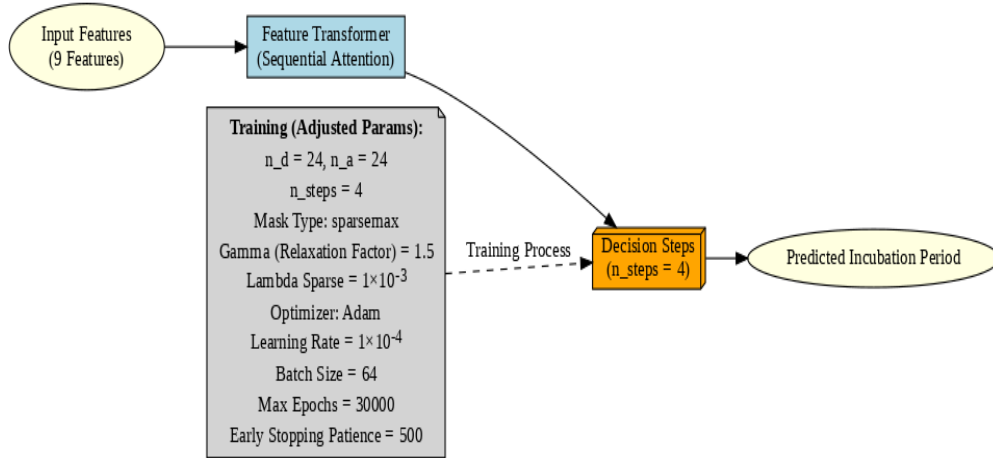


Figure 5.4.TabNet method for incubation period prediction

The key tuning parameters for TabNet include decision and attention widths (24), number of decision steps (4), relaxation factor (1.5), sparsity regularization strength (1×10^{-3}), learning rate (1×10^{-4}), and batch size (64). The final model utilized sparsemax masks to enhance interpretability and enable automatic feature selection. Early stopping was employed with a patience of 500 epochs, and, similar to FT-Transformer, the model was trained for 30000 epochs using the optimal hyperparameters.

5.3.8.3 SAINT (Self-Attention and INtersample Attention Transformer)

SAINT (Self-Attention and INtersample Attention Transformer), described by Somepalli et al. [71], was implemented as a transformer-based architecture tailored for tabular regression tasks. The model architecture in Figure 5.5 includes a linear embedding layer that projected features into hidden dimensions (32), two encoder layers, and four attention heads.

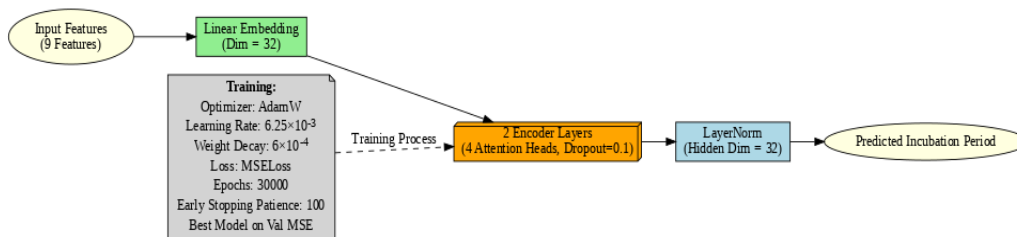


Figure 5.5.SAINT method for incubation period prediction

Dropout regularization with a rate of 0.1 was applied, whereby 10% of neurons were randomly deactivated during training. This strategy was employed to enhance neural network performance and convergence by reducing overfitting and encouraging more robust feature learning. Hyperparameter tuning further explored learning rates (6.25×10^{-3}), weight decay (6×10^{-4}), and the number of layers, attention heads, dropout rates, and batch sizes. Each configuration was trained for up to 30,000 epochs. The best-performing setup was determined through early stopping based on the validation mean squared error (MSE). Validation MSE measures the average squared difference between predicted and actual values on the validation dataset during training. Finally, the selected model's performance was evaluated on the test dataset using MSE to assess its generalization beyond the training and validation phases.

5.3.8.4 Tabular Prior-data Fitted Networks (TabPFN)

Tabular Prior-data Fitted Networks (TabPFN), introduced by Hollmann et al. [74], employ a zero-shot meta-learning strategy [156], allowing the model to make accurate predictions on new tasks without any fine-tuning or additional gradient updates. This pre-trained transformer-based model, shown in Figure 5.6, generalizes from prior knowledge learned on synthetic datasets, making it particularly effective for small datasets by reducing the risk of overfitting and enabling rapid, reliable inference. The model was implemented using the ‘tabpfn’ client library.

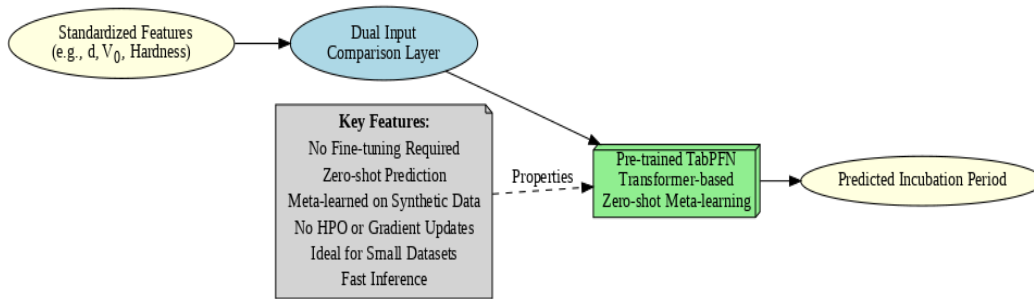


Figure 5.6. TabPFN method for incubation period prediction

The TabPFN model was tested using standardized input features, including droplet diameter, impact velocity, and material hardness. Unlike conventional models, TabPFN requires no

additional training or hyperparameter tuning, making it especially well-suited for small datasets. This capability reduces the risk of overfitting while enabling rapid and reliable predictions.

5.3.9 Model Evaluation and Validation

To ensure a fair and consistent comparison, all models were evaluated using standard regression metrics calculated on the test dataset [163]. These include the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The R^2 metric indicates how well the input features explain the variation in the target variable, with values closer to 1 suggesting effective capture of underlying data patterns. RMSE estimates the average magnitude of prediction errors, with larger errors receiving greater weight due to the squaring operation. MAE calculates the average absolute difference between predicted and actual values, providing an intuitive measure of prediction accuracy that does not overemphasize larger errors. These metrics collectively assess each model's accuracy and robustness, particularly emphasizing the generalization of unseen data.

To further evaluate model generalizability beyond the training and testing phases, an additional validation was conducted using a separate set of previously unseen data. This validation set comprised three test cases with distinct impact velocities of 350, 300, and 250 m/s, using 463 μm droplets impacting austenitic stainless steel 304L. The corresponding observed incubation periods were 3, 5, and 30 minutes, providing a practical test of model performance across varying impact conditions for a material not included in the original dataset. Water droplet erosion tests were conducted using a custom erosion test rig equipped with an 18-inch rotating disc system following ASTM G73 [23] standard procedures. Test samples were mounted diametrically opposite to reference samples on the rotating disc, which was driven by a compressed air turbine to achieve the desired linear impact velocities of 250, 300, and 350 m/s. The test chamber was evacuated to 30-50 mbar vacuum pressure to minimize air friction effects during testing. Deionized water droplets with an average diameter of 463 μm were generated using a 400 μm orifice nozzle positioned 50 mm above the samples, maintaining a water line pressure of 30 psi and a flow rate of 0.05 L/min. Each erosion cycle involved accelerating the disc to the target speed, introducing water droplets for a predetermined duration, then gradually decelerating the system. Mass loss

measurements were recorded after each cycle using a precision balance accurate to 0.2 mg, and multiple cycles constituted a complete erosion test for each impact velocity condition.

This comprehensive methodology, which integrates dimensionless analysis, principal component analysis with neural networks, and advanced machine learning techniques, addresses the limitations of traditional approaches to predicting the WDE incubation period identified in the literature review. By systematically comparing these diverse approaches, this study aims to determine the most effective approach and gain insights into the physical mechanisms governing erosion behavior. This will ultimately contribute to improved material selection, component design, and maintenance planning in applications where water droplet erosion is encountered [3].

5.4. Results and Discussion

This section presents the performance of various machine learning models used to predict the incubation period (N_0) in water droplet erosion. The results are evaluated using standard regression metrics and visual analyses to assess accuracy, generalization capability, and physical interpretability.

5.4.1 Physics-Informed Modeling and Dimensionless Analysis

To develop a predictive model for the water droplet erosion incubation period N_0 , we employed Buckingham- Π dimensional analysis. As derived in section 5.3.5, the incubation period depends fundamentally on the following relevant properties: droplet impact velocity V_0 , droplet diameter d , material yield strength σ_y , fracture toughness K_{IC} , speed of sound in both the material (C_s) and water (C_w), material density (ρ_s), and water density (ρ_w). By fitting the model in Eq. (5.3) on the experimental data, we obtain the dimensionless power-law model for the incubation period, N_0 (in minutes):

$$N_0 = 0.01 \left(\frac{d}{V_0} \right)^{1/2} \left(\frac{\sigma_y}{\rho_w C_w V_0} \right)^{3/2} \left(\frac{K_{IC}}{\rho_w V_0^2 \sqrt{d}} \right)^{1/2} \left(\frac{E}{\rho_w V_0^2} \right)^{3/2} \left(\frac{\rho_s C_s}{\rho_w C_w} \right)^{-1/4} \quad (5.5)$$

Figure 5.7 illustrates the comparison between the measured and predicted values of N_0 (in minutes) on a logarithmic scale. Although the log-scale correlation shows a strong relationship ($R_{\log}^2 = 0.83$), the linear-scale correlation is weaker ($R^2 = 0.62$). Thus, integrating dimensionless features

into machine learning models may be an effective strategy for enhancing the accuracy of incubation period predictions.

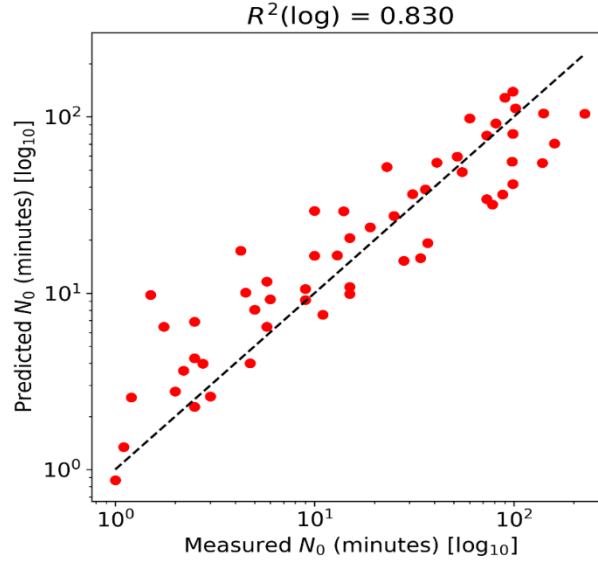


Figure 5.7. Comparison between measured and predicted incubation period N_o (in minutes), displayed on a logarithmic scale

Finally, it is worth noting that the derived relationship indicates that the incubation period is inversely proportional to the velocity $N_o \propto V_o^{-6}$ ($N_o \propto \left(\frac{C1}{V_o}\right)^{1/2} \left(\frac{C2}{V_o}\right)^{3/2} \left(\frac{C3}{V_o^2}\right)^{1/2} \left(\frac{C4}{V_o^2}\right)^{3/2}$), indicating that increased impact velocity significantly reduces the incubation period [2, 14, 145]. This is consistent with WDE experimental observations, where higher V_o raises impact pressure, leading to faster erosion [2, 145]. This model exemplifies the utility of physics-informed, dimensionless analysis for predicting WDE incubation time.

5.4.2 ML Model with Dimensionless Groups

To predict N_o , a MRL regression model (ridge) was trained on dimensionless groups derived via the Buckingham PI theorem (Section 5.3.6). Using these groups, the model achieved $R^2 = 0.93$ and $MAE = 7.50$ min (Figure 5.8a). However, when trained on the original dimensional features, performance declined to $R^2 = 0.61$ and $MAE = 21.81$ min (Figure 5.8b), indicating that dimensionless representations improve predictive accuracy and enhance physical interpretability of the ML model.

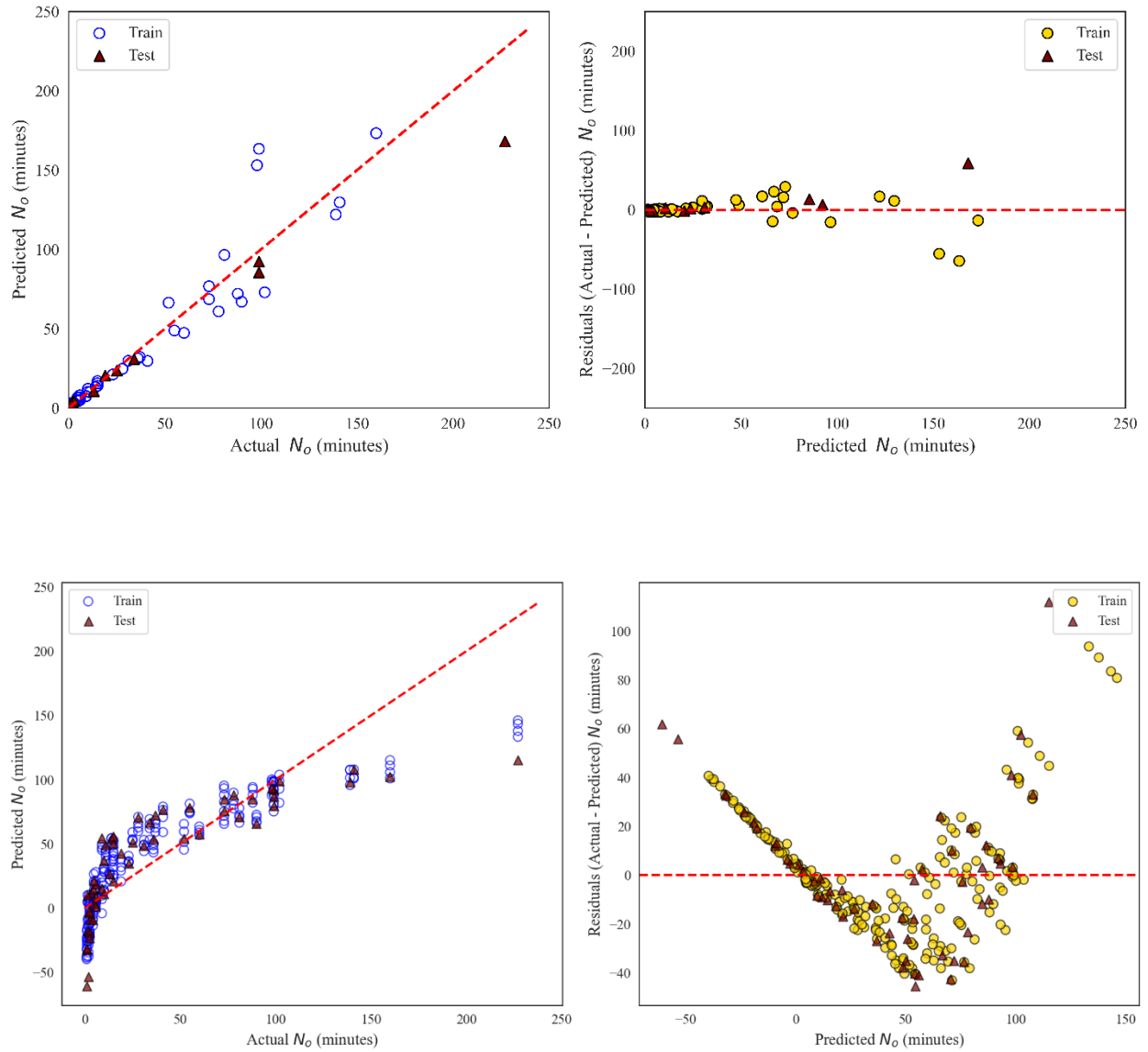


Figure 5.8. MLR model performance for N_o (minutes) using (top) dimensionless groups from the Buckingham PI theorem and (bottom) the original features

Finally, errors increase for long incubation times, indicating the need for a better model or for applying the data transformations adopted in Chapter 3 to stabilize variance and improve long-duration accuracy.

5.4.3 Principal Component Analysis for Improving Neural Network Performance

The PCA (Principal Component Analysis) Explained Variance plot (Figure 5.9) demonstrates how much of the total variance in the dataset is captured as more principal components are included. The x-axis indicates the number of principal components considered, while the y-axis shows the cumulative explained variance ratio, which rises from approximately 0.60 to nearly 1.0.

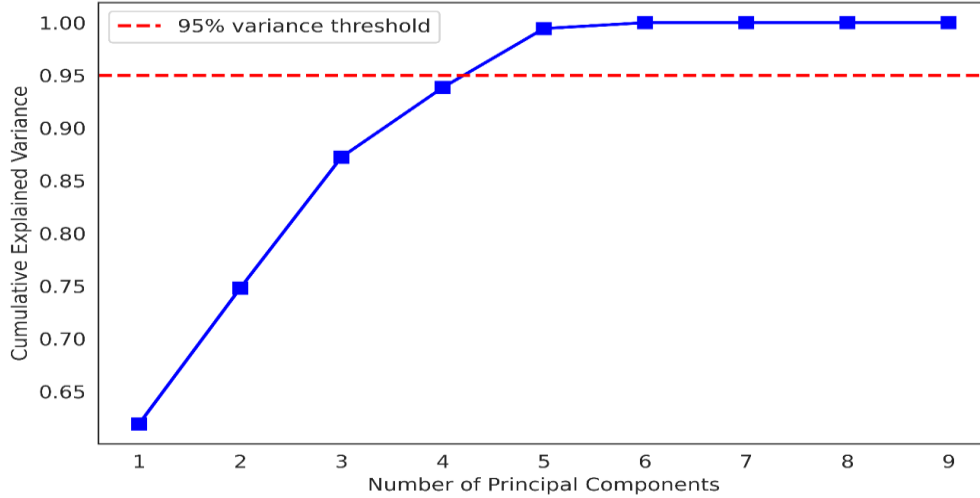


Figure 5.9. Cumulative explained variance ratio as a function of the number of principal components

The red reference line at 0.95 indicates the threshold for retaining most information [150]. Analysis confirms that five principal components capture over 95% of the dataset’s variability, effectively reducing dimensionality while preserving essential content.

PCA reduced the original nine input features to five orthogonal components to enhance generalization and mitigate overfitting—an essential step given the limited WDE dataset. These components were used to efficiently train a multilayer perceptron (MLP), resulting in rapid convergence [154, 155].

In the present study, the network was trained for 3000 epochs using the Adam optimizer (learning rate: 6×10^{-3}) and L2 regularization (weight decay 1×10^{-4}) to penalize overly complex weights, encouraging simpler solutions that capture nonlinear interactions among material properties (e.g., hardness, fracture toughness, elastic modulus, density) and impact velocity.

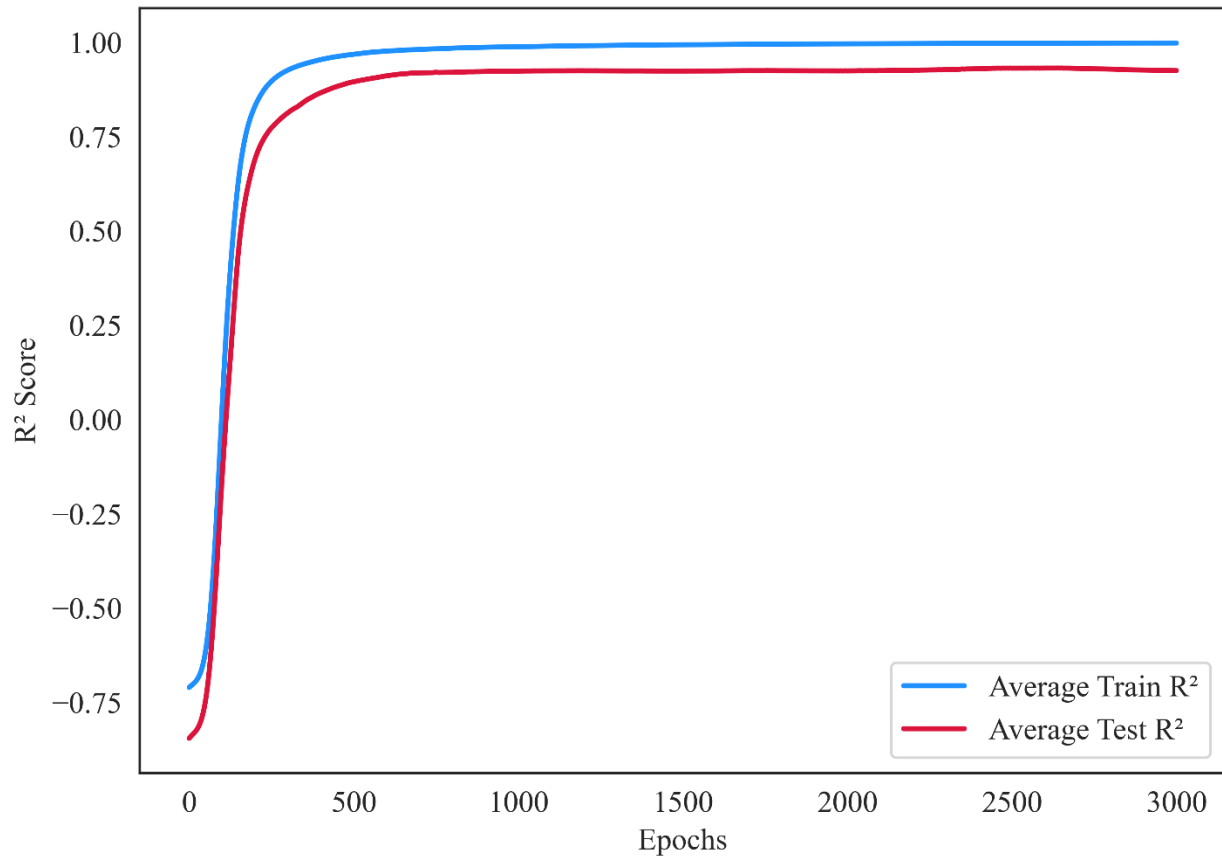


Figure 5.10. Training and testing process for the PCA-NN model on the WDE dataset

Figure 5.10 shows the Coefficient of Determination (R^2) over training epochs for both training (blue) and validation (red) datasets after applying PCA to the input features. PCA significantly accelerates the learning process by reducing noise and redundant correlations, allowing the model to identify the most informative components early in training. As a result, both curves rise steeply, with the training R^2 approaching 1.0 and the test R^2 surpassing 0.85 within the first few hundred epochs.

PCA effectively reduces the dimensionality of the feature space, enhancing model generalization by focusing on the most informative components. This compression minimizes overfitting, stabilizes training, and promotes robust performance with fewer epochs, leading to faster convergence and improved generalization.

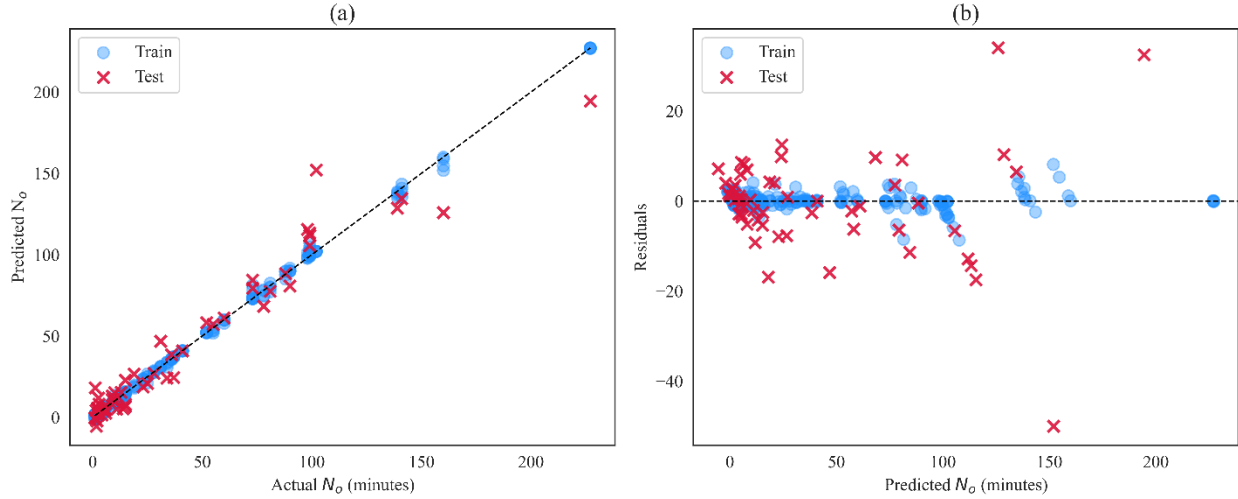


Figure 5.11. PCA-NN model: (a) Actual vs. predicted N_o values (minutes) for the test set; (b) Residual plot

Through complementary visualizations, Figure 5.11 evaluates the PCA-NN model's predictive performance for the incubation period. Figure 5.11(a) presents predicted versus actual N_o values, with training data (blue) exhibiting closer adherence to the 45° reference line compared to testing data (red), which displays greater scatter—indicating reduced generalization capability. Both datasets demonstrate increasing vertical dispersion at higher N_o values, reflecting diminished predictive precision in extended incubation periods. Quantitatively, the PCA-NN approach achieved an R^2 of 0.91, MAE of 7.72, and RMSE of 14.20, confirming strong overall predictive ability while revealing limitations in capturing variability at extreme values.

Figure 5.11(b) provides residual analysis (Actual – Predicted N_o), revealing critical error patterns: residuals cluster tightly around zero for shorter incubation periods, signifying accurate predictions, but become increasingly dispersed at higher N_o values. This pattern indicates systematic overestimation bias in extended incubation periods, highlighting a fundamental limitation in capturing extreme-value behavior under data scarcity conditions. The observed deterioration in predictive accuracy—characterized by both increased scatter magnitude and directional bias—persists across modeling approaches, including multiple linear regression (MLR) with dimensionless parameters and PCA-NN. This consistent challenge underscores the necessity for advanced tabular data modeling techniques with enhanced generalization capacity, particularly

when working with limited experimental datasets. Subsequent sections evaluate the efficacy of such approaches in addressing these identified limitations.

5.4.4 Advanced Tabular Learning Architectures

We evaluated advanced deep learning models—FT-Transformer [70], TabNet, SAINT [164], and TabPFN [156]—optimized for tabular data to predict WDE incubation time. These state-of-the-art architectures excel at capturing complex feature interactions, crucial for WDE given limited data and nonlinear relationships among velocity, droplet size, and material properties (e.g., plane strain fracture toughness, elastic modulus). Validation revealed that while these models achieve high fitting performance, their generalization to novel material systems remains challenging, underscoring the need for robust cross-validation beyond standard train-test splits.

5.4.4.1 FT-Transformer

Using a Feature Tokenizer architecture, FT-Transformer converted tabular features into embeddings for transformer-based processing, with attention mechanisms capturing nonlinear interactions. It achieved an excellent fit. Figure 5.12 evaluates the FT-Transformer model's predictive performance for the N_o through three complementary visualizations.

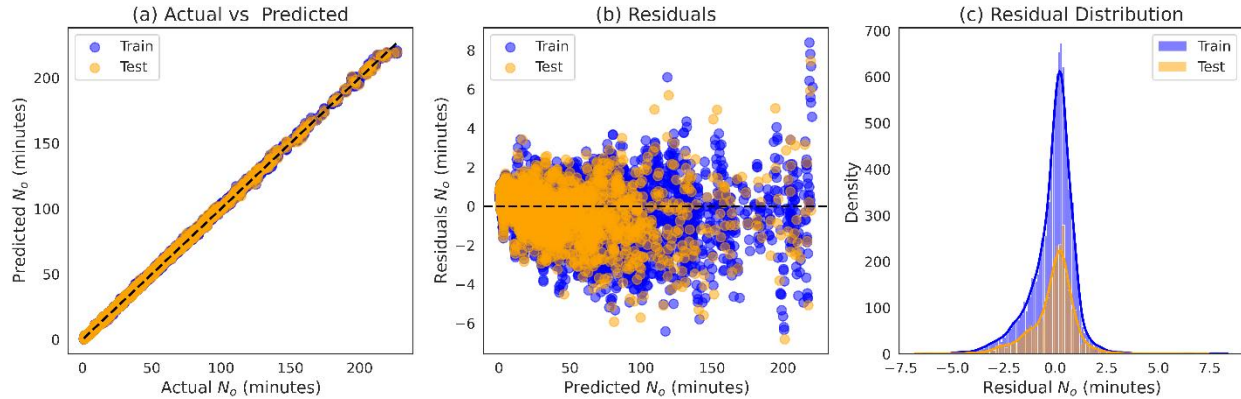


Figure 5.12. FT-Transformer: (a) Actual vs. predicted N_o values (minutes); (b) Residual plot; (c) Residual distribution

Figure 5.12(a) presents predicted versus actual N_o values, demonstrating near-perfect alignment with the diagonal reference line, with only minor deviations observed in the 150–200-minute range. This indicates exceptional overall predictive accuracy with localized precision limitations

at extended incubation periods. Figure 5.12(b) shows residuals that are symmetrically distributed around zero, with a slight asymmetry in the test set, suggesting room for further refinement. Figure 5.12(c) presents a residual distribution analysis, illustrating the pronounced, zero-centered distributions of both datasets. The symmetry and concentration of residuals around zero confirm negligible systematic bias and substantiate the model's strong generalization performance across both seen and unseen data. This represents a significant improvement over the PCA-NN approach evaluated in Figure 5.11.

5.4.4.2 TabNet

The TabNet model exhibited strong predictive performance, reflecting its ability to explain a significant share of the target variable's variance with relatively low average prediction error. Figure 5.13 illustrates the TabNet model's predictive performance for the incubation period (N_0) using three complementary visualizations.

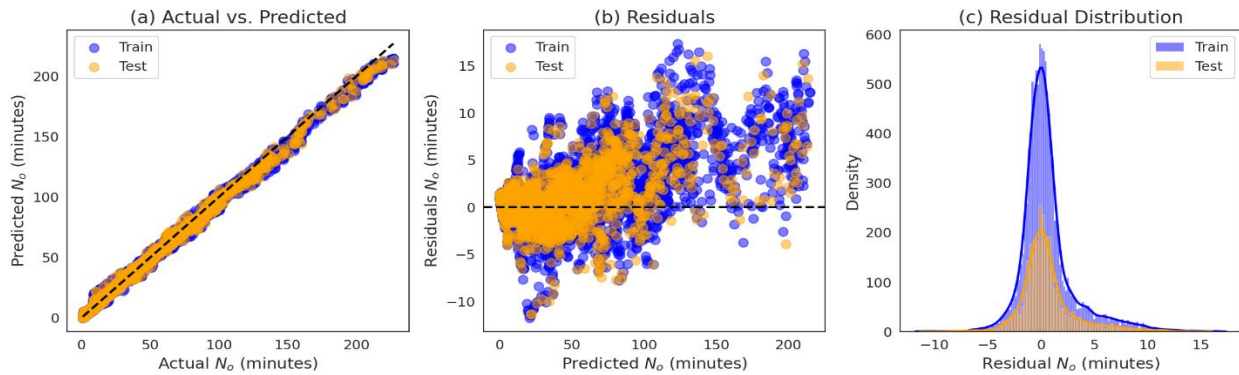


Figure 5.13. TabNet: (a) Actual vs. predicted N_0 values (minutes); (b) Residual plot; (c) Residual distribution

Figure 5.13(a) presents predicted versus actual N_0 values, demonstrating close alignment along the diagonal reference line across most of the data range. This indicates robust predictive agreement with localized precision variations. Figure 5.13(b) displays residual analysis, revealing the distribution of prediction errors across different N_0 . Figure 5.13(c) provides residual distribution analysis, illustrating a sharp, symmetric distribution centered at zero. This pattern confirms that most prediction errors are small and unbiased, with only a few larger deviations appearing in the distribution tails. The rarity of these extreme errors highlights the model's overall robustness and

represents continued advancement in predictive performance compared to previous approaches, including the FT-Transformer evaluated in Figure 5.12.

5.4.4.3 SAINT

The SAINT model, which incorporates inter-sample attention to capture relationships between samples with similar material behaviors, demonstrates performance comparable to the FT-Transformer. Figure 5.14 evaluates the SAINT model's predictive performance for incubation period through three complementary visualizations. Figure 5.14(a) presents predicted versus actual N_o values, demonstrating strong alignment along the diagonal reference line for both training and test datasets. This confirms the model's overall predictive accuracy, though a slight increase in dispersion is observed at higher N_o values. These maximum deviations remain within acceptable limits, particularly when considering the relatively high experimental errors inherent in water droplet erosion (WDE) measurements.

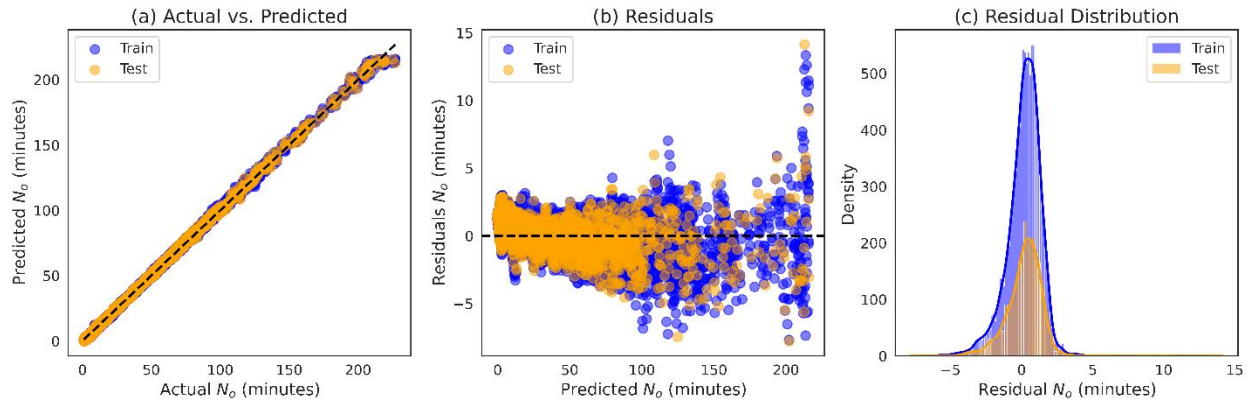


Figure 5.14. SAINT: (a) Actual vs. predicted N_o values (minutes); (b) Residual plot; (c) Residual distribution

Figure 5.14(b) displays residual analysis, revealing larger residuals for the test set compared to the training set. This indicates reduced generalization capability, though the differences remain small and within acceptable performance bounds. Figure 5.14(c) provides residual distribution analysis, highlighting the overall pattern of prediction errors across both datasets. The SAINT model demonstrates competitive performance in the methodological progression, showing improvements over the PCA-NN approach while maintaining robustness comparable to the FT-Transformer and TabNet models evaluated previously.

5.4.4.4 TabPFN

The TabPFN model, representing an advanced approach for tabular data prediction, was evaluated using K-fold cross-validation to assess its performance in predicting the incubation period (N_o). Figure 5.15 presents the comprehensive results through three complementary visualizations.

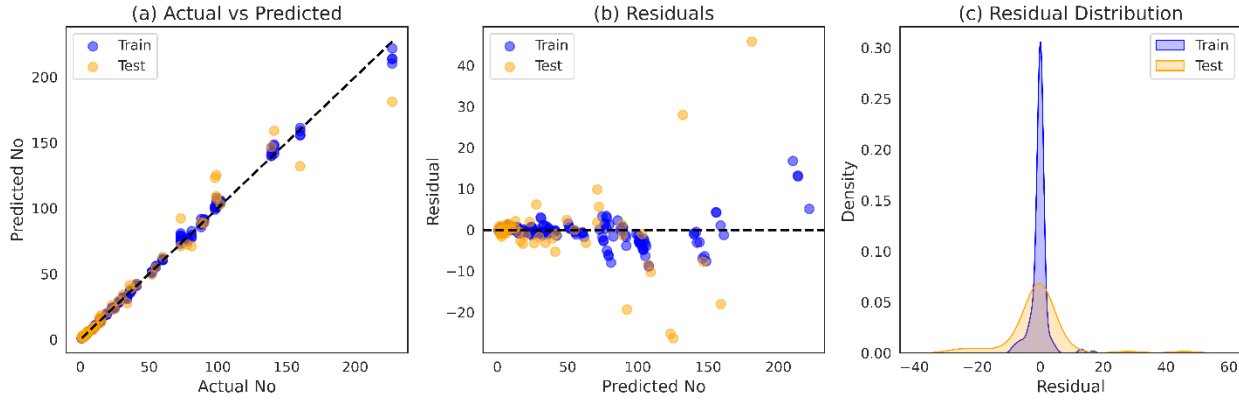


Figure 5.15. TabPFN: (a) Actual vs. predicted N_o values (minutes); (b) Residual plot; (c) Residual distribution

Figure 5.15(a) displays predicted versus actual N_o values, demonstrating close alignment along the ideal diagonal reference line for both training and test datasets. This indicates accurate predictions across data splits, confirming the model's strong performance under cross-validation conditions. Figure 5.15(b) presents residual analysis, revealing residuals randomly scattered around zero across the predicted range. Figure 5.15(c) provides residual distribution analysis, showing a sharply peaked, zero-centered distribution for training data, contrasted with a broader spread in the test set. This differential behavior highlights the model's strong predictive capacity on training data while also reflecting the presence of some generalization error when applied to unseen test data. The TabPFN model demonstrates mixed performance within the methodological progression, showing unbiased predictions but with substantial error magnitudes that may limit its practical utility compared to models with more constrained error ranges.

5.4.5 Model Comparison

Following the evaluation of multiple linear regression (MLR), PCA-NN, and advanced tabular deep learning models (FT-Transformer, TabNet, SAINT, and TabPFN) for predicting the incubation period (N_o) in water droplet erosion (WDE), this section compares their predictive

performance. Table 5.2 summarizes their results on the primary test set using standard regression metrics (R^2 , MAE, RMSE).

Table 5.2. Performance comparison of different models

Model	R^2	MAE (min)	RMSE (min)
MLR on Dimensionless Groups	0.93	7.50	17.20
PCA-NN approach on Original Features	0.91	7.72	14.20
FT-Transformer on Original Features	0.99	0.80	1.15
TabNet on Original Features	0.99	1.50	2.20
SAINT on Original Features	0.99	0.85	1.25
TabPFN on Original Features	0.96	4.61	9.30

Among the transformer-based models, FT-Transformer achieved the strongest performance ($R^2 = 0.99$, MAE = 0.80 min, RMSE = 1.15 min), reflecting its ability to capture complex nonlinear dependencies. SAINT matched this accuracy, with slightly higher errors (MAE = 0.85 min, RMSE = 1.25 min), while TabNet also reached $R^2 = 0.99$ but with larger error margins (MAE = 1.50 min, RMSE = 2.20 min).

TabPFN, a task-agnostic meta-learning model, performed well ($R^2 = 0.96$, MAE = 4.61 min, RMSE = 9.30 min) despite the absence of task-specific fine-tuning. In contrast, PCA-NN ($R^2 = 0.91$) and MLR ($R^2 = 0.93$) delivered less accuracy but offer interpretability: PCA-NN captures key data structure through dimensionality reduction, while MLR provides physics-informed insight via dimensionless groups. In summary, the test set results highlight the superiority of transformer-based models, with FT-Transformer and SAINT achieving the highest predictive accuracy, TabNet balancing accuracy with interpretability, and TabPFN demonstrating the promise of meta-learning approaches.

5.4.6 Validation of Incubation Period Prediction for Austenitic Stainless Steel 304L

To evaluate model generalizability beyond the training data, external validation was conducted using a separate dataset comprising three unseen test cases (impact velocities of 350, 300, and 250 m/s with 463 μm droplets impacting 304L austenitic stainless steel. Figure 5.16 presents model predictions relative to experimental incubation times.

TabPFN achieved the strongest generalization on external dataset prediction ($R^2 = 0.948$, MAE = 2.25 min), producing symmetric, zero-centered residuals. FT-Transformer ($R^2 = 0.929$, MAE = 2.65 min) and SAINT ($R^2 = 0.941$, MAE = 2.45 min) also performed well, maintaining balanced predictions across cases. TabNet, while highly accurate on the training dataset, showed reduced generalization ($R^2 = 0.909$, MAE = 3.67 min).

In contrast, PCA-NN ($R^2 = 0.251$, MAE = 7.83 min) displayed wider residual spreads, while MLR failed to generalize ($R^2 = -0.971$, MAE = 12.22 min), consistently underestimating incubation times.

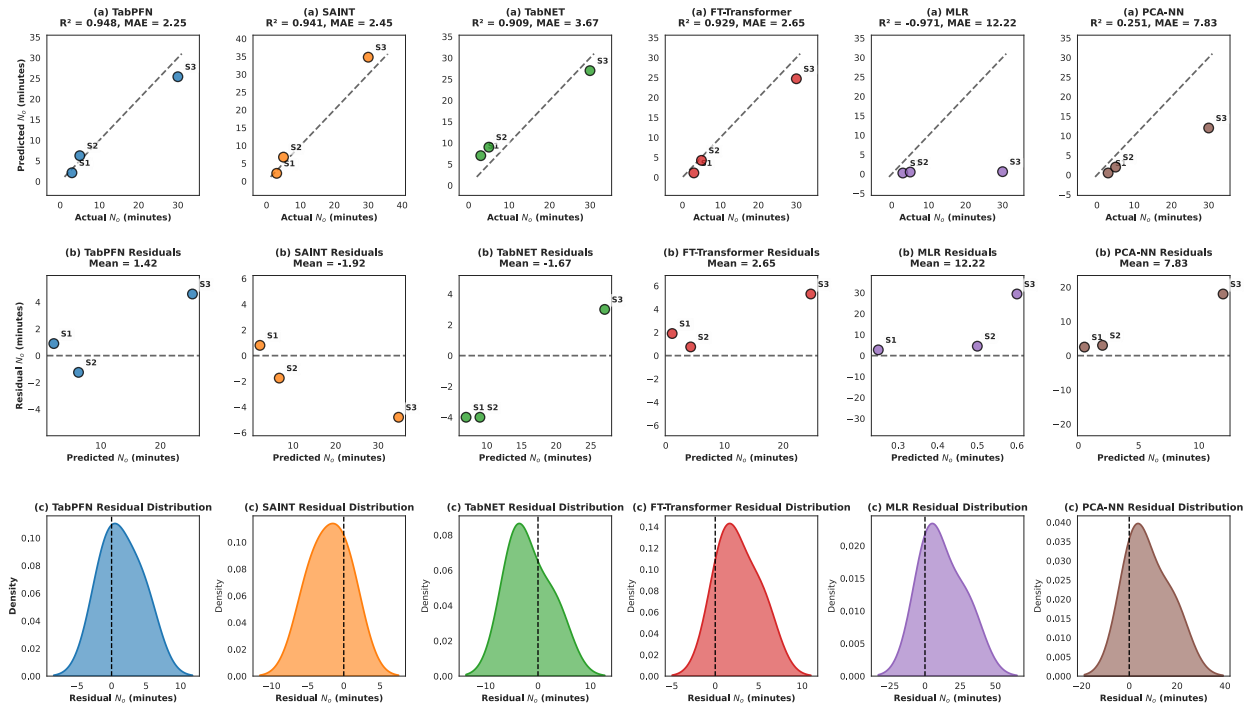


Figure 5.16. Predicted unseen data results for a new material (Austenitic Stainless Steel 304L)

Overall, the validation confirms that transformer-based and meta-learning architectures (FT-Transformer, SAINT, TabNet, TabPFN) adapt more effectively to novel materials, while simpler models such as PCA-NN and MLR face significant limitations when applied beyond their calibration range.

5.4.7 Performance Analysis and Model Ranking

This study systematically compared six approaches—MLR, PCA-NN, FT-Transformer, TabNet, SAINT, and TabPFN—by integrating results from both the primary test dataset (Section 5.4.5) and external validation (Section 5.4.6). The evaluation considered three dimensions: predictive accuracy, interpretability, and generalizability.

On the primary test set, transformer-based models (FT-Transformer, SAINT, TabNet) achieved the highest accuracy ($R^2 = 0.99$), while TabPFN also performed strongly ($R^2 = 0.96$). PCA-NN ($R^2 = 0.91$) and MLR ($R^2 = 0.93$) lagged behind but offered greater interpretability.

On the external 304L dataset, performance patterns differed. TabPFN achieved the strongest generalization ($R^2 = 0.948$), followed by SAINT ($R^2 = 0.941$), FT-Transformer ($R^2 = 0.929$) and TabNet ($R^2 = 0.909$). while PCA-NN ($R^2 = 0.291$) and MLR ($R^2 = -0.971$) performed poorly. Taken together, advanced tabular architectures—FT-Transformer, SAINT, TabNet, and TabPFN—delivered superior predictive accuracy, with TabPFN highlighting the potential of meta-learning for generalization. In contrast, PCA-NN and MLR, while interpretable, lacked robustness when applied to unseen material.

5.5 Conclusion

This study addresses the challenge of predicting the incubation period (N_0) in Water Droplet Erosion (WDE) of metallic materials by combining physics-informed regression with advanced machine learning. A key contribution is the application of dimensionless analysis via the Buckingham π theorem, which enabled the formulation of a physically interpretable power-law model linking elastic modulus, fracture toughness, impact velocity, hardness, yield strength, and Poisson's ratio to erosion resistance.

The MLR model based on these dimensionless groups achieved good accuracy and offered strong interpretability, but it generalized poorly to unseen data. By contrast, transformer-based models (FT-Transformer, SAINT, TabNet) delivered exceptional test-set accuracy. TabPFN demonstrated the strongest external validation performance, underscoring the ability of meta-learning to handle domain shifts in small tabular datasets. Although PCA-NN and MLR provided interpretability, they lacked robustness when applied to new materials.

Overall, the findings suggest that advanced tabular deep learning and meta-learning approaches provide the most reliable predictions of WDE incubation, while physics-informed models remain valuable for mechanistic understanding. Future research should expand experimental datasets and explore hybrid models that integrate physical interpretability with the predictive power of deep learning to improve generalization across diverse material systems.

Acknowledgments: Hashemite University is acknowledged for sponsoring the first author's PhD studies. MT and MM gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC). Finally, the paper benefited from the use of large language models (LLMs) for assistance with language editing and polishing. All scientific content, data collection, analysis, and conclusions are the sole work of the authors.

Author contributions: All authors contributed to the study conception and design. **Khaled AlHammad:** Data curation, Formal analysis, Investigation, Methodology, Writing –original draft. **Mamoun Medraj** and **Moussa Tembely:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. All authors have read and approved the final manuscript. **Julio. C Franco-Correa:** Performed water droplet erosion experiments on austenitic stainless steel 304L; these experimental results were used as an external validation dataset for ML models.

Funding: Not applicable.

Data Availability: The datasets analyzed during the current study are publicly available in the Github repository at the following link: <https://github.com/khaledalhammad/WDE-DATA>

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests

Chapter 6: Synthesis and Discussion

6.1 WDE Prediction Insights

This thesis presents a framework for enhancing the prediction and understanding of WDE in metallic materials through the application of machine learning (ML). Drawing upon the insights gained from three distinct yet interconnected research studies, this work offers a comprehensive perspective on WDE behavior—encompassing both the incubation period and the maximum erosion rate. The integration of these investigations highlights the potential of ML in addressing the complex, multi-parameter nature of WDE that was not possible to address through traditional modeling approaches.

A fundamental challenge in WDE research lies in the nature of experimental data, which often exhibits non-normality or significant skewness, particularly in target variables such as the incubation period and erosion rate. These statistical characteristics can hinder the performance of many machine learning algorithms, which perform better with normally or uniformly distributed data for optimal learning. Addressing this issue was a critical first step toward developing robust and reliable predictive models.

The prediction of the incubation period (N_0) serves as the focal point of the first paper, which has been published. This work demonstrated that statistical transformations, including logarithmic and power-based transformation methods such as Box-Cox and Yeo-Johnson, effectively improve the distribution of the target variable. These preprocessing steps significantly improved data linearity and variance homogeneity, thereby enhancing model performance. Notably, applying the Box-Cox transformation substantially improved the Linear Regression (LR) model, yielding an R^2 value exceeding 90%. This result underscores the value of rigorous data conditioning in maximizing ML model efficacy when working with raw experimental data characterized by non-linear relationships and diverse scales. In addition, using SHAP (SHapley Additive exPlanations) values offered critical interpretability, identifying impact velocity and surface hardness as the most influential factors governing the incubation period. These findings not only align with physical intuition but also provide insights for material design and selection.

The second paper expanded the investigation to the prediction of the maximum erosion rate (R_e), a key indicator of long-term component degradation. Similar to the incubation period, the maximum erosion rate exhibited skewed distributions; however, in this case, standard data transformation techniques yielded limited improvements in model accuracy. This limitation was attributed to data scarcity and the inherently complex nature of erosion phenomena. In response, this study explored the use of data augmentation techniques to address these challenges. The implementation of synthetic data generation strategies, such as SMOTER (Synthetic Minority Over-sampling TEchnique for Regression) and Multivariate Interpolation, substantially enhanced the performance and generalization of predictive models. For instance, the k-nearest neighbors (KNN) model trained on SMOTER-augmented data achieved an R^2 exceeding 98% during cross-validation and retained 97% accuracy on previously unseen validation data. These results demonstrate the critical role of data augmentation in overcoming the constraints of limited experimental datasets, which are a common obstacle in materials science research due to the high cost and effort associated with physical testing.

The third paper evaluated a wide range of ML models—from conventional Multiple Linear Regression (MLR) employing dimensionless analysis to advanced architectures designed for tabular data, including FT-Transformer, TabNet, SAINT, and Tabular Prior-data Fitted Networks (TabPFN). A key outcome of this study was the clear distinction between models that achieve high accuracy on test sets and those that generalize effectively to entirely new, unseen materials. While transformer-based architectures demonstrated near-perfect accuracy on holdout test data (e.g., FT-Transformer and SAINT with $R^2 \approx 0.99$), only models with adaptive learning mechanisms—such as TabPFN—exhibited superior generalization to entirely unseen datasets. In contrast, simpler models informed by physical principles, such as Multiple Linear Regression (MLR) with dimensionless variables, were more interpretable but struggled to maintain predictive accuracy on unseen data. These findings underscore a vital insight: in the context of materials degradation application, where data is often limited and heterogeneous, models that balance interpretability with robust generalization are essential. The use of dimensionless analysis in this study also produced a physically interpretable power-law relationship, reinforcing the value of integrating domain knowledge into machine learning workflows.

Collectively, these studies demonstrate the versatility and strength of ML approaches in tackling the ongoing challenges of WDE prediction. The success of this framework derives from a combined strategy that includes rigorous data preprocessing, data augmentation, and the judicious selection of ML models that balance predictive accuracy with physical interpretability and generalization. This integrated approach not only advances the understanding of WDE mechanisms but also equips engineers with reliable tools for materials selection, performance assessment, and design optimization.

6.2 Balancing Accuracy, Interpretability, and Generalization

The journey through the application of machine learning to WDE prediction, as detailed in this thesis, consistently highlights a fundamental challenge in materials engineering: the need to balance predictive accuracy with model interpretability and generalization capabilities. While highly accurate models are desirable, their utility in real-world engineering applications is significantly enhanced when their predictions can be understood and trusted, and when they perform reliably on new, unseen data.

As demonstrated in all three papers, accuracy is achievable through various ML techniques. The first paper showed that with appropriate data transformations, even simpler models such as Linear Regression can achieve high R^2 values for incubation period prediction. The second paper further pushed the boundaries of accuracy for maximum erosion rate prediction by effectively employing data augmentation, leading to R^2 values exceeding 95% for specific models. This pursuit of high accuracy is essential for developing reliable predictive tools that can inform critical engineering decisions, such as maintenance scheduling and material selection.

However, high predictive accuracy alone is often insufficient. Interpretability is essential for engineers to comprehend the rationale underlying model predictions. In the context of WDE, elucidating the governing physical mechanisms and parameter influences is critical for guiding material design and process optimization. This aspect was addressed in the first paper through the use of SHAP values to identify impact velocity and surface hardness as principal factors affecting the incubation period, thereby yielding physically interpretable insights. A subsequent research effort, currently submitted for publication to the *Journal of Computational Materials Science*. Further emphasized interpretability by applying dimensionless analysis to derive a power-law

relationship that is both predictive and physically grounded. These interpretable models provide a deeper understanding of the nonlinear relationships governing WDE, moving beyond “black-box” predictions. Interpretable models enable engineers to optimize material selection, predict component lifetimes, and design erosion-resistant surfaces.

Perhaps the most critical aspect for practical application, especially in materials science, where new materials and operating conditions are constantly emerging, is generalization. A model that performs exceptionally well on its training data but fails on unseen data is of limited practical value. The third paper provided an illustration of this: while complex transformer-based models achieved superior performance on the held-out test set, lighter meta-learned approaches such as TabPFN exhibited stronger generalization to entirely new unseen materials. It is worth noting that PCA-based models (e.g., PCA-NN) can aid interpretability and noise suppression; however, they did not match the external generalization achieved by transformer and meta-learning methods. The capacity to generalize indicates that a model has captured underlying physical regularities rather than simply fitting training examples. This is especially important for WDE, where exhaustive experiments across all materials and conditions are infeasible. Consequently, models that reliably predict behavior for novel materials or new regimes are invaluable. In essence, the optimal ML approach for WDE—indeed for many materials-science problems—must strike a balance among three pillars: high predictive accuracy, sufficient interpretability, and generalization for practical applications. This thesis advocates a modeling strategy that treats these aspects as interconnected, moving toward ML systems that predict material behavior and deepen our understanding of WDE.

6.3 Practical Implications for Engineering Applications

The advancements in WDE prediction through machine learning, as detailed in this thesis, carry significant practical implications for various engineering applications. By providing more accurate, interpretable, and generalizable predictive tools, this research directly contributes to improved material selection, optimized maintenance strategies, and enhanced component design, ultimately leading to increased operational efficiency and reduced economic losses.

6.3.1 Improved Material Selection

The ability to accurately predict the incubation period and maximum erosion rate of different metallic materials under various impact conditions empowers engineers to make more informed decisions during the design phase. For instance, the insights from the first paper regarding the influence of impact velocity and surface hardness on the incubation period, coupled with the predictive capabilities for both incubation period and maximum erosion rate, allow for a data-driven approach to selecting materials with superior WDE resistance for specific applications. This can lead to using more durable materials, extending the lifespan of components in erosion-prone environments such as steam turbine blades, aircraft components, and wind turbines.

6.3.2 Optimized Maintenance Strategies

Accurate prediction of WDE progression enables proactive and predictive maintenance rather than reactive repairs. By predicting the incubation period, operators can schedule interventions before significant material loss occurs, preventing catastrophic failures and minimizing downtime. The models developed in this thesis can serve as valuable tools for predicting when a component is likely to enter the acceleration or steady-state erosion stages, allowing for timely inspection, repair, or replacement. This shift from time-based to condition-based maintenance can lead to substantial cost savings and improved operational safety.

6.3.3 Enhanced Component Design

The interpretability offered by some ML models, particularly those incorporating dimensionless analysis (third paper) and feature importance (first paper), provides designers with a deeper understanding of the critical parameters influencing WDE. This knowledge can be leveraged to optimize component geometries, surface treatments, or even introduce protective coatings more effectively. For example, understanding the non-linear dependence of the incubation period on material properties and impact conditions can guide the development of new alloys or surface modification techniques specifically tailored for WDE resistance.

6.3.4 Mitigation of Data Scarcity:

As demonstrated in the second paper, the successful application of data augmentation techniques represents a significant advancement for industries where data collection is inherently challenging,

time-consuming, and costly. This research demonstrates that robust and accurate WDE prediction models can be developed even with limited initial datasets. This reduces the reliance on extensive physical testing, accelerating the research and development cycle for new materials and designs. It also opens avenues for leveraging existing, smaller datasets more effectively.

By enabling better prediction and understanding of WDE, this research directly contributes to reducing the risk of component failure, improving safety, and extending the operational life of industrial assets. The potential economic benefits are substantial, including reduced maintenance costs, minimized unplanned downtime, improved energy efficiency (e.g., for wind turbines with less eroded blades), and prolonged asset value. The ability to predict and mitigate WDE effectively translates into more reliable and cost-effective operations across WDE-affected industries.

In conclusion, the machine learning–driven framework for WDE prediction presented in this thesis offers tangible and practical benefits for engineering applications. Enhancing the understanding and predictive capabilities of WDE enables engineers to design, operate, and maintain systems with greater confidence, reliability, and efficiency.

Chapter 7: Conclusion

7.1 Summary of Research

Water droplet erosion (WDE) remains a significant challenge across various industrial applications, underscoring the need for accurate and reliable predictive models to guide material selection, design optimization, and maintenance strategies. This thesis comprehensively demonstrates the transformative potential of advanced machine learning (ML) techniques in addressing the complexities and data limitations associated with WDE prediction. By synthesizing the findings from three interconnected research studies, this work establishes a robust framework for understanding and predicting both the incubation period and the maximum erosion rate of metallic materials.

The first paper tackled the prediction of the incubation period (N_0), a critical early stage of WDE. It revealed that experimental data for N_0 often exhibited significant skewness, hindering the performance of many ML algorithms. To address this, various data transformation techniques, including logarithmic, Box-Cox, and Yeo-Johnson transformations, were successfully applied to transform skewed data to a more normal or uniform distribution, leading to substantial improvements in model accuracy. This highlighted the essential role of data preprocessing in enhancing ML model robustness and predictive power.

Building on this foundation, the second paper focused on predicting the maximum erosion rate (R_e). Similar skewness was observed, but data transformation yielded only marginal improvements, indicating that data scarcity posed a greater challenge. Consequently, the study explored and validated several data augmentation techniques, including SMOTER (Synthetic Minority Over-sampling Technique for Regression) and multivariate interpolation. These techniques improved dataset quality and size, resulting in substantial gains in ML model performance and generalization, particularly with SMOTER.

The third paper revisited incubation period prediction with a new focus on generalization and interpretability for small datasets. In addition to the models previously applied, this study introduced dimensionless analysis using MLR with Pi theorem, Principal Component Analysis

combined with neural networks (PCA-NN), and advanced tabular deep learning architectures, including FT-Transformer, TabNet, SAINT, and TabPFN. The results highlighted the importance of physically interpretable models and architectures with strong inductive biases. Although Multiple Linear Regression (MLR) performed reasonably well on seen data, it failed to generalize effectively to entirely unseen experimental data. In contrast, advanced tabular models—particularly TabPFN, SAINT, and FT-Transformer—demonstrated strong predictive performance on unseen data, making them more suitable for real-world applications where robustness and generalization are critical.

Collectively, these papers form an integrated framework for WDE prediction, demonstrating that combining physics-informed reasoning, appropriate data preprocessing, transformation approaches, augmentation techniques, and modern ML architectures enables significant advances in modeling erosion behavior in metallic materials.

7.2 Contributions to Knowledge

This thesis advances the scientific understanding and predictive modeling of WDE through the following key contributions

7.2.1 Effective Handling of Skewed Data

Applying statistical transformation methods (logarithmic, Box-Cox, Yeo-Johnson) to skewed experimental WDE data, particularly for the incubation period (N_o), significantly improves ML model accuracy and robustness. This provides a valuable methodology for researchers handling non-normally or non-uniformly distributed data in materials degradation.

7.2.2 Use of Data Augmentation to Address Scarcity

The use of data augmentation methods, including SMOTER and multivariate interpolation, to overcome limitations of sparse experimental datasets for predicting maximum erosion rate (R_e) has been accomplished in this thesis. These techniques yielded substantial generalization and predictive performance improvements, proving their value in experimental sciences where data acquisition is costly.

7.2.3 Integration of Interpretability and Generalization

The importance of model interpretability alongside predictive accuracy has been demonstrated in this work. Using the Pi theorem, PCA-NN model, and simple yet generalizable models such as TabPFN demonstrated that physics-informed and bias-encoded models often outperform complex black-box models in extrapolating to new materials.

7.2.4 Development of a Unified ML Framework for WDE

A cohesive framework for predicting both N_o and R_e has been presented here. This framework not only advances predictive accuracy but also facilitates a deeper understanding of erosion behavior and supports more informed decision-making in engineering applications.

7.3 Suggestions for Future Work

Building on the insights and outcomes of this thesis, several promising directions for future research are proposed:

7.3.1 Hybrid Physics-ML Modeling

Develop models that explicitly integrate physical laws within ML architectures, such as Physics-Informed Neural Networks (PINNs) or custom loss functions based on mechanistic constraints. This could enhance model generalization across diverse material systems.

7.3.2 Expansion of Experimental Datasets

Prioritize the collection of broader and more diverse experimental WDE datasets across varying materials, droplet conditions, and environmental factors. Collaborative, open-access WDE databases could accelerate progress and enable more generalized models.

7.3.3 Deployment and Real-World Validation

Evaluate the developed predictive models' long-term robustness and practical applicability in operational environments. For the incubation period model, this includes deployment in real-time monitoring of process variables to anticipate critical timelines. For the maximum erosion rate model, implementation in engineering systems subject to WDE can support predictive maintenance and material selection decisions.

7.3.4 Advanced Interpretability Techniques

Explore explainable AI (XAI) methods beyond SHAP values to uncover the internal reasoning of complex deep learning models such as transformers. Improving interpretability will enhance trust and adoption of these models in industry

7.3.5 Transfer Learning Across Rigs and Degradation Mechanisms

Future research should expand experimental datasets and explore hybrid models that integrate physical interpretability with the predictive power of deep learning to improve generalization across diverse material systems. Furthermore, an important direction involves leveraging machine learning to reconcile WDE experimental data generated using different test rigs. This would address the inherent variability and inconsistencies often observed across various experimental setups, leading to more robust and universally applicable datasets for model training. Another critical area is the application of ML to bridge the gap between laboratory results and real-life engineering applications in the field. This involves developing models capable of translating controlled lab conditions to the complex, dynamic environments encountered in actual industrial settings, thereby enhancing the practical utility and reliability of WDE predictions. Finally, examining the adaptability of WDE-trained models to other related degradation phenomena, such as cavitation or particle erosion, could pave the way toward a unified predictive framework for diverse material degradation processes.

7.4 Final Remarks

In conclusion, this thesis offers a machine learning-based framework for predicting WDE, addressing data scarcity, interpretability, and generalization challenges. The work provides methodological advancements and practical tools for engineers and scientists aiming to mitigate WDE in real-world applications. It lays a foundation for future research that blends domain knowledge with data-driven insights to enhance materials design and equipment maintenance.

References

- [1] Field, J. E., Lesser, M. B., & Dear, J. P. (1985). Studies of two-dimensional liquid-wedge impact and their relevance to liquid-drop impact problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 401(1821), 225–249. <https://doi.org/10.1098/rspa.1985.0096>
- [2] Heymann, F. J. (1992). Liquid impingement erosion. In *ASM Handbook: Vol. 18. Friction, lubrication, and wear technology* (pp. 221–232). ASM International.
- [3] Ibrahim, M. E., & Medraj, M. (2019). Water droplet erosion of wind turbine blades: Mechanics, testing, modeling and future perspectives. *Materials*, 13(1), 157. <https://doi.org/10.3390/ma13010157>
- [4] Rempel, L. (2012). Rotor blade leading edge erosion-real-life experiences. *Wind Systems Magazine*, 11, 22–24.
- [5] Eisenberg, D., Laustsen, S., & Stege, J. (2018). Wind turbine blade coating leading edge rain erosion model: Development and validation. *Wind Energy*, 21(10), 942–951. <https://doi.org/10.1002/we.2200>
- [6] Gohardani, O. (2011). Impact of erosion testing aspects on current and future flight conditions. *Progress in Aerospace Sciences*, 47(4), 280–303. <https://doi.org/10.1016/j.paerosci.2011.04.001>
- [7] Yasugahira, N., Namura, K., Kaneko, R., & Satoh, T. (1990). Erosion resistance of titanium alloys for steam turbine blades as measured by water droplet impingement. In R. I. Jaffee (Ed.), *Titanium steam turbine blading* (pp. 383–402). Pergamon. <https://doi.org/10.1016/B978-0-08-037301-0.50025-7>
- [8] Özçakmak, Ö. S., Bretos, D., Méndez, B., & Bak, C. (2024). Determination of annual energy production loss due to erosion on wind turbine blades. *Journal of Physics: Conference Series*, 2767(2), 022066. <https://doi.org/10.1088/1742-6596/2767/2/022066>
- [9] Malik, T. H., & Bak, C. (2025). Challenges in detecting wind turbine power loss: The effects of blade erosion, turbulence, and time averaging. *Wind Energy Science*, 10(1), 227–243. <https://doi.org/10.5194/wes-10-227-2025>
- [10] Honegger, E. (1927). Tests on erosion caused by jets. *Brown Boveri Review*, 14(4), 95–104.
- [11] Cook, S. S. (1928). Erosion by water-hammer. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 119(783), 481–488. <https://doi.org/10.1098/rspa.1928.0107>
- [12] Springer, G. S., & Baxil, C. B. (1974). A model for rain erosion of homogeneous materials. In A. Thiruvengadam (Ed.), *Erosion, wear, and interfaces with corrosion* (ASTM STP 567, pp. 106–124). ASTM International. <https://doi.org/10.1520/STP32223S>

- [13] Springer, G. S. (1976). *Erosion by liquid impact*. Washington, DC: Scripta Publishing Co.; distributed by Halsted Press Division, Wiley.
- [14] Heymann, F. J. (1979). Conclusions from the ASTM interlaboratory test program with liquid impact erosion facilities. In *International Conference on Erosion by Liquid and Solid Impact* (pp. 20_1–20_10). <https://ui.adsabs.harvard.edu/abs/1979elsi.conf...20H>
- [15] Liu, Y., Chu, X., Yang, G., & Weigand, B. (2024). Simulation and analytical modeling of high-speed droplet impact onto a surface. *Physics of Fluids*, 36(1), 012137. <https://doi.org/10.1063/5.0186883>
- [16] Adler, W. F. (1973). *Analytical modeling of liquid and solid particle erosion* (Report No. AFML-TR-73-174). Air Force Materials Laboratory. <https://doi.org/10.21236/AD0768652>
- [17] Lesser, M. B. (1981). Analytic solution of liquid-drop impact problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 377(1768), 289–308. <https://doi.org/10.1098/rspa.1981.0125>
- [18] Li, R., Ninokata, H., & Mori, M. (2011). A numerical study of impact force caused by liquid droplet impingement onto a rigid wall. *Progress in Nuclear Energy*, 53(7), 881–885. <https://doi.org/10.1016/j.pnucene.2011.03.002>
- [19] Dhami, M. K., Pasricha, I., Kaur, R., & Sidhu, A. S. (2022). A review of machine learning applications in materials science. In *Advances in Manufacturing Technology* (pp. 165–174). Routledge. <https://doi.org/10.1201/9781003203681-21>
- [20] Jasthi, B. K., Gadhamshetty, V., Sereda, G. A., & Lipatov, A. (2023). Machine learning for materials science. In *Machine learning in 2D materials science* (pp. 201–216). CRC Press. <https://doi.org/10.1201/9781003132981-11>
- [21] Gao, G., Zhang, Z., Cai, C., Zhang, J., & Nie, B. (2019). Cavitation damage prediction of stainless steels using an artificial neural network approach. *Metals*, 9(5), 506. <https://doi.org/10.3390/met9050506>
- [22] Vozhakov, I. S., Misyura, S. Y., Morozov, V. S., Piskunov, M. V., Piskunova, A. E., Orlova, E. G., & Mullyadzhyanov, R. I. (2025). High-speed droplet impact on solid surfaces: review on the prospects of modeling fluid flow. *Experiments in Fluids: Experimental Methods and Their Applications to Fluid Flow*, 66(3). <https://doi.org/10.1007/s00348-025-03996-0>
- [23] ASTM International. (2017). *Standard test method for liquid impingement erosion using rotating apparatus* (ASTM G73-10). <https://doi.org/10.1520/G0073-10R17>
- [24] Anderson, K., Azimy, N., Lichty, E., Karimi, S. (2024). Experimental Investigation of Water Droplet Erosion on Metallic and Non-Metallic Materials. In *Proceedings of the ASME 2024 Fluids Engineering Division Summer Meeting*

collocated with the ASME 2024 Heat Transfer Summer Conference and the ASME 2024 18th International Conference on Energy Sustainability (FEDSM2024-131543). American Society of Mechanical Engineers. <https://doi.org/10.1115/FEDSM2024-131543>

[25] Kennedy, C. F., & Field, J. E. (2000). Damage threshold velocities for liquid impact. *Journal of Materials Science*, 35(21), 5331–5339. <https://doi.org/10.1023/A:1004842828161>

[26] Ibrahim, M. E., & Medraj, M. (2022). Prediction and experimental evaluation of the threshold velocity in water droplet erosion. *Materials & Design*, 213, 110312. <https://doi.org/10.1016/j.matdes.2021.110312>

[27] Thiruvengadam, A., Rudy, S., & Gunasekaran, M. (1970). Experimental and analytical investigations on liquid impact erosion. In *Characterization and determination of erosion resistance* (ASTM STP 474, pp. 249–280). ASTM International

[28] Mahdipoor, M., Kirols, H., Kevorkov, D., Jedrzejowski, P., & Medraj, M. (2015). Influence of impact speed on water droplet erosion of TiAl compared with Ti6Al4V. *Scientific Reports*, 5, Article 14182. <https://doi.org/10.1038/srep14182>

[29] Heymann, F. J. (1968). On the shock wave velocity and impact pressure in high-speed liquid-solid impact. *ASME Journal of Basic Engineering*, 90(3), 400–402. <https://doi.org/10.1115/1.3605114>

[30] Gujba, A. K., Hackel, L., Kevorkov, D., & Medraj, M. (2016). Water droplet erosion behaviour of Ti–6Al–4V and mechanisms of material damage at the early and advanced stages. *Wear*, 358–359, 109–122. <https://doi.org/10.1016/j.wear.2016.04.008>

[31] Heymann, F. J. (1967). A survey of clues to the relation between erosion rate and impact parameters. In *Second rain erosion conference* (Vol. 2, pp. 683–760).

[32] Lee, B., Riu, K., Shin, S., & Kwon, S. (2003). Development of a water droplet erosion model for large steam turbine blades. *KSME International Journal*, 17(1), 114–121. <https://doi.org/10.1007/BF02984292>

[33] Ahmad, M., Schatz, M., & Casey, M. V. (2013). Experimental investigation of droplet size influence on low pressure steam turbine blade erosion. *Wear*, 303(1–2), 83–86. <https://doi.org/10.1016/j.wear.2013.03.013>

[34] Kirols, H. S., Mahdipoor, M. S., Kevorkov, D., Uihlein, A., & Medraj, M. (2016). Energy-based approach for understanding water droplet erosion. *Materials & Design*, 104, 76–86. <https://doi.org/10.1016/j.matdes.2016.04.089>

- [35] Heymann, F. J. (1970). Toward quantitative prediction of liquid impact erosion. In *Characterization and determination of erosion resistance* (ASTM STP 474, pp. 212–248). ASTM International. <https://doi.org/10.1520/STP26871S>
- [36] Hancox, N. L., & Brunton, J. H. (1966). The erosion of solids by the repeated impact of liquid drops. *Philosophical Transactions of the Royal Society A*, 260(1110), 121–139. <https://doi.org/10.1098/RSTA.1966.0036>
- [37] Hattori, S., & Lin, G. (2012). Effect of droplet diameter on liquid impingement erosion. *AIP Conference Proceedings*, 1428(1), 191–198. <https://doi.org/10.1063/1.3694706>
- [38] Ahmad, M., Casey, M., & Sürken, N. (2009). Experimental assessment of droplet impact erosion resistance of steam turbine blade materials. *Wear*, 267(9–10), 1605–1618. <https://doi.org/10.1016/j.wear.2009.06.012>
- [39] Ahmad, M., Schatz, M., & Casey, M. V. (2018). An empirical approach to predict droplet impact erosion in low-pressure stages of steam turbines. *Wear*, 402–403, 57–63. <https://doi.org/10.1016/j.wear.2018.02.004>
- [40] Bassim, M. N., & Hay, D. R. (1974). The impact of fracture mechanics on engineering practice. *Canadian Journal of Civil Engineering*, 1(2), 158–166. <https://doi.org/10.1139/L74-015>
- [41] Anderson, K., Karimi, S., & Shirazi, S. (2021). Erosion testing and modeling of several non-metallic materials. *Wear*, 477, 203811. <https://doi.org/10.1016/j.wear.2021.203811>
- [42] Lin, B.-S. (1992). Effect of the Poisson ratio on the perfectly plastic stress field at a stationary plane-strain crack tip. *Applied Mathematics and Mechanics*, 13(3), 289–295. <https://doi.org/10.1007/BF02457375>
- [43] Field, J. E. (1966). Stress waves, deformation and fracture caused by liquid impact. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 260(1110), 86–93. <https://doi.org/10.1098/rsta.1966.0032>
- [44] Itoh, H., & Okabe, N. (1993). Evaluation of erosion by liquid droplet impingement for metallic materials. *Transactions of the Japan Society of Mechanical Engineers Series A*, 59(567), 2736–2741. <https://doi.org/10.1299/kikaia.59.2736>
- [45] Slot, H., Matthews, D., Schipper, D., & Van Der Heide, E. (2021). Fatigue-based model for the droplet impingement erosion incubation period of metallic surfaces. *Fatigue & Fracture of Engineering Materials & Structures*, 44(1), 199–211. <https://doi.org/10.1111/ffe.13352>

- [46] Mora, A., Xu, R., & Schmauder, S. (2023). Hybrid smoothed-particle hydrodynamics/finite element method simulation of water droplet erosion on ductile metallic targets. *Metals*, 13(12), 1937. <https://doi.org/10.3390/met13121937>
- [47] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547–555. <https://doi.org/10.1038/s41586-018-0337-2>
- [48] Liu, Y., Zhao, T., Ju, W., & Shi, S. (2017). Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3), 159–177. <https://doi.org/10.1016/j.jmat.2017.08.002>
- [49] Guo, Y., Sun, M., Zhang, W., & Wang, L. (2023). Machine learning in enhancing corrosion resistance of magnesium alloys: A comprehensive review. *Metals*, 13(10), 1790. <https://doi.org/10.3390/met13101790>
- [50] Chen, D., Li, Y., Liu, K., & Li, Y. (2023). A physics-informed neural network approach to fatigue life prediction using small quantity of samples. *International Journal of Fatigue*, 166, 107270. <https://doi.org/10.1016/j.ijfatigue.2022.107270>
- [51] Karimi, S., Xu, B., Asgharpour, A., Shirazi, S., & Sen, S. (2020). Predicting solid particle erosion and uncertainty in elbows by artificial intelligence methods. *Proceedings of the ASME 2020 Fluids Engineering Division Summer Meeting (FEDSM2020)*. American Society of Mechanical Engineers. <https://doi.org/10.1115/FEDSM2020-20458>
- [52] Hariprasad, T., Shivalingappa, D., Nagaraj, A., & Manivasagam, G. (2015). The use of artificial neural network for the prediction of wear loss of aluminium–magnesium alloys. *International Journal of Computer Aided Engineering and Technology*, 7(1), 72–80. <https://doi.org/10.1504/IJCAET.2015.066174>
- [53] Pardoe, I. (2012). Multiple linear regression. In *Applied regression modeling* (2nd ed., pp. 83–135). John Wiley & Sons. <https://doi.org/10.1002/9781118274415.ch3>
- [54] Clark, L. A., & Pregibon, D. (2017). Tree-based models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 377–419). Routledge. <https://doi.org/10.1201/9780203738535-9>
- [55] Keprate, A., & Ratnayake, R. M. C. (2017). Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1331–1336). IEEE. <https://doi.org/10.1109/IEEM.2017.8290109>
- [56] Lovie, A. D. (1995). Who discovered Spearman’s rank correlation? *British Journal of Mathematical and Statistical Psychology*, 48(2), 255–269. <https://doi.org/10.1111/j.2044-8317.1995.tb01063.x>

- [57] Mach, P., Thuring, J., & Samal, D. (2006). Transformation of data for statistical processing. In *2006 29th International Spring Seminar on Electronics Technology* (pp. 278–282). IEEE. <https://doi.org/10.1109/ISSE.2006.365112>
- [58] Vičar, B. (2023). A comprehensive implementation of the log, Box-Cox and log-sinh transformations for skewed and censored precipitation data. *Journal of Hydrology*, 620, 129347. <https://doi.org/10.1016/j.jhydrol.2023.129347>
- [59] Oh, C., Han, S., & Jeong, J. (2020). Time-series data augmentation based on interpolation. *Procedia Computer Science*, 175, 64–71. <https://doi.org/10.1016/j.procs.2020.07.012>
- [60] Zhu, Y., Su, H., Xu, P., Xu, Y., Wang, Y., Dong, C.-H., Lu, J., Le, Z., Yang, X., Xuan, Q., Zou, C.-L., & Ren, H. (2023). Data augmentation using continuous conditional generative adversarial networks for regression and its application to improved spectral sensing. *Optics Express*, 31(23), 37722–37739. <https://doi.org/10.1364/OE.502709>
- [61] Dong, H., Wang, J., Wu, X., Zhou, M., & Lu, J. (2023). Gaussian Noise Data Augmentation-Based Delay Prediction for High-Speed Railways. *IEEE Intelligent Transportation Systems Magazine*, 15(6). <https://doi.org/10.1109/MTS.2023.3274787>
- [62] Zhang, Z., Wang, H., Geng, J., Deng, X., & Jiang, W. (2023). A new data augmentation method based on mixup and Dempster-Shafer theory. *IEEE Transactions on Multimedia*, 26, 4998–5013. <https://doi.org/10.1109/TMM.2023.3330106>
- [63] Marnissi, Y., Abboud, D., Chouzenoux, E., Pesquet, J.-C., El-Badaoui, M., & Benazza-Benyahia, A. (2019). A data augmentation approach for sampling Gaussian models in high dimension. In *2019 27th European Signal Processing Conference (EUSIPCO)* (pp. 1–5). IEEE. <https://doi.org/10.23919/EUSIPCO.2019.8902496>
- [64] Jalili, M., & Talebi, S. (2012). Analysis of electrostatically actuated microbeams using dimensionless groups. *International Journal of Physical Sciences*, 7(6), 906–912. <https://doi.org/10.5897/IJPS11.905>
- [65] Lewis, B. J., Onder, E. N., & Prudil, A. A. (2022). Nondimensionalization. In B. J. Lewis, E. N. Onder, & A. A. Prudil (Eds.), *Advanced mathematics for engineering students* (pp. 311–327). Butterworth-Heinemann. <https://doi.org/10.1016/B978-0-12-823681-9.00019-8>
- [66] Reddy, G. M., & Reddy, V. D. (2014). Theoretical Investigations on Dimensional Analysis of Ball Bearing Parameters by Using Buckingham Pi-Theorem. *Procedia Engineering*, 97, 1305–1311. <https://doi.org/10.1016/j.proeng.2014.12.410>

- [67] Groth, D., Hartmann, S., Klie, S., & Selbig, J. (2013). Principal components analysis. In J. M. Walker (Ed.), *Methods in molecular biology: Vol. 930. Data mining in systems biology* (pp. 527–547). Humana Press. https://doi.org/10.1007/978-1-62703-059-5_22
- [68] Suh, C., Rajagopalan, A., Li, X., & Rajan, K. (2002). The application of Principal Component Analysis to materials science data. *Data Science Journal*, 1(1), 19–26. <https://doi.org/10.2481/dsj.1.19>
- [69] Zhao, Y. (2012). Applications of principal component analysis (PCA) in materials science. In P. Sanguansat (Ed.), *Principal component analysis – engineering applications*. IntechOpen. <https://doi.org/10.5772/37523>
- [70] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932–18943.
- [71] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2022). *SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training*. arXiv. <https://doi.org/10.48550/arXiv.2106.01342>
- [72] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
- [73] Zhuohan, W. (2022). Tabnet with data augmentation approach in stock return prediction task. In *2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 1–5). IEEE. <https://doi.org/10.1109/ICCWAMTIP56608.2022.10016580>
- [74] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). *TabPFN: A transformer that solves small tabular classification problems in a second*. arXiv. <https://doi.org/10.48550/arXiv.2207.01848>
- [75] Hollmann, N., Müller, S., Salinas, D., & Hutter, F. (2025). *The tabular foundation model TabPFN outperforms specialized time series forecasting models based on simple features*. arXiv. <https://doi.org/10.48550/arXiv.2501.02945>
- [76] Bhushan, B. (2013). *Introduction to tribology* (2nd ed.). Wiley. <https://doi.org/10.1002/9781118403259>
- [77] Wood, R. J. K. (2017). Liquid impingement erosion. In G. E. Totten (Ed.), *ASM Handbook: Vol. 18. Friction, lubrication, and wear technology* (pp. 302–312). ASM International. <https://doi.org/10.31399/asm.hb.v18.a0006378>
- [78] Engel, O. G. (1955). Waterdrop collisions with solid surfaces. *Journal of Research of the National Bureau of Standards*, 54(5), 281–298. <https://doi.org/10.6028/jres.054.033>

- [79] Rum, Z., & Straka, F. (1995). A new model for steam turbine blade materials erosion. *Wear*, 186–187, 421–424. [https://doi.org/10.1016/0043-1648\(95\)07164-4](https://doi.org/10.1016/0043-1648(95)07164-4)
- [80] Keegan, M. H., Nash, D. H., & Stack, M. M. (2013). On erosion issues associated with the leading edge of wind turbine blades. *Journal of Physics D: Applied Physics*, 46(38), 383001. <https://doi.org/10.1088/0022-3727/46/38/383001>
- [81] Mednikov, A. F., Ryzhenkov, V. A., Seleznev, L. I., & Lebedeva, A. I. (2012). Studying the variation of parameters characterizing the material surface during the droplet erosion incubation period. *Thermal Engineering*, 59(5), 414–420. <https://doi.org/10.1134/S0040601512050072>
- [82] Durmuş, H. K., Özkaya, E., & Meriç, C. (2006). The use of neural networks for the prediction of wear loss and surface roughness of AA 6351 aluminium alloy. *Materials & Design*, 27(2), 156–159. <https://doi.org/10.1016/j.matdes.2004.09.011>
- [83] Pleune, T. T., & Chopra, O. K. (2000). Using artificial neural networks to predict the fatigue life of carbon and low-alloy steels. *Nuclear Engineering and Design*, 197(1–2), 1–12. [https://doi.org/10.1016/S0029-5493\(99\)00252-6](https://doi.org/10.1016/S0029-5493(99)00252-6)
- [84] Kang, J.-Y., & Song, J.-H. (1998). Neural network applications in determining the fatigue crack opening load. *International Journal of Fatigue*, 20(1), 57–69. [https://doi.org/10.1016/S0142-1123\(97\)00119-9](https://doi.org/10.1016/S0142-1123(97)00119-9)
- [85] Hassan, A. M., Alrashdan, A., Hayajneh, M. T., & Mayyas, A. T. (2009). Prediction of density, porosity, and hardness in aluminum–copper-based composite materials using artificial neural network. *Journal of Materials Processing Technology*, 209(2), 894–899. <https://doi.org/10.1016/j.jmatprotec.2008.02.066>
- [86] Oka, Y. I., & Miyata, H. (2009). Erosion behaviour of ceramic bulk and coating materials caused by water droplet impingement. *Wear*, 267(11), 1804–1810. <https://doi.org/10.1016/j.wear.2009.02.009>
- [87] Wang, X., Chen, S., Xu, J., Li, Y., Gao, G., & Chen, J. (2024). Comparative study on the water droplet erosion behavior between stretched and casted aeronautical PMMA. *International Journal of Impact Engineering*, 183, 104806. <https://doi.org/10.1016/j.ijimpeng.2023.104806>
- [88] Jatavallabhula, J. K., Shabana, S., & Pappula, B. (2025). Development and evaluation of machine learning based predictive models for tribological properties of blended coatings at elevated temperature. *Journal of Bio- and Tribo-Corrosion*, 11(1), Article 7. <https://doi.org/10.1007/s40735-025-00952-7>
- [89] Harouz, R., Zelmatı, D., & Khelil, K. (2024). Machine learning application for wear rate prediction of WC/Co-based cermet with different content of Ni, Cr, TiC, TaC, and NbC. *The International Journal of Advanced Manufacturing Technology*, 135(11–12), 5945–5959. <https://doi.org/10.1007/s00170-024-14862-4>

- [90] Espinoza-Jara, A., Wilk, I., Aguirre, J., & Walczak, M. (2023). An AI-extended prediction of erosion-corrosion degradation of API 5L X65 steel. *Lubricants*, 11(10), 431. <https://doi.org/10.3390/lubricants11100431>
- [91] Thiruvengadam, A. (Ed.). (1974). *Erosion, wear, and interfaces with corrosion* (ASTM STP 567). ASTM International. <https://doi.org/10.1520/STP567-EB>
- [92] MatWeb. (n.d.). *Material property data*. Retrieved December 20, 2024, from <https://www.matweb.com>
- [93] El Garchani, F. E., & Kabiri, M. R. (2023). Intergranular corrosion and mechanical property evolution in AA2024 alloy through heat treatment. *The International Journal of Advanced Manufacturing Technology*, 128(7–8), 3273–3282. <https://doi.org/10.1007/s00170-023-12161-y>
- [94] Tiryakioğlu, M. (2015). On the relationship between Vickers hardness and yield stress in Al–Zn–Mg–Cu alloys. *Materials Science and Engineering: A*, 633, 17–19. <https://doi.org/10.1016/j.msea.2015.02.073>
- [95] Anderson, D. D., & Rosakis, A. J. (2006). Dynamic fracture properties of titanium alloys. *Experimental Mechanics*, 46(3), 399–406. <https://doi.org/10.1007/s11340-006-6421-9>
- [96] Mhetre, H., Pande, S., Singla, B., & Kumar, A. (2025). Data-driven modeling of TiO₂ PVP nanofiber diameter using LSTM and regression for enhanced functional performance. *Discover Applied Sciences*, 7, Article 336. <https://doi.org/10.1007/s42452-025-06823-7>
- [97] Curran-Everett, D. (2018). Explorations in statistics: The log transformation. *Advances in Physiology Education*, 42(2), 343–347. <https://doi.org/10.1152/advan.00018.2018>
- [98] Huang, Z., Zhao, T., Lai, R., Tian, Y., & Yang, F. (2023). A comprehensive implementation of the log, Box-Cox, and log-sinh transformations for skewed and censored precipitation data. *Journal of Hydrology*, 620, 129347. <https://doi.org/10.1016/j.jhydrol.2023.129347>
- [99] Yeo, I.-K. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954–959. <https://doi.org/10.1093/biomet/87.4.954>
- [100] Stoto, M. A., & Emerson, J. D. (1983). Power transformations for data analysis. *Sociological Methodology*, 14, 126–168. <https://doi.org/10.2307/270905>
- [101] Ewald, F. K., Bothmann, L., Wright, M. N., Bischl, B., Casalicchio, G., & König, G. (2024). A guide to feature importance methods for scientific inference. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Lecture notes in computer science: Vol. 2154. Explainable artificial intelligence* (pp. 440–464). Springer. https://doi.org/10.1007/978-3-031-63797-1_22

- [102] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. <https://doi.org/10.1017/CBO9781107298019>
- [103] Ma, K. (2024). Integrated hybrid modeling and SHAP (SHapley Additive exPlanations) to predict and explain the adsorption properties of thermoplastic polyurethane (TPU) porous materials. *RSC Advances*, 14(15), 10348–10357. <https://doi.org/10.1039/d4ra00010b>
- [104] Mahmood, Z., & Khan, S. (2009). On the use of K-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression. *The International Journal of Biostatistics*, 5(1), Article 23. <https://doi.org/10.2202/1557-4679.1105>
- [105] Jiang, G., & Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition*, 69, 94–106. <https://doi.org/10.1016/j.patcog.2017.03.025>
- [106] Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(1), 477–508. <https://doi.org/10.1146/annurev-fluid-010719-060214>
- [107] Carranza-Abaid, A., & Jakobsen, J. P. (2022). Neural network programming: integrating first principles into machine learning models. *Computational Chemical Engineering*, 163, 107858. <https://doi.org/10.1016/j.compchemeng.2022.107858>
- [108] Khanal, A., & Dik, M. (2022). Comparative analysis of first and second order methods for optimization in neural networks. *Proceedings of International Mathematical Sciences*, 4(2), 77–87. <https://doi.org/10.47086/pims.1170457>
- [109] Naser, M. Z., & Alavi, A. H. (2023). Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. *Architectural Structure and Construction*, 3(4), 499–517. <https://doi.org/10.1007/s44150-021-00015-8>
- [110] Singh, S., Singh, D. S., & Kumar, S. (2014). Modified mean square error algorithm with reduced cost of training and simulation time for character recognition in backpropagation neural network. In S. C. Satapathy, S. K. Udgata, & B. N. Biswal (Eds.), *Lecture notes in computer science: Vol. 247. Frontiers of intelligent computing: Theory and applications* (pp. 13-20). Springer. https://doi.org/10.1007/978-3-319-02931-3_17
- [111] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295–316. <https://doi.org/10.1016/j.neucom.2020.07.061>
- [112] Ali, Y., Awwad, E., Al-Razgan, M., & Maarouf, A. (2023). Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes*, 11(2), 349. <https://doi.org/10.3390/pr11020349>

- [113] Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, 16, e01059. <https://doi.org/10.1016/j.cscm.2022.e01059>
- [114] Del Buono, N., Esposito, F., & Selicato, L. (2020). Methods for hyperparameters optimization in learning approaches: An overview. In G. R. G. Notarnicola & A. M. Fanelli (Eds.), *Lecture notes in computer science: Vol. 12565. Machine learning, optimization, and data science* (pp. 100–112). Springer. https://doi.org/10.1007/978-3-030-64583-0_11
- [115] Arnold, C., Biedebach, L., Küpfer, A., & Neunhoeffler, M. (2024). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*, 12(4), 841–848. <https://doi.org/10.1017/psrm.2023.61>
- [116] George, C. G., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications*, 11(9). <https://doi.org/10.14569/IJACSA.2020.0110920>
- [117] Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, 8(4), 79. <https://doi.org/10.3390/informatics8040079>
- [118] Peterson, R. A., & Cavanaugh, J. E. (2020). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 47(13–15), 2312–2327. <https://doi.org/10.1080/02664763.2019.1630372>
- [119] Hamasha, M. M., Ali, H., Hamasha, S., & Ahmed, A. (2022). Ultra-fine transformation of data for normality. *Heliyon*, 8(5), e09370. <https://doi.org/10.1016/j.heliyon.2022.e09370>
- [120] Peng, Y.-L., & Lee, W.-P. (2024). Practical guidelines for resolving the loss divergence caused by the root-mean-squared propagation optimizer. *Applied Soft Computing*, 153, 111335. <https://doi.org/10.1016/j.asoc.2024.111335>
- [121] Elshamy, R., Abu-Elnasr, O., Elhoseny, M., & Elmougy, S. (2023). Improving the efficiency of RMSProp optimizer by utilizing Nesterov in deep learning. *Scientific Reports*, 13(1), Article 8279. <https://doi.org/10.1038/s41598-023-35663-x>
- [122] Ryzhenkov, V. A., Lebedeva, A. I., & Mednikov, A. F. (2011). Erosion wear of the blades of wet-steam turbine stages: Present state of the problem and methods for solving it. *Thermal Engineering*, 58(9), 713–718. <https://doi.org/10.1134/S0040601511090138>

- [123] Richter, A. N., & Khoshgoftaar, T. M. (2019). Learning curve estimation with large imbalanced datasets. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 763–768). IEEE. <https://doi.org/10.1109/ICMLA.2019.00135>
- [124] Liu, Y., Xu, N., Li, M., & Jin, J. (2024). Intelligent prediction model for ship roll damping based on CTGAN-XGBoost. In *2024 13th International Conference on Communications, Circuits and Systems (ICCCAS)* (pp. 154–158). IEEE. <https://doi.org/10.1109/ICCCAS62034.2024.10652856>
- [125] Zhu, Y., Su, H., Xu, P., Xu, Y., Wang, Y., Dong, C. H., Lu, J., Le, Z., Yang, X., Xuan, Q., Zou, C. L., & Ren, H. (2023). Data augmentation using continuous conditional generative adversarial networks for regression and its application to improved spectral sensing. *Optics Express*, 31(23), 37722–37739. <https://doi.org/10.1364/OE.502709>
- [126] Horñas, J., Materna, A., Glinz, J., Yosifov, M., & Senck, S. (2025). Multivariate interpolation and machine learning models for extreme defects-based fatigue life prediction of Ti6Al4V specimens fabricated by SLM. *Engineering Fracture Mechanics*, 314, 110756. <https://doi.org/10.1016/j.engfracmech.2024.110756>
- [127] Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1), 91–99. <https://doi.org/10.1016/j.gltp.2022.04.020>
- [128] AlHammad, K., Medraj, M., & Tembely, M. (2025). Application of machine learning for predicting the incubation period of water droplet erosion in metals. *Discover Applied Sciences*, 7(7). <https://doi.org/10.1007/s42452-025-07268-8>
- [129] Rust, T. M., & Steltz, W. G. (1982). Titanium for steam turbine blades. *JOM*, 34(9), 42–47. <https://doi.org/10.1007/BF03338096>
- [130] Wilcox, R. R. (2016). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, 69(3), 215–224. <https://doi.org/10.1111/bmsp.12069>
- [131] Zhang, S., Lei, H., Zhou, Z., Wang, G., & Qiu, B. (2023). Fatigue life analysis of high-strength bolts based on machine learning method and SHapley Additive exPlanations (SHAP) approach. *Structures*, 51, 275–287. <https://doi.org/10.1016/j.istruc.2023.03.060>
- [132] Trentin, E. (2006). A novel connectionist-oriented feature normalization technique. In R. G. F. D. & M. K. (Eds.), *Lecture notes in computer science: Vol. 4132. Artificial neural networks – ICANN 2006* (pp. 410–416). Springer. https://doi.org/10.1007/11840930_42

- [133] Rodriguez, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 569–575. <https://doi.org/10.1109/TPAMI.2009.187>
- [134] Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109–139). Springer. https://doi.org/10.1007/978-3-030-89010-0_4
- [135] Demir, S., Mincev, K., Kok, K., & Paterakis, N. G. (2021). Data augmentation for time series regression: Applying transformations, autoencoders and adversarial networks to electricity price forecasting. *Applied Energy*, 304, 117695. <https://doi.org/10.1016/j.apenergy.2021.117695>
- [136] Jawa, M., & Meena, S. (2022). Software effort estimation using synthetic minority over-sampling technique for regression (SMOTER). In *2022 3rd International Conference for Emerging Technology (INCET)* (pp. 1–6). IEEE. <https://doi.org/10.1109/INCET54531.2022.9824043>
- [137] Triggiano, F., & Romito, M. (2024). Gaussian processes based data augmentation and expected signature for time series classification. *IEEE Access*, 12, 80884–80895. <https://doi.org/10.1109/ACCESS.2024.3408712>
- [138] Tian, Y., Shen, J., Wang, A., Li, Z., & Huang, X. (2024). Data augmentation and fault diagnosis for imbalanced industrial process data based on residual Wasserstein generative adversarial network with gradient penalty. *Journal of Chemometrics*, 38(12), e3624. <https://doi.org/10.1002/cem.3624>
- [139] Dablain, D. A., & Chawla, N. V. (2024). Data augmentation's effect on machine learning models when learning with imbalanced data. In *Proceedings of the 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–10). IEEE. <https://doi.org/10.1109/DSAA61799.2024.10722815>
- [140] Ying, D., Hua, P., & Hao, M. (2023). Research and application of SMOTE-based method with XGBoost regression prediction. In *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA)* (pp. 1737–1740). IEEE. <https://doi.org/10.1109/ICIPCA59209.2023.10257809>
- [141] Dorner, G. C., Ferrard, J. M., & Lemberg, H. (2000). Interpolation and approximation. In *Visual mathematics, illustrated by the TI-92 and the TI-89* (pp. 159–219). Springer. https://doi.org/10.1007/978-2-8178-0201-5_4
- [142] Jetter, K. (2006). *Topics in multivariate approximation and interpolation* (1st ed.). Elsevier.
- [143] Khosravi, H., Farhadpour, S., Grandhi, M., Shoyeb Raihan, A., Das, S., & Ahmed, I. (2024). Strategic data augmentation with CTGAN for smart manufacturing: Enhancing ML predictions of paper breaks in pulp-and-paper production. *Manufacturing Letters*, 41, 1312–1323. <https://doi.org/10.1016/j.mfglet.2024.09.158>

- [144] Kadhim, Z. S., Abdullah, H. S., & Ghathwan, K. I. (2022). Artificial Neural Network Hyperparameters Optimization: A Survey. *International Journal of Online and Biomedical Engineering (iJOE)*, 18(15), pp. 59–87. <https://doi.org/10.3991/ijoe.v18i15.34399>
- [145] Field, J. E. (1999). ELSI conference: Invited lecture: Liquid impact: Theory, experiment, applications. *Wear*, 233–235, 1–12. [https://doi.org/10.1016/S0043-1648\(99\)00189-1](https://doi.org/10.1016/S0043-1648(99)00189-1)
- [146] Buckingham, E. (1914). On physically similar systems: Illustrations of the use of dimensional equations. *Physical Review*, 4(4), 345–376. <https://doi.org/10.1103/PhysRev.4.345>
- [147] Dym, C. L. (2004). Dimensional analysis. In *Principles of mathematical modeling* (2nd ed., pp. 13–32). Elsevier. <https://doi.org/10.1016/B978-012226551-8/50003-X>
- [148] Zhang, Z., Liu, T., Zhang, D., & Xie, Y. (2021). Water droplet erosion life prediction method for steam turbine blade materials based on image recognition and machine learning. *Journal of Engineering for Gas Turbines and Power*, 143(3), 031009. <https://doi.org/10.1115/1.4049768>
- [149] Zhang, Y., & Xu, X. (2021). Solid particle erosion rate predictions through LSBoost. *Powder Technology*, 388, 517–525. <https://doi.org/10.1016/j.powtec.2021.04.072>
- [150] Rebala, G., Ravi, A., & Churiwala, S. (2019). Principal component analysis. In *An introduction to machine learning* (pp. 141–152). Springer. https://doi.org/10.1007/978-3-030-15729-6_12
- [151] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- [152] Lee, H., Park, Y. M., & Lee, S. (2015). Principal component regression by principal component selection. *Communications for Statistical Applications and Methods*, 22(2), 173–180. <https://doi.org/10.5351/CSAM.2015.22.2.173>
- [153] Sun, Y., Zhou, S., Meng, S., Wang, M., & Mu, H. (2023). Principal component analysis–artificial neural network-based model for predicting the static strength of seasonally frozen soils. *Scientific Reports*, 13(1), Article 16085. <https://doi.org/10.1038/s41598-023-43462-7>
- [154] Fan, C., Zhang, N., Jiang, B., & Liu, W. V. (2024). Using deep neural networks coupled with principal component analysis for ore production forecasting at open-pit mines. *Journal of Rock Mechanics and Geotechnical Engineering*, 16(3), 727–740. <https://doi.org/10.1016/j.jrmge.2023.06.005>

- [155] Zhang, G., Wang, C., Xu, B., & Grosse, R. (2018). *Three mechanisms of weight decay regularization*. arXiv. <https://doi.org/10.48550/arXiv.1810.12281>
- [156] Babu, R. A., Priya V, V., Kumar Mishra, M., Ramesh Raja, I., Kiran Chebrolu, S., & Swarna, B. (2025). Transformer-Based Tabular Foundation Models: Outperforming Traditional Methods with TabPFN. *International Journal of Engineering, Science and Information Technology*, 5(3), 448–455. <https://doi.org/10.52088/ijesty.v5i3.1146>
- [157] Bhagat, M., & Bakariya, B. (2025). A comprehensive review of cross-validation techniques in machine learning. *International Journal on Science and Technology*, 16(1), 1305–1312. <https://doi.org/10.71097/ijstat.v16.i1.1305>
- [158] Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., & Tallón-Ballesteros, A. J. (2023). The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In P. García Bringas et al. (Eds.), *18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023) (Lecture Notes in Networks and Systems, Vol. 750, pp. 353–365)*. https://doi.org/10.1007/978-3-031-42536-3_33
- [159] Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for regression. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), *Progress in artificial intelligence: EPIA 2013 (Lecture Notes in Computer Science, Vol. 8154, pp. 378–389)*. Springer. https://doi.org/10.1007/978-3-642-40669-0_33
- [160] Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *mixup: Beyond empirical risk minimization*. arXiv. <https://doi.org/10.48550/arXiv.1710.09412>
- [161] EL Bilali, A., Taleb, A., Bahlaoui, M. A., & Brouziyne, Y. (2021). An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset. *Journal of Hydrology*, 599. <https://doi.org/10.1016/j.jhydrol.2021.126510>
- [162] Nartsev, D. Y., Gneushev, A. N., & Matveev, I. A. (2022). Adam adaptive optimization method for neural network models regression in image recognition tasks. In *Machine learning for biometrics* (pp. 147–163). Wiley <https://doi.org/10.1002/9781119798798.ch7>
- [163] Suleiman, A., Abdullahi, U. A., & Ahmad, A. (2015). An analysis of residuals in multiple regressions. *International Journal of Advanced Technology in Engineering and Science*, 3(1), 563–570.
- [164] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7499–7519. <https://doi.org/10.1109/TNNLS.2022.3229161>