ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Review Article

Open Access

Machine Learning Insights into the Dynamics of Cusp Catastrophe Instability Region

Pascal Stiefenhofer

Department of Economics, Newcastle University, UK

ABSTRACT

This study investigates a novel application of Random Forest regression for analyzing the unstable set of the cusp catastrophe, a mathematical model describing abrupt, nonlinear transitions in dynamical systems. The cusp catastrophe effectively captures critical phenomena such as bifurcation, hysteresis, and multistability; however, modeling its unstable region remains challenging due to noise, sparsity, and the localized nature of transitions in real world data.

Random Forest's ability to approximate complex, nonlinear relationships was evaluated across varying noise levels and data availabilities. The results demonstrate that the model excels in low noise scenarios, accurately capturing the critical features of the unstable set. However, performance declines with increasing noise and limited data, highlighting the need for noise tolerant strategies. This work provides new insights into leveraging machine learning for robust modeling and prediction of unstable regions in non-linear dynamical systems, offering a foundation for future advancements in catastrophe theory applications.

***Corresponding author**

Pascal Stiefenhofer, Department of Economics, Newcastle University, UK.

Received: December 24, 2024; **Accepted:** December 27, 2024, **Published:** January 07, 2025

Keywords: Cusp Catastrophe, Random Forest, Machine Learning

Introduction

The study of dynamical systems with abrupt, discontinuous transitions plays a vital role in understanding non-linear phenomena across disciplines such as physics, biology, engineering, and economics [1-3]. These transitions often signal critical thresholds or tipping points, where small, continuous changes in control parameters lead to sudden, dramatic shifts in system behavior. Among the theoretical frameworks used to model such phenomena, the cusp catastrophe, introduced by Thom and popularized by Zeeman, stands out as a cornerstone of catastrophe theory [4, 5]. The cusp catastrophe describes systems characterized by multistability, bifurcations, and hysteresis, making it applicable to real-world scenarios such as ecological thresholds, market crashes, and biological tipping points [6, 7].

Mathematically, it is defined through a potential function $V(x)$:

$$V(x) = \frac{1}{4}x^4 - \frac{1}{2}ux^2 - vx,$$

where u and v are control parameters, and x is the response variable. The equilibrium states of the system are determined by the stationary condition:

$$\frac{\partial V}{\partial x} = x^3 - ux - v = 0,$$

Resulting in a cusp-shaped surface in (u, v, x)-space [8]. This model captures the interplay of control parameters and response, offering insights into the bifurcations and sudden state transitions inherent in non-linear systems.

Despite its theoretical elegance, accurately modeling cusp catastrophe behavior from real-world data remains a formidable challenge. Nonlinearities, coupled with noise, sparsity, and irregular sampling, complicate the approximation of the mapping between control parameters (u, v) and the response variable x. sudden transitions and bifurcations require techniques that can effectively capture localized features in data [9]. Traditional regression methods often fall short in addressing these complexities.

Recent advancements in machine learning offer promising alternatives for tackling these challenges. Techniques such as deep neural networks and reservoir computing have demon started success in approximating catastrophe surfaces and predicting tipping points directly from data [10, 11]. Among these, the Random Forest algorithm, an ensemble-based learning method introduced by Breima , is particularly well-suited for predicting cusp catastrophe surfaces [12]. By aggregating predictions from multiple decision trees trained on bootstrap samples, Random Forest can model complex, non-linear relationships while maintaining robustness to noise and outliers. Its ability to identify feature importance further enhances its utility for understanding the influence of control parameters [13].

Previous studies have validated the application of machine learning to catastrophe prediction. For example, Chen and Chen applied logistic cusp regression to binary out comes, highlighting the model's capacity to handle abrupt transitions [14]. Similarly, Cross and Wheat demonstrated the effectiveness of Random

Forest in analyzing cusp like dynamics. Building upon these foundations, this study ventures into the largely uncharted territory of predicting the unstable regions of the cusp surface using Random Forest regression [9]. By tackling this intricate challenge, it seeks to deepen our understanding of these complex systems, shedding light on their chaotic dynamics and pushing the boundaries of machine learning's applicability in modeling nonlinear phenomena. Understanding this region enables a deeper insight into the mechanisms driving system instability and the onset of catastrophic transitions. This study focuses specifically on the instability region within the broader spectrum of multiple equilibria, assessing the model's capability to approximate the complex mapping between (u, v) and x. The analysis encompasses the model's effectiveness in capturing bifurcations, hysteresis, and abrupt state transitions defining characteristics of cusp catastrophe systems. By systematically evaluating predictive accuracy within these critical regions, this work seeks to rigorously examine the robustness and applicability of Random Forest regression in modeling the intricate dynamics underpinning catastrophe theory.

The paper is organized as follows: Section 2 outlines the methodology, detailing the Random Forest framework and experimental setup. Section 3 examines the impact of data uncertainty on model performance, while Section 4 investigates the effects of data availability. Section 5 considers model complexity and computational effectiveness. Finally, the conclusions and implications of the findings are discussed in Section 6.

Methodology

Random Forest Regression is an ensemble learning technique designed to approximate complex, non-linear relationships in data by aggregating predictions from multiple decision trees. It combines the principles of bootstrap sampling, random feature selection, and recursive partitioning to deliver robust predictions while minimizing overfitting [12].

The model starts with a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i = [u_i, v_i]$ are the input features and y_i is the corresponding target value. Random Forest constructs an ensemble of T decision trees, where each tree is trained on a unique bootstrap sample a randomly drawn subset of the original training data with replacement. This ensures that each tree sees a slightly different dataset, introducing diversity into the ensemble.

At each node within a decision tree, the algorithm identifies an optimal split of the data. The split is determined by selecting a feature $j \in Ft$ from a randomly chosen subset of features $Ft \subseteq F$ and a threshold value s. The data is divided into two child nodes based on whether the feature value satisfies $xi,j \leq s$ or $xi,j > s$. To evaluate the quality of a split, the algorithm minimizes the weighted Mean Squared Error (MSE) for the two resulting nodes:

$$\mathrm{MSE}_{\mathrm{split}} = \frac{N_{\mathrm{left}}}{N}\mathrm{MSE}_{\mathrm{left}} + \frac{N_{\mathrm{right}}}{N}\mathrm{MSE}_{\mathrm{right}}, \qquad (1)$$

where Nleft and NRight are the number of samples in the left and right nodes, respectively.
The MSE for a given node is defined as:

$$\mathrm{MSE}_{\mathrm{node}} = \frac{1}{N_{\mathrm{node}}} \sum_{i \in \mathrm{node}} (y_i - \bar{y}_{\mathrm{node}})^2, \qquad (2)$$

where y_{node} is the mean of the target values in the node.

This splitting process continues recursively, partitioning the data into smaller and more homogeneous subsets at each step. The recursion stops when one of the following conditions is met: the maximum tree depth is reached, the number of samples in a node falls below a predefined threshold, or further splitting does not significantly reduce the MSE.

Random Forest introduces two key sources of randomness to reduce overfitting and improve generalization: bootstrap sampling and random feature selection. Bootstrap sampling ensures variability by training each tree on a slightly different dataset. Random feature selection further enhances diversity by limiting the candidate features for splits at each node. By combining these strategies, Random Forest reduces the correlation between trees and improves the model's ability to generalize to unseen data.

Once all T decision trees are constructed, the Random Forest prediction for a new input x is obtained by averaging the outputs of all individual trees:

$$\hat{f}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}), \qquad (3)$$

where $ft(x)$ is the prediction from the t-th tree in the ensemble. This averaging reduces variance and smooths out the predictions, leading to a more stable and accurate model.

To estimate the model's generalization error, Random Forest uses the Out-of-Bag (OOB) error. During training, each tree is built on a bootstrap sample, leaving out approximately 36.8% of the data points, referred to as OOB samples. The OOB error is calculated as the mean squared error of predictions for these OOB samples:

$$\text{OOB Error} = \frac{1}{|\mathcal{O}|}\sum_{i \in \mathcal{O}} \left(y_i - \hat{f}_{\mathrm{OOB}}(\mathbf{x}_i)\right)^2, \qquad (4)$$

where $\hat{f}_{\mathrm{OOB}}(\mathbf{x}_i)$ is the average prediction from all trees where xi was not included in the bootstrap sample. The OOB error serves as a reliable estimate of the model's performance without requiring a separate validation set.

The computational complexity of Random Forest depends on the number of trees T, the number of samples N, and the number of features M. For a dataset with N samples and M features, the time complexity for building the Random Forest is approximately:

$$\mathcal{O}(T \cdot N \cdot M \cdot \log N), \qquad (5)$$

where the log$N$ term corresponds to the average depth of a decision tree in a balanced split. This makes Random Forest efficient for moderately sized datasets and scalable to larger problems when combined with parallel computation.

The Random Forest algorithm builds multiple decision trees using randomized data and features, aggregates their predictions through averaging, and estimates its generalization error using the OOB samples. By incorporating randomness and ensemble averaging, Random Forest provides a robust and versatile solution for regression tasks, capable of capturing complex, non-linear relationships while minimizing overfitting and reducing variance.

Analysis of Model Performance: Data Noise

The Random Forest model was evaluated on datasets with increasing levels of randomness, ranging from 0.1% to 30%. To

assess the model's performance, three key metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Additionally, residual plots, actual vs. predicted plots, and feature importance plots were analyzed to understand the behavior of the model under varying
levels of noise.

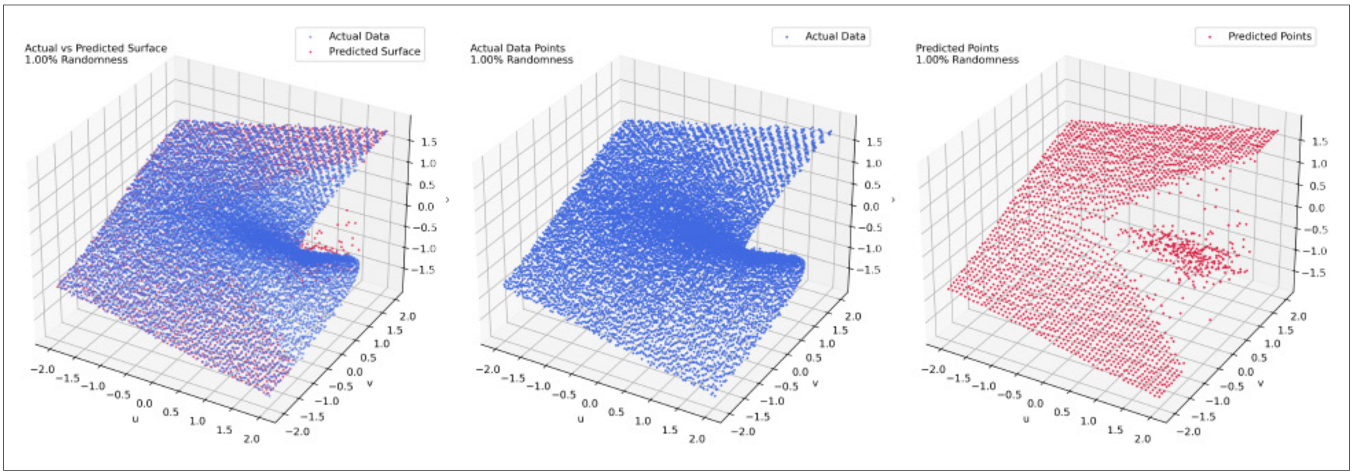

Figure 1: 1% Randomness: Actual Versus Predicted Values

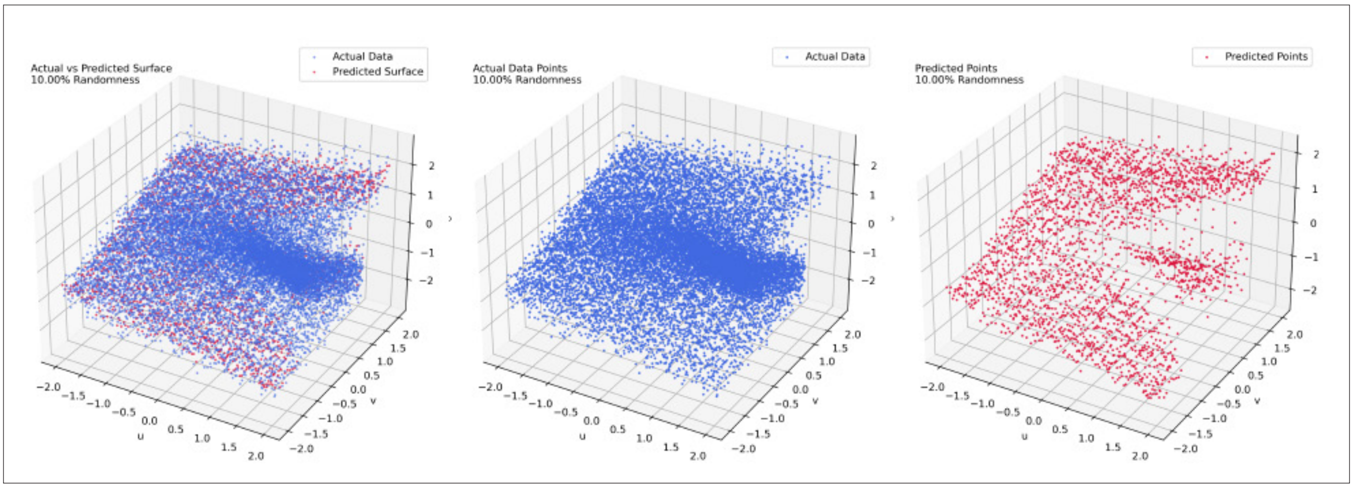

Figure 2: 10% Randomness: Actual Versus Predicted Values

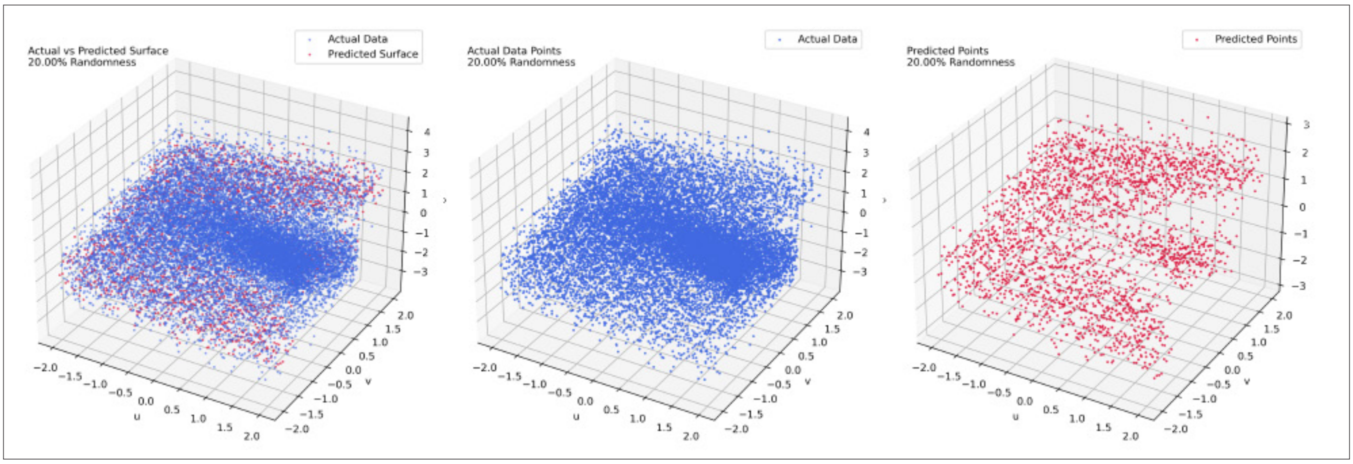

Figure 3: 20% Randomness: Actual Versus Predicted Values

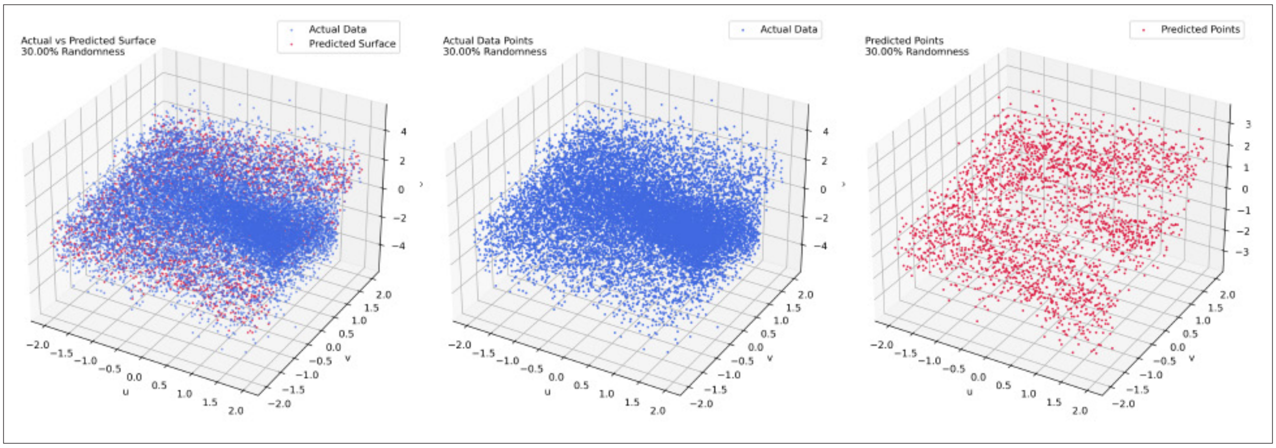

Figure 4: 30% Randomness: Actual Versus Predicted Values

The performance of the model, measured by MAE, RMSE, and R^2, exhibits a clear trend as the level of randomness increases. At low randomness levels (0.1%–2%), the model performs exceptionally well, achieving low MAE and RMSE values, and R2 scores close to

This indicates that the model accurately captures the relationships between the input features *u*, *v*, and the target variable *x*. As the randomness level increases (5%–10%), the errors (MAE and RMSE) gradually rise, and the R^2 score declines. This behavior reflects the model's decreasing ability to predict accurately due to the noise in the data. At high randomness levels (20%–30%), the model performance deteriorates significantly, with large errors and R^2 scores approaching zero or negative values, indicating that the model fails to explain the variance in the target variable.

Scatter plots comparing actual and predicted values provide further insight into the model's accuracy [Figures: 1, 2, 3, 4]. at low levels of randomness, the predicted values align closely with the actual values along the diagonal line $y = x$, indicating highly accurate predictions. However, as randomness increases, the scatter around the diagonal line becomes more pronounced, with points deviating significantly from the line. At very high random- ness levels (e.g., 20% and above), the scatter plot appears almost random, reflecting the model's inability to capture meaningful relationships due to noise dominance.

Residual plots were used to visualize the prediction errors (Residual = $y_{\text{actual}} - y_{\text{predicted}}$) across different levels of randomness. At low randomness levels, the residuals are small and randomly distributed around zero, indicating a good model fit with no significant bias. As the noise in the data increases, the residuals become more widely scattered, with a larger magnitude of errors. At high randomness levels, the residuals exhibit no discernible structure and deviate significantly from the zero line, further confirming the model's reduced predictive accuracy [Figures: 5, 6, 7, 8].

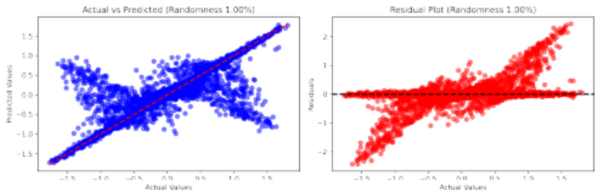

Figure 5: Residual at 1%

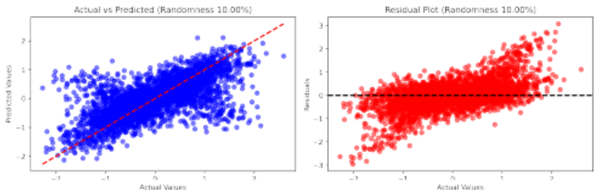

Figure 6: Residual at 10%

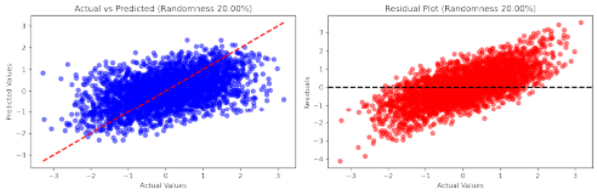

Figure 7: Residual at 20%

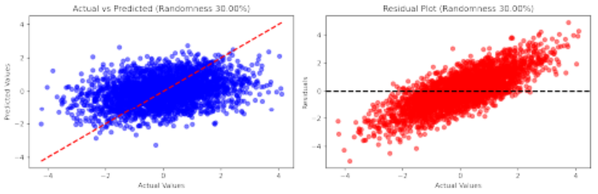

Figure 8: Residual at 30%

The feature importance plot highlights the relative contributions of the input features *u* and *v* in the Random Forest model [Figures: 9, 10, 11, 12]. At low noise levels, both features contribute meaningfully to the model predictions, with relatively balanced importance scores. This suggests that the model effectively leverages the information provided by both input features. However, as randomness increases, the noise obscures the relationships be- tween the features and the target variable, reducing their relative importance.

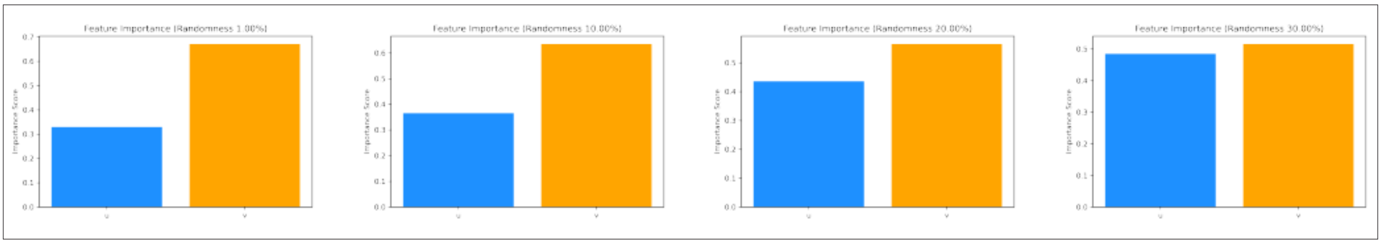

Figure 9: 1% Noise **Figure 10:** 10% Noise **Figure 11:** 20% Noise **Figure 12:** 30% Noise

Figure 13: Feature Importance Analysis

Figure 13 shows feature analysis reveals a notable shift in the balance of importance between features *u* and *v* as data noise increases. At low noise levels (e.g., 1%), the model assigns disproportionate importance to v, with the feature balance skewed at 30-70 in favor of *v*. However, as noise increases to 30%, the balance gradually equalizes, approaching a near 50-50 distribution. This suggests that higher noise levels compel the model to leverage both features more equitably, possibly as a compensatory mechanism to preserve predictive performance amidst diminishing signal clarity. These findings highlight the adaptive interplay between feature importance and data quality in Random Forest models.

Figure 14 summarizes the behavior of MAE, RMSE, and R^2 across all levels of randomness. At low randomness (0.1%–2%), the model achieves low error values (MAE and RMSE) and high R^2 scores, indicating strong predictive performance. At moderate randomness (5%–10%), prediction errors begin to increase, and the R^2 score starts to decline, reflecting reduced accuracy. At high randomness (20%–30%), MAE and RMSE increase significantly, and R^2 approaches zero or negative values, indicating that the model is unable to explain the variance in the target variable.

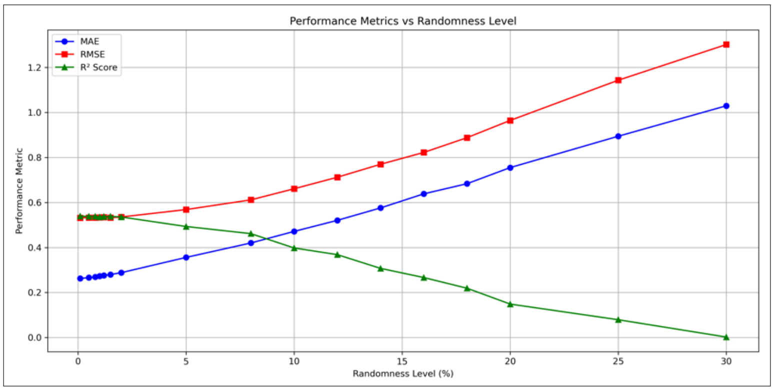

Figure 14: Performance Metrics (Mae, Rmse, and R2) As a Function of Randomness Levels.

The analysis of the performance metrics, residual plots, and feature importance high lights the impact of increasing noise on the Random Forest model as shown in figure 14. At low levels of randomness, the model effectively captures the underlying relationships in the data and produces accurate predictions. However, as randomness increases, the model's performance declines due to the noise overwhelming the signal. This behavior is consistent with expectations, as high noise levels reduce the signal to noise ratio, making it more challenging for the model to generalize.

In summary, the Random Forest model demonstrates strong performance at low noise levels but struggles as the randomness increases. To improve model performance un- der noisy conditions, the following recommendations are proposed: preprocess the data to reduce noise through smoothing or denoising techniques, experiment with alternative models such as Gradient Boosting or noise-tolerant techniques, and investigate additional features or transformations that may help the model better generalize under noisy conditions. These findings provide a comprehensive understanding of the model's behavior and its limitations, informing future improvements for handling noisy datasets.

Analysis of Model Performance: Data Reducation

The performance of the Random Forest regression model was systematically evaluated across datasets with varying levels of randomness and data availability to assess its robust- ness and predictive accuracy. Randomness levels ranged from 0.1% to 30%, simulating increasing levels of noise in the data, while subset percentages ranged from 100% to 30%, representing different levels of data availability [Figures: 15, 16, 17, 18].

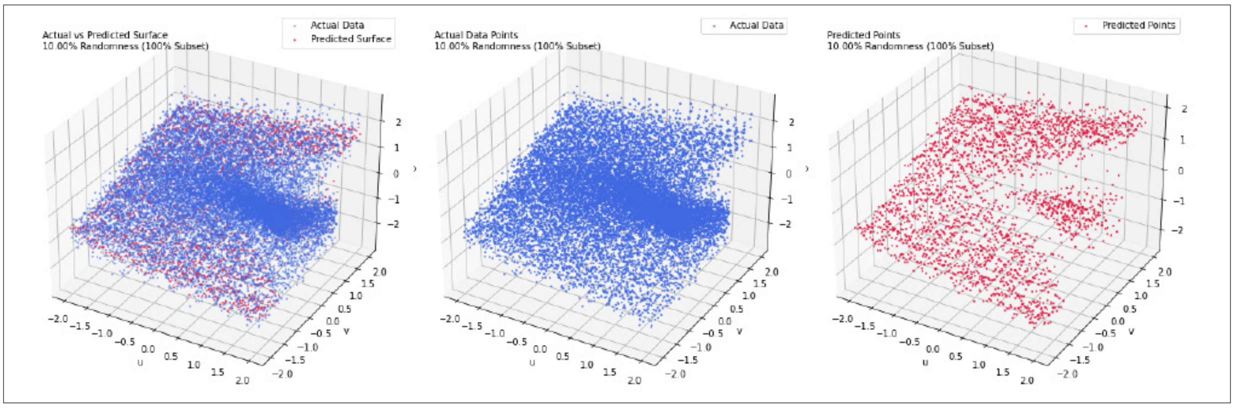

Figure 15: 100% Data: Actual Versus Predicted Data, 10% Noise

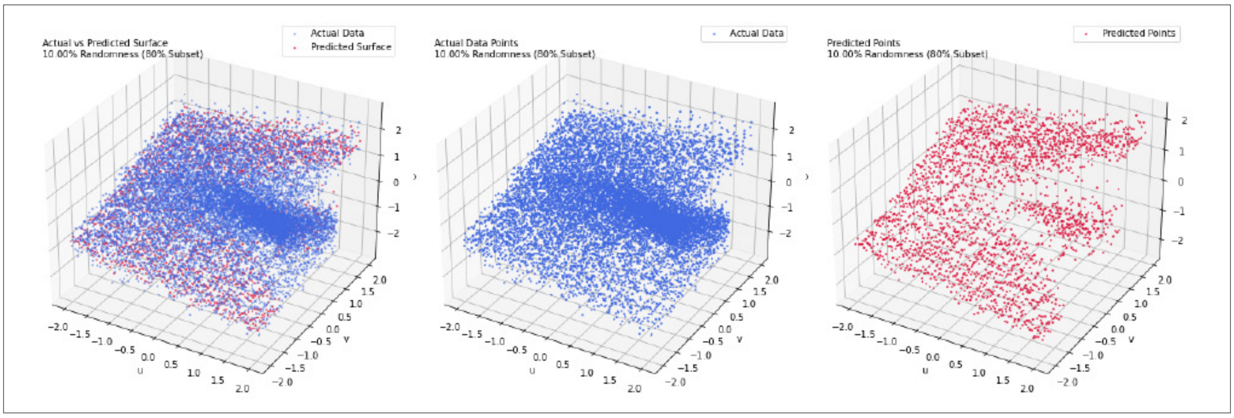

Figure 16: 80% Data: Actual Versus Predicted Data, 10% Noise

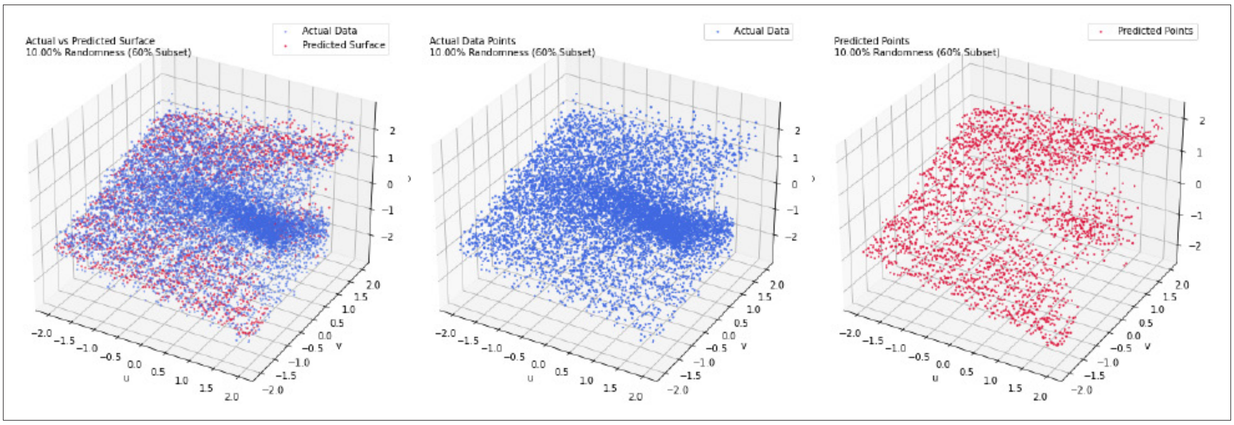

Figure 17: 60% Data: Actual Versus Predicted Data, 10% Noise

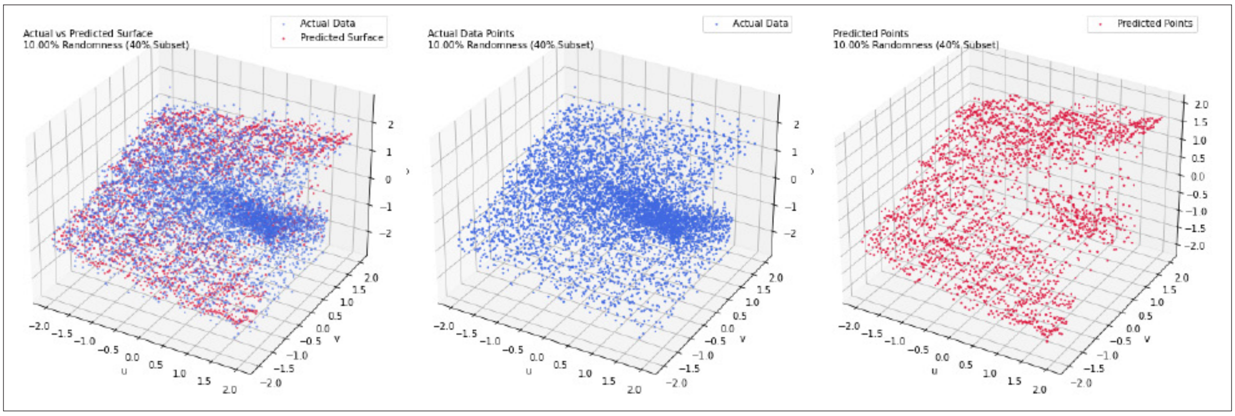

Figure 18: 40% Data: Actual Versus Predicted Data, 10% Noise

At low randomness levels (0.1% and 1%), the model exhibited excellent performance, achieving high R^2 values (above 0.98) and low error metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These results highlight the model's capacity to accurately capture the underlying relationships in the data when noise is minimal. However, as randomness increased to higher levels (10%–30%), performance deteriorated significantly. MAE and RMSE values rose steadily, while R^2 scores dropped below 0.7 in some cases, demonstrating the sensitivity of the model to noise in the data. This performance degradation was particularly evident in smaller subsets, where the combined effect of limited data and high noise further hindered the model's ability to generalize [Figures: 19, 20, 21, 22].

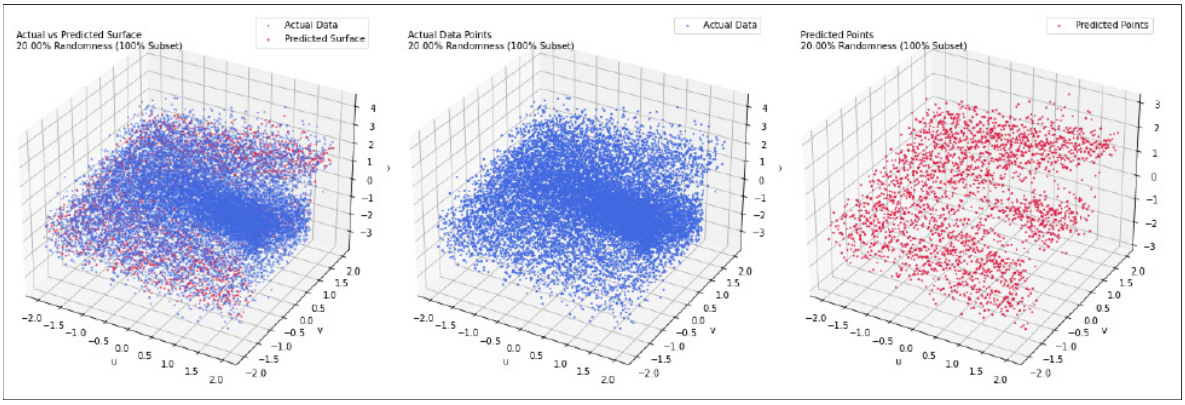

Figure 19: 100% Data: Actual Versus Predicted Data, 20% Noise

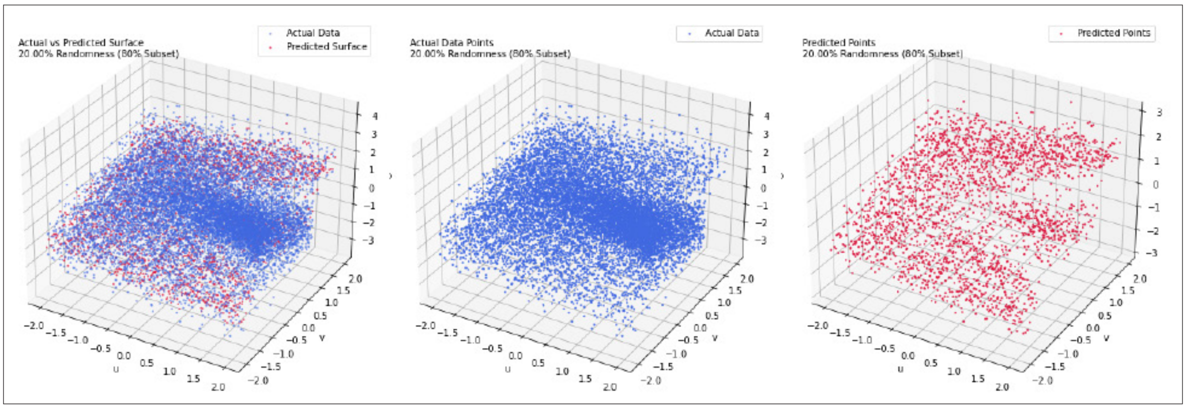

Figure 20: 80% Data: Actual Versus Predicted Data, 20% Noise

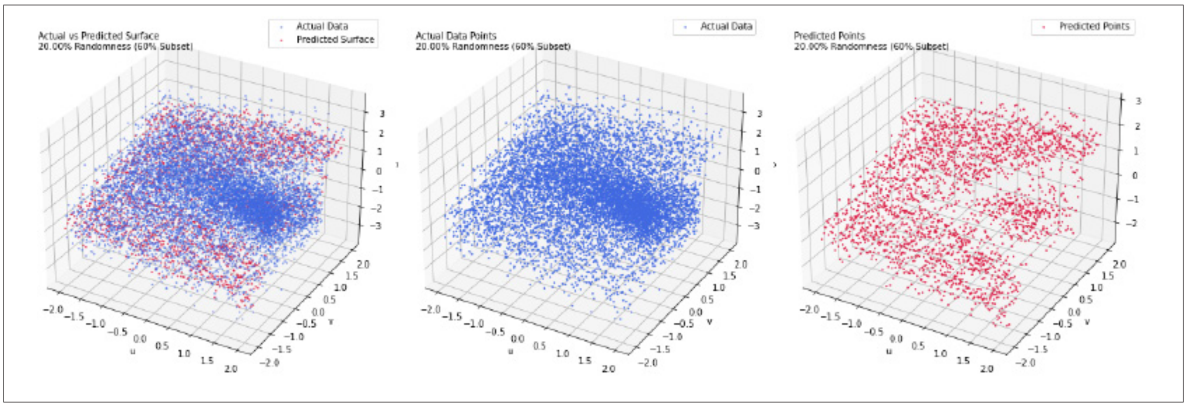

Figure 21: 60% Data: Actual Versus Predicted Data, 20% Noise

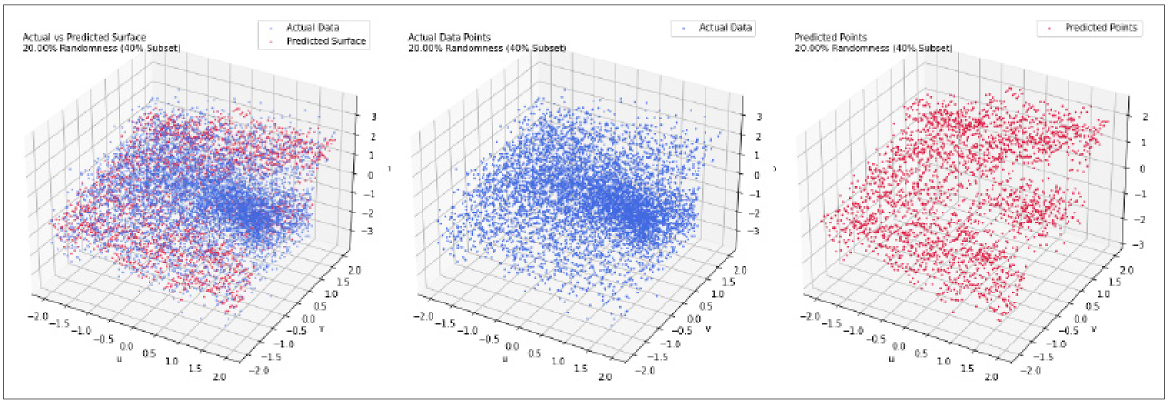

Figure 22: 40% Data: Actual Versus Predicted Data, 20% Noise

The model's performance was strongly influenced by the size of the data subsets. Larger subsets (100%–80%) consistently outperformed smaller ones, particularly at higher randomness levels. With access to more training data, the model effectively mitigated the adverse effects of noise, resulting in better predictive accuracy and generalizability. Conversely, smaller subsets (50%–30%) showed significant drops in performance metrics, especially when coupled with high randomness. For instance, at a randomness level of 20%, the RMSE for the 30% subset increased by over 50% compared to the full dataset, and the R2 score fell below acceptable thresholds ($R^2 < 0.5$) [Figures: 23, 24].

The inclusion of cross validation during training provided reliable performance estimates and enhanced the model's robustness across different data configurations. Despite the challenges posed by noise and reduced data availability, the model maintained a consistent level of stability in its predictions, reflecting the inherent strengths of the Random Forest algorithm in ensemble learning.

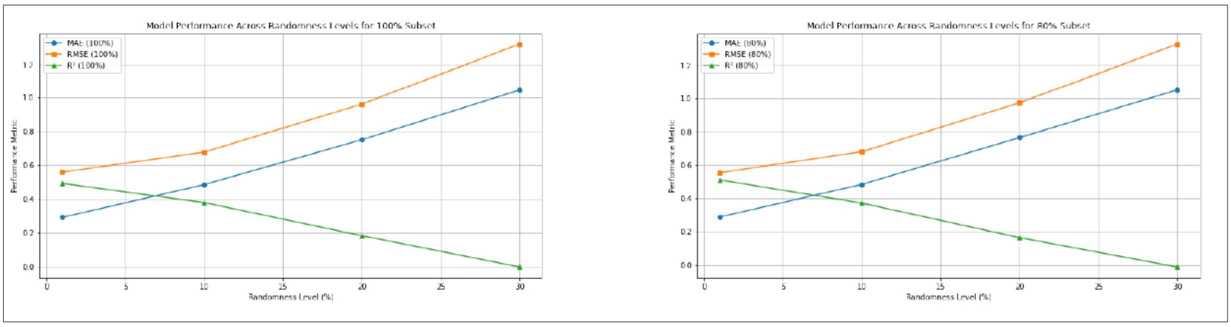

Figure 23: Model Performance for 100% and 80% Data

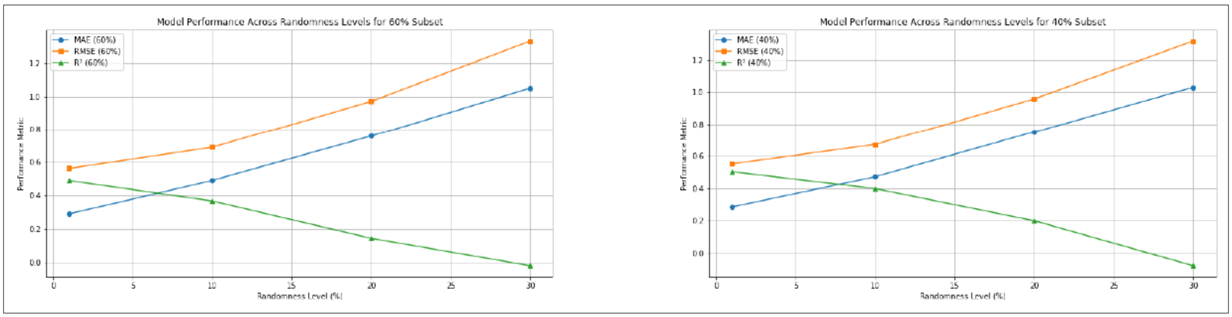

Figure 24: Model Performance for 60% and 40% Data

The analysis highlights the Random Forest regression model's robustness in scenarios with adequate data and low noise, where it reliably captures complex patterns and relation- ships. However, its sensitivity to noise and reduced data availability underscores the need for additional measures, such as noise reduction techniques, feature engineering, or model ensembling, to further enhance performance in challenging conditions. These findings provide valuable insights for deploying Random Forest models in real-world applications, particularly in noisy or resource-constrained environments.

Analysis of Model Performance: Model Complexity

Random Forest models are celebrated for their predictive robustness and versatility; how- ever, their computational complexity is intrinsically tied to hyperparameter choices, particularly tree depth and the number of trees. In this study, we systematically analyze the time complexity of Random Forest regressors by examining training and prediction times across varying model configurations. Specifically, tree depth (5, 10, and 15) and the number of trees (50, 100, and 200) were varied, using synthetic datasets with controlled randomness to ensure consistency. For each combination, average training and prediction times were meticulously recorded and analyzed.

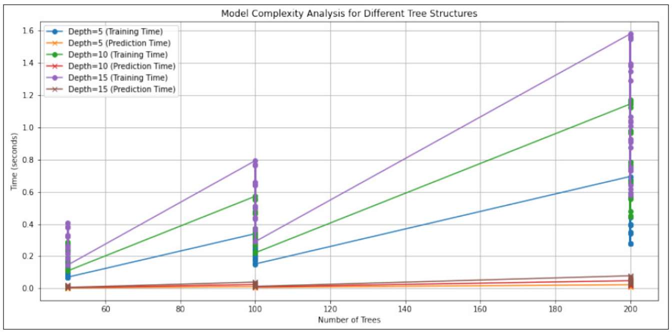

Figure 25: Model Complexity Analysis

The results reveal a linear increase in training time with the number of trees, regardless of tree depth. However, deeper trees (e.g., depth = 15) consistently demand significantly more computational resources during training compared to shallower counterparts (e.g., depth = 5), due to the greater complexity of their branching structures. Prediction time, while substantially lower than training time, also increases with the number of trees but is less affected by tree depth. This is attributed to the marginally higher traversal cost associated with deeper trees during prediction [Figure: 25].

These findings illuminate the delicate balance between model complexity and computational efficiency. Deeper trees and a larger number of estimators enhance the model's capacity to capture intricate data patterns, but at the expense of increased computational overhead. This trade off is particularly critical in time-sensitive applications, such as real time predictions, where efficiency must be weighed against accuracy. Practitioners are advised to judiciously calibrate hyperparameters to align with the specific demands of their use cases, ensuring optimal performance without undue computational burden.

Conclusion

This study delved into the utility of Random Forest regression models for predicting cusp instability regions of catastrophe surfaces under varying conditions of noise and data avail ability. The findings affirm that Random Forests are highly effective in capturing the intricate, non-linear dynamics of cusp catastrophe systems, particularly in scenarios with abundant data and minimal noise. At lower randomness levels (e.g., 0.1%–2%), the model consistently achieved superior predictive accuracy, as evidenced by low MAE and RMSE values and R2 scores approaching unity. These results underscore the model's ability to reliably characterize hallmark features of cusp catastrophe systems, including bifurcation and hysteresis. Abrupt state transitions are not considered in this paper.

Feature analysis further revealed an adaptive shift in the balance of feature importance, with the model leveraging both *u* and *v* more equitably as noise levels increased. This dynamic interplay underscores the model's inherent flexibility in adjusting feature reliance to maintain predictive robustness under varying data conditions.

The model's performance declined significantly as randomness increased (10%– 30%), with rising errors and decreasing R^2 scores, particularly when data availability was limited. This sensitivity to noise was most pronounced in smaller data subsets (30%–50%), where the combined effects of noise and sparsity led to marked performance degradation. These observations highlight the critical role of data quality and quantity in maintaining predictive robustness, as well as the value of techniques such as cross-validation to bolster the model's generalizability.

The study also explored the relationship between model complexity and computational efficiency, revealing critical trade-offs inherent in Random Forest configurations. While increasing tree depth and the number of trees enhanced the model's capacity to capture complex data patterns, these adjustments came with substantial computational costs. Training times scaled almost linearly with the number of trees, while deeper trees (e.g., depth = 15) disproportionately increased computational overhead due to their complex branching structures. Prediction times, though less sensitive to tree depth, also scaled with the number of trees, underscoring the importance of balancing these hyperparameters to achieve both efficiency and accuracy. These findings emphasize that optimal model performance is not solely dependent on predictive accuracy but must also account for computational feasibility, especially in resource-constrained or time-sensitive applications. To address the challenges posed by high noise and limited data, several strategies merit consideration. Preprocessing methods for noise reduction, alternative machine learning models such as Gradient Boosting, and advanced feature engineering could further improve performance.These refinements may help mitigate the model's sensitivity to adverse conditions and expand its applicability to more complex real world scenarios.

In conclusion, Random Forest regression offers a robust framework for modeling cusp catastrophe surfaces under favorable conditions. However, its sensitivity to noise, data sparsity, and computational complexity underscores the need for refinement and careful hyperparameter tuning. Future research should prioritize integrating noise resistant methodologies, leveraging domain-specific insights, and developing strategies to manage computational costs. By addressing these challenges, the potential of machine learning to elucidate nonlinear dynamical systems can be more fully realized, broadening its applicability to a wider array of scientific and practical problems [15, 16].

References

1. Stiefenhofer P (2013) The catastrophe map of a two-period production model with uncertainty. Applied Mathematics 4: 8.
2. Stiefenhofer P (2014) Topological properties of the catastrophe map of a general equilibrium production model with uncertain states of nature. Applied Mathematics 5: 2719-2727.
3. Stiefenhofer P (2017) Catastrophe theory and postmodern general equilibrium: Some critical observations. Applied Mathematical Sciences 11: 2383-2391.
4. Thom R (1972) Stabilité structurelle and morphogenèse. Paris: W A Benjamin 5: 1-4.
5. Zeeman EC (1976) Catastrophe theory. Scientific American 234 : 65-83.
6. Gilmore R (1993) Catastrophe theory for scientists and

engineers. New York: Dover Publications https://www.amazon.in/Catastrophe-Theory-Scientists-Engineers-Gilmore/dp/0486675394.

7. Chen S, Chen H (2017) Cusp catastrophe regression and its application in public health and behavior research. International Journal of Environmental Research and Public Health 14 : 1220.
8. Arnold VI (1986) Catastrophe theory (3rd ed.). Berlin: Springer https://link.springer.com/book/10.1007/978-3-642-58124-3.
9. Cross DJ, Wheat SA (2019). Ghost hunting in the nonlinear dynamic machine. Plos One 14: e0226572.
10. Daw R, He Z (2020) Deep neural network in cusp catastrophe model. arXiv preprint arXiv: 2004.02359.
11. Köglmayr D, Räth C (2023) Extrapolating tipping points and simulating non-stationary dynamics of complex systems using efficient machine learning. arXiv preprint arXiv: 2312.06283.
12. Breiman L (2001) Random forests. Machine Learning 45 : 5-32.
13. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York: Springer 002E.
14. Chen S, Chen H (2020) Logistic cusp catastrophe regression for binary out- come: Method and application. In Statistical Models for Data Analysis 187-200.
15. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, et al. (2007) Random forests for classification in ecology. Ecology 88: 2783-2792.
16. Wager S, Hastie T (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. Journal of Machine Learning Research 15: 1625-1651.