Comparative Analysis of Interpretability of Simple and Complex Machine Learning Models in Presence of Noise

Tong Zhen Hao

Huazhong University of Science and Technology

Abstract： This paper offers a comprehensive analysis of the interpretability of key Machine Learning models, including ElasticNet regression, Random Forest, and Neural Networks, when faced with various types of noise. Focusing on both synthetic and real-world datasets of diverse sizes (385 to 15,000 samples), the study probes the models' ability to detect hidden patterns, especially in the presence of varied noise conditions (Gaussian, Perlin, and Simplex). Through systematic evaluation using Permutation Feature Importance (PFI) and SHAP summary plots, our research reveals a strong correlation between dataset size and model robustness to noise perturbations. The results demonstrate that larger datasets consistently lead to more stable feature importance rankings and better preservation of model interpretability under noise conditions. While ElasticNet shows superior performance on larger datasets, Neural Networks prove most sensitive to noise, particularly with smaller datasets. The findings provide valuable insights for practical applications of machine learning, suggesting that emphasis should be placed on acquiring larger training datasets to ensure robust and trustworthy model interpretations in noisy environments. This work contributes to the broader understanding of ML model interpretability and provides guidance for model selection in real-world applications where data noise is inevitable.

Keywords: Machine Learning, Model Interpretability, Noise Robustness, Feature Importance, SHAP Analysis

DOI：10.63887/jtie.2025.1.1.8

1.Introduction

Machine Learning (ML), as a part of artificial intelligence, focuses on creating algorithms that help computers learn from data. This learning imitates human patterns to refine predictions or decisions by the ML model [1]. ML is vital in data science, helping to draw insights from large datasets using statistical and computational methods [2]. These insights from data mining guide decision-making, expanding the use of ML [3].

Understanding the trustworthiness and interpretability of models, especially in noisy environments, has become increasingly crucial. Interpretable ML extracts key knowledge from models about data relationships, offering insights for specific audiences on chosen issues, guiding actions, and is displayed as visuals, language, or equations based on context. Recent research has focused extensively on explainable AI, with some studies providing comprehensive overviews of interpretation methods, particularly

emphasizing post hoc deep learning interpretations[4,5]. Other research evaluates interpretation qualities[6-8], while some explore method similarities[9,10]. The field has also become central to discussions about ML bias and fairness [11-13].

However, a significant gap exists in understanding how different ML models maintain their interpretability and trustworthiness when faced with various types of noise. While traditional evaluation metrics provide quantitative measures of model performance, they fail to capture the nuanced ways in which noise affects a model's decision-making process.This research aims to analyze and compare the interpretability of three notable ML models: ElasticNet regression, Random Forest, and Neural Networks, particularly in the presence of noise. By utilizing both synthetic and real-world datasets, the study examines the capacity of these models to identify hidden patterns under various noise conditions. The central concern revolves around assessing the precision and trustworthiness of these models when their interpretability is challenged by noise.

Specifically, this study focuses on three objectives: Examining how different noise patterns affect model interpretability across varying dataset sizes. Understanding the relationship between dataset complexity and model robustness to noise. Evaluating which models maintain better interpretability under noisy conditions. The findings from this research have significant implications for both theoretical understanding and practical applications of ML, particularly in domains where both accuracy and interpretability are crucial, such as healthcare, finance, and industrial applications.

2. Related Work

2.1 Machine Learning

Machine Learning (ML) is a part of artificial intelligence focusing on creating algorithms that help computers learn from data. This learning imitates human patterns to refine predictions or decisions by the ML model. ML is vital in data science, helping to draw insights from large datasets using statistical and computational methods.

2.2 Regression Algorithms and Model Selection

Regression algorithms help uncover relationships between an outcome and its features. Linear regression assumes a straight-line relationship between variables - its simplicity makes it efficient, but it may not fit non-linear data. Regularization techniques like Lasso and Ridge help prevent overfitting. The Random Forest algorithm[14] uses ensemble learning[15] to combine classifiers, boosting model robustness and accuracy. Neural Networks are models mimicking biological neural systems, used to understand complex data patterns.

2.3 Machine Learning Interpretability

Interpretable ML extracts key knowledge from models about data relationships. Recent research has focused extensively on explainable AI, with some studies providing comprehensive overviews of interpretation methods, particularly emphasizing post hoc deep learning interpretations. Other research evaluates interpretation qualities while some explore method similarities. The field has also become central to discussions about ML bias and

fairness.

2.4 Effects of Noise on Model Performance

Limited research has been conducted on how different noise patterns affect model interpretability and performance. While traditional evaluation metrics provide quantitative measures of model performance, they fail to capture the nuanced ways in which noise affects a model's decision-making process. The relationship between dataset size, model complexity, and noise resilience remains an understudied area that this research aims to address.

3.Methodology

3.1 Noise Pattern Design

To comprehensively evaluate model interpretability under different disturbance conditions, three distinct noise patterns were systematically introduced. The “level” (low, medium, high) relates to the magnitude or intensity of the noise in relation to the original data's spread.

3.1.1 Gaussian Noise

For the “low”level noise implementation, Gaussian noise was selected due to its well-defined statistical properties. The probability density function (PDF) of the Gaussian distribution is defined as:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (1)$$

In our experimental setup, we introduced Gaussian noise with μ = 1 and σ proportional to the original feature's variability. Figure 1 illustrates the distribution changes in the Random dataset's top five features after introducing Gaussian noise.

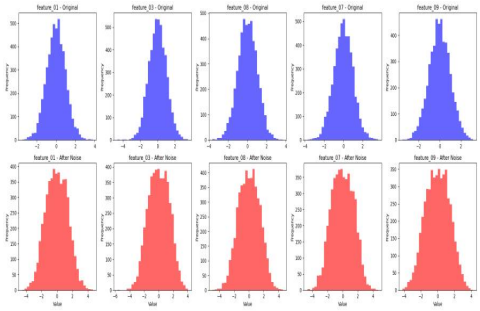


Fig. 8: Perturbation of the five paramount features in the Random dataset using Gaussian noise. Post perturbation, discernible alterations in the central tendencies and dispersions (μ and σ) of the features are evident.

3.1.2 Perlin Noise

The “medium” level noise employs Perlin noise[16], a gradient noise function that provides continuous, smooth variations across its domain. The implementation follows a three-step process:

Grid Creation: Establishing a grid of random gradient vectors;

Dot Product: Computing dot products between distance and gradient vectors;

Interpolation: Smooth interpolation between calculated values.

The Perlin noise function can be mathematically represented as:

$$f(x) = \sum_{i=0}^{n} a_i \cdot g(b_i \cdot x) \quad (2)$$

Where n is set to 100 octaves, ai represents the amplitude, and bi denotes the frequency for the ith octave. Figure 2 demonstrates the effects of Perlin noise on the Random dataset's features.

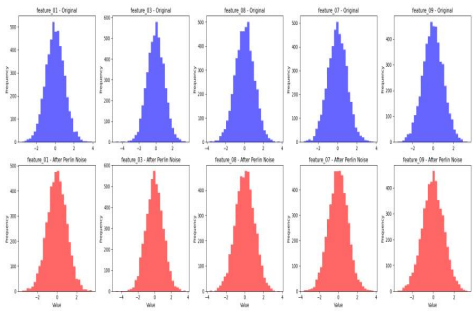


Fig. 2: Influence of Perlin noise on the predominant five features of the Random dataset.

3.1.3 Simplex Noise

For the “high” level noise, Simplex noise was implemented as an n-dimensional noise function. The space is divided into simplexes (triangles in 2D, tetrahedra in 3D), with each vertex associated with a gradient. The noise value computation involves: Computing dot products between position vectors and gradients; Interpolating across the simplex structure; Combining multiple octaves for the final noise value.The comparative effects of all three noise patterns are visualized in Figure 3, showing their impact on feature distributions.

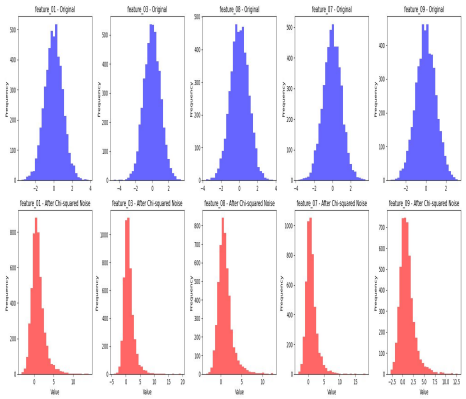


Fig. 3: Influence of Simplex noise on the predominant five features of the Random dataset.

3.2 Datasets

In order to conduct a comparative analysis of the robustness of ML interpretability, this study employs three distinct datasets of varying sizes, each applied to a regression problem. These datasets have been chosen to represent a range of scenarios, from small-scale to large-scale data. Detailed information about each dataset is provided in Table 1.

Table 1: Datasets comparison.

Name	Number of instances	Number of features	Target name	Domain
Elongation	385	17	Elong	Steel
Random	5000	10	label	Synthetic
Chemical	15000	15	D	Chemical

3.2.1 Elongation Dataset

The Elongation dataset, originating from the metallurgical foundry domain, encompasses 385 observations with 17 predictor variables representing the chemical constituents of a distinct steel alloy (like C, Si, Mn, P, S, Cu, Mg, etc.). The dependent variable, designated as "Elong", quantifies the extent of elongation for the specific steel type.

Feature correlation analysis (Figure 4) reveals that 'C' positively correlates with 'SI', while negatively correlating with 'MG'. The target 'Elong' exhibits varied correlations with features, particularly negative correlations with 'C' and 'SI'. Most features show minimal inter-correlation.

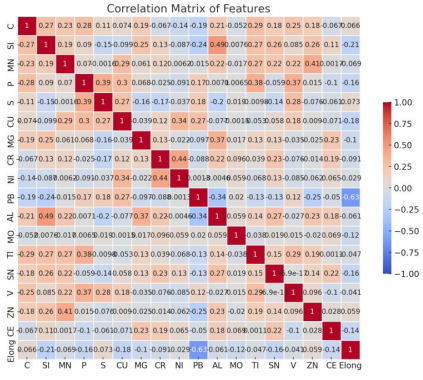


Fig. 4: Correlation matrix of the features in the elongation dataset

For evaluating model trustworthiness, five pivotal determinants were identified based on domain expertise: “PB”,“SI”, “P”, “AL”, and “MO”. The relationship was defined as:

$$y = \sin(\mathrm{PB}) + \mathrm{SI}^2 + \log(\mathrm{P} + 5) - \sqrt{|\mathrm{AL}|} + \exp(-\mathrm{MO})$$

3.2.2 Random Dataset

Using Scikit-learn [17], a synthetic dataset was generated with 5,000 samples and 10 features, of which 5 are informative. The make_regression method was employed to create this controlled environment. Correlation analysis (Figure 5) shows most features have low to moderate correlations, indicating limited multicollinearity. The target 'label' demonstrates strong correlations with specific features while maintaining weak associations with others.

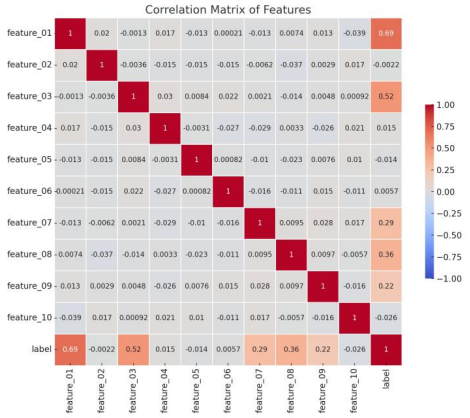


Fig. 5: Correlation matrix of the features in the random dataset.

3.2.3 Chemical Dataset

The Chemical dataset comprises 15 features, labeled as 'I1' through 'I15', representing specific chemical properties or measurements relevant to the industry. The target variable 'D' represents a specific outcome property. Correlation analysis (Figure 18) shows low to moderate correlations between features, with the target 'D' showing notable correlations with 'I1', 'I5', and 'I14'.

To examine the machine learning's ability to find hidden patterns, a non-linear relationship between top 5 features and the label was explicitly defined as:

$$y = I1 + I2^2 + e^{I3} + I4 \cdot I5 + \log(I2 + 1)$$

This diverse dataset selection allows for a comprehensive evaluation of model interpretability across different scales and domains, providing insights into how dataset characteristics influence model robustness under noise conditions.

3.3 Model Training and Hyperparameter Settings

Hyperparameters in ML algorithms significantly influence the performance and predictive accuracy of models. These parameters, which are set prior to the commencement of the learning process, govern the behavior of the algorithms. In this study, Bayesian optimization was selected as the hyperparameter tuning method due to its efficiency and ability to handle high-dimensional parameter spaces [18-22].

3.3.1 Hyperparameter Search Space

The dimensionality and configuration of the search space for Bayesian optimization are contingent upon the dataset size, as detailed in Tables 2-4.

Table 2: Hyperparameter search spaces for ElasticNet on different datasets.

Dataset	α	l1_ratio	max_iter
Elongation	(0.001, ..., 10)	(0.1, ..., 1.0)	(100, ..., 500)
Random	(0.001, ..., 10)	(0.1, ..., 1.0)	(100, ..., 800)
Chemical	(0.001, ..., 10)	(0.1, ..., 1.0)	(100, ..., 1000)

Table 3: Hyperparameter search spaces for Random Forest on different datasets.

Dataset	n_estimators	max_depth	min_sample_split	criterion
Elongation	(50, ..., 200)	(3, ..., 20)	(2, ..., 10)	squared_error
Random	(50, ..., 1000)	(3, ..., 30)	(2, ..., 20)	squared_error
Chemical	(50, ..., 1000)	(3, ..., 30)	(2, ..., 20)	absolute_error

Table 4: Hyperparameter search spaces for Neural Networks on different datasets.

Dataset	number_layers	number_neurons	dropout_rate	learning_rate
Elongation	(1, 2, 3)	(10, ..., 200)	(0.1, ..., 0.5)	(0.001, ..., 0.1)
Random	(1, 2, 3)	(10, ..., 300)	(0.1, ..., 0.5)	(0.001, ..., 0.1)
Chemical	(1, 2, 3)	(10, ..., 500)	(0.1, ..., 0.5)	(0.001, ..., 0.1)

3.3.2 Optimization Settings

The training data was preprocessed using StandardScaler to ensure feature standardization :$z = (x - \mu)/\sigma$

Where μ denotes the mean of training samples, and σ represents the standard deviation. The optimization process followed specific settings for each model type, as shown in Table 5.

Table 5: Bayesian optimization settings.

Algorithm	Cross-validation folds	Iterations
ElasticNet	5	20
Random Forest	3	20
Neural Network	5	10

3.4 Model Trustworthiness

This dissertation focuses on global interpretability, which aims to provide an overall understanding of the model across all instances. PFI plot and the SHAP summary plot are utilized to assess model interpretability under noise conditions. The PFI plot offers a measure of the importance of each feature by observing the increase in the model's prediction error after permuting the feature's values. This allows for an understanding of the overall impact of each feature on the model's predictions.

Conversely, the SHAP summary plot provides a bird's eye view of the model by displaying the impact of all features on the model's output for every instance in the dataset. The SHAP summary plot amalgamates feature importance with feature effects. Each point on the summary plot represents a Shapley value for a feature and an instance. The color gradient signifies the value of the feature, ranging from low to high, with features arranged in order of their importance.

A marked disparity in the order or significance of features, as discerned from these plots before and after the introduction of noise to the dataset, can be indicative of the model's susceptibility to random perturbations. Such a behavior may cast doubt on the model's reliability and suggest potential overfitting or sensitivity to irrelevant features. Conversely, if the perturbations, in the form of noise, yield minimal to no alterations in the derived feature importance, it can be posited that the model possesses a commendable degree of robustness and can be deemed trustworthy in its predictive capacity.

4.Results

4.1 Elongation Dataset

4.1.1 Model Performance

In the context of the elongation dataset, hyperparameters described in Tables 6 through 8 have been used for model validation on the corresponding test dataset. The performance outcomes across different noise conditions are shown in Table 9.

Table 9: Elongation dataset model performances (R^2)

Algorithm	Noise-free	Gaussian	Perlin	Simplex
ElasticNet	0.58	0.4	0.39	0.36
Random Forest	0.63	0.46	0.48	0.55
Neural Network	0.49	0.07	0.33	0.43

From Table 9, several key observations can be made:All algorithms experience a decrease in performance with noise introduction. The Neural Network is severely impacted by Gaussian noise, with its R^2 dropping dramatically from 0.49 to 0.07. Random Forest demonstrates the most stable performance across different noise patterns. ElasticNet shows consistent degradation as noise complexity increases.

4.1.2 Global Interpretability

Feature importance analysis reveals significant changes in model interpretability under noise conditions. For the ElasticNet model, initial analysis in the absence of noise highlights “Pb”, “Si”, “Al”, “P”, and “C” as the important features (Figure 6). However, after noise incorporation, there is a perceptible alteration in the hierarchy of feature importance.

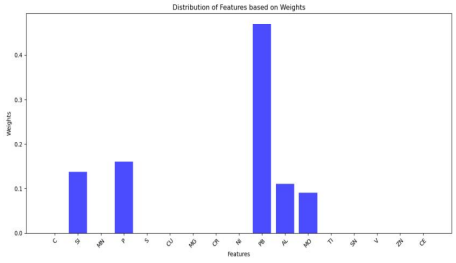


Fig. 6: Distribution of feature weights as determined from domain expertise for the elongation dataset.

The PFI plots (Figures 7-9) demonstrate that:ElasticNet model's feature importance rankings become unstable under all noise types. Random Forest shows significant changes in feature importance hierarchy. Neural Network exhibits the most dramatic shifts in feature importance patterns.

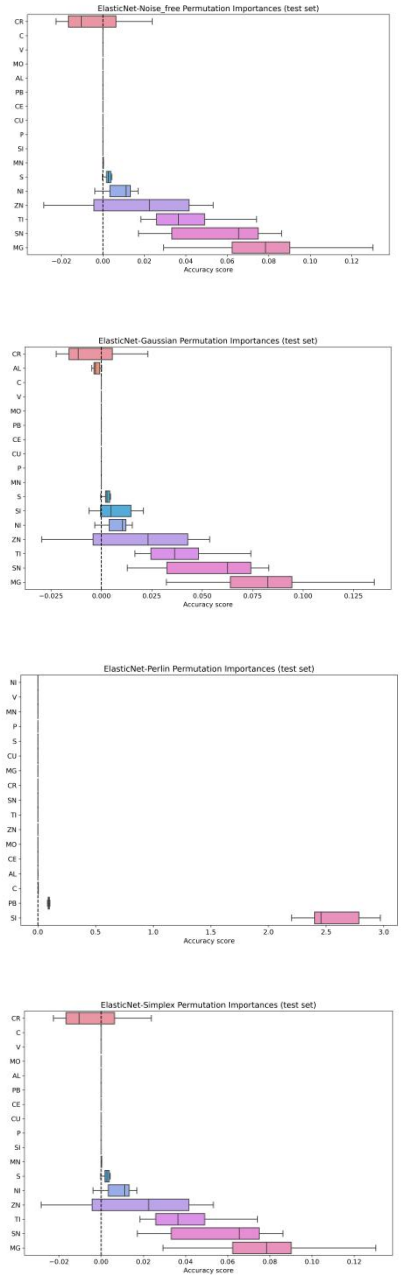


Fig. 7: PFI plot for elongation dataset of ElasticNet model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

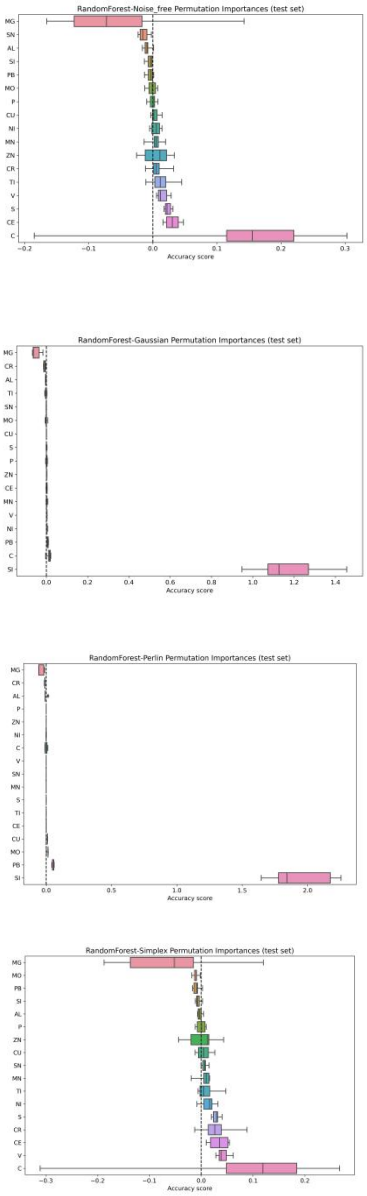
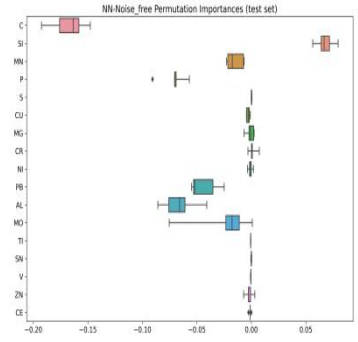


Fig. 8: PFI plot for elongation dataset of Random Forest model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.



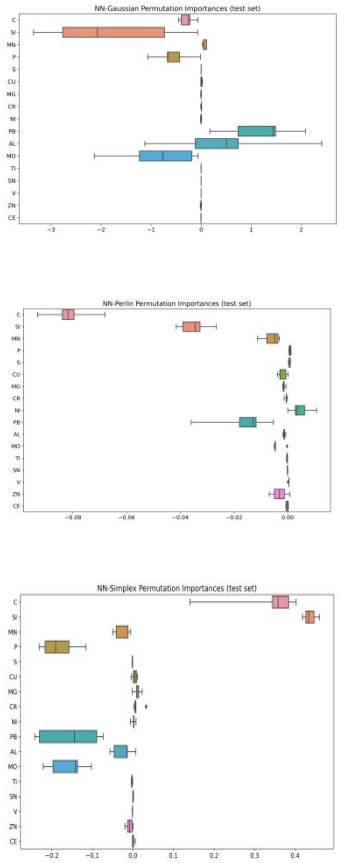


Fig. 9: PFI plot for elongation dataset of Neural Network model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

SHAP summary plots (Figures 10-12) further confirm these findings, showing that:All three algorithms change their order of feature importance after noise introduction. The feature importance rankings become inconsistent with the original domain knowledge-based pattern. The models fail to maintain stable feature attribution patterns under noise conditions.

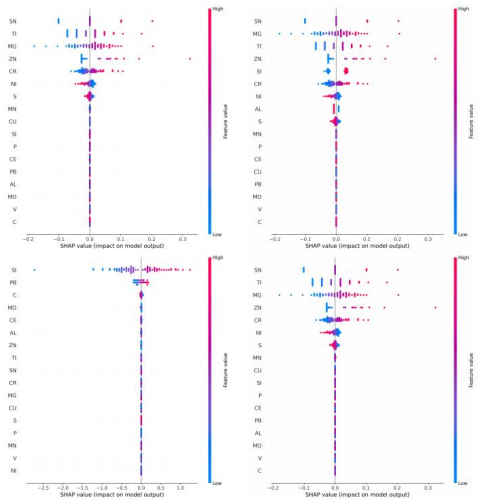


Fig. 10: SHAP summary plot for elongation dataset of ElasticNet model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex

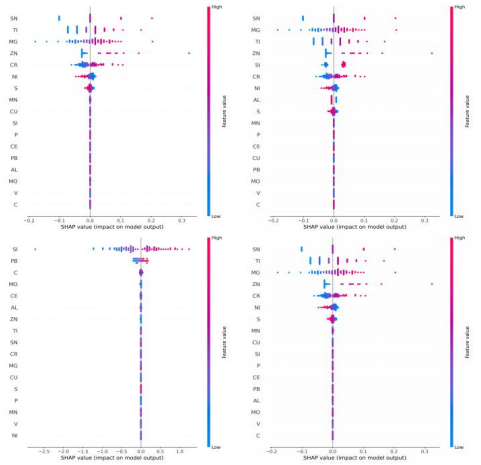


Fig. 11: SHAP summary plot for elongation dataset of Random Forest model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex

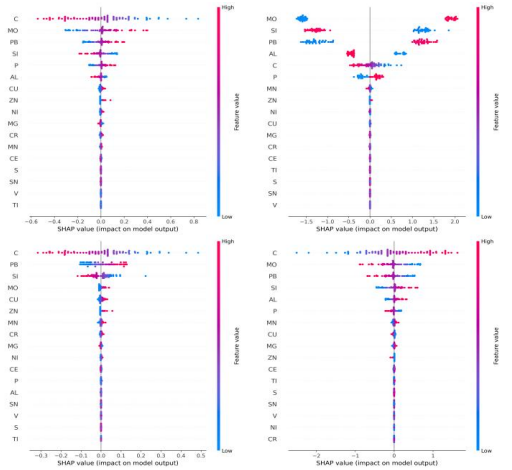


Fig. 12: SHAP summary for elongation dataset of Neural Network model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

4.1.3 Model Trustworthiness

For the elongation dataset (385 samples), upon the introduction of three distinct levels of noise, the model's efficacy markedly deteriorated. Both the PFI plots and SHAP summary plots show that, irrespective of the noise type—whether "low" level Gaussian or "high" level Simplex—the machine learning models failed to discern the underlying patterns robustly. This implies that for this relatively small dataset, the models, under the influence of these noise perturbations, are not reliable.

The exogenous noise significantly impairs both the performance and interpretability of the models, suggesting that smaller datasets may be more susceptible to noise-induced degradation in both predictive accuracy and interpretability.

4.2 Random Dataset

4.2.1 Model Performance

In the analysis of the random dataset, variations in noise patterns were systematically introduced to the raw data. The performance metrics for each model under different noise conditions are presented in Table 10

Table 10: Random dataset model performances (R^2).

Algorithm	Noise-free	Gaussian	Perlin	Simplex
ElasticNet	1	0.97	0.94	0.9
Random Forest	0.92	0.91	0.89	0.91
Neural Network	0.96	0.92	0.91	0.91

Key observations from Table 10 include: The ElasticNet model exhibits superior performance under noise-free conditions (R^2 = 1.0). All models demonstrate higher resilience to noise compared to the Elongation dataset. Performance degradation is less severe across all noise types. Random Forest shows remarkable stability with minimal performance variation

4.2.2 Global Interpretability

Both PFI and SHAP summary plots were used on the pristine, noise-free training dataset to discern the five paramount features (Figure 13). In Figure 24, the SHAP summary plot for the Random Forest model demonstrates remarkable stability - the feature importance rankings remain consistent before and after noise introduction.

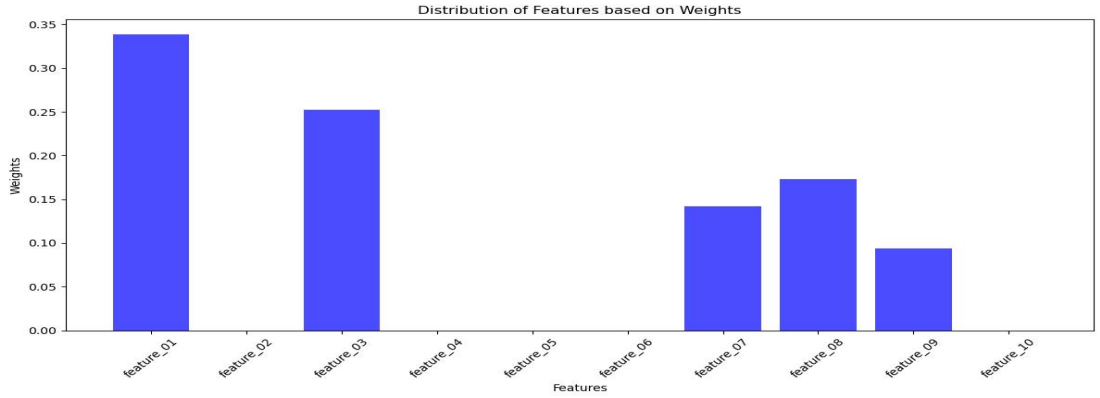


Fig. 13: Coefficient magnitudes of the quintet of informative features within the random dataset.

The evidence of model stability is further supported by:PFI plots (Figures 14-16) showing consistent feature importance rankings across all noise levels. SHAP summary plots (Figures 17) maintaining stable feature attribution patterns. Clear preservation of the original feature importance hierarchy despite noise perturbations.

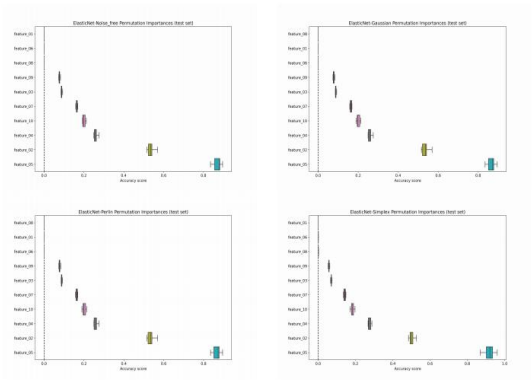


Fig. 14: PFI plot for random dataset of ElasticNet model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

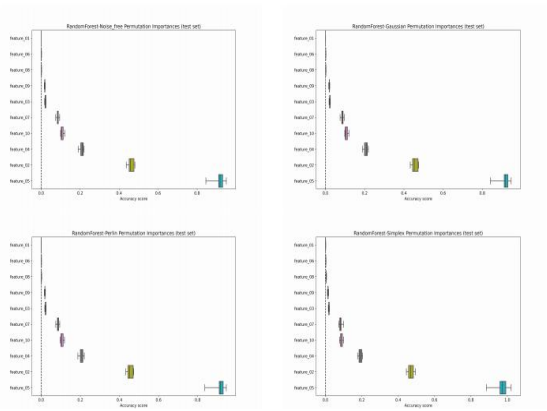


Fig. 15: PFI plot for random dataset of Random Forest model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

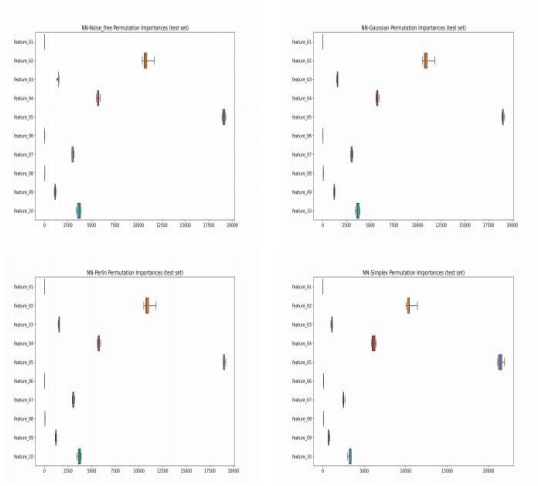


Fig. 16: PFI plot for random dataset of Neural Network model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

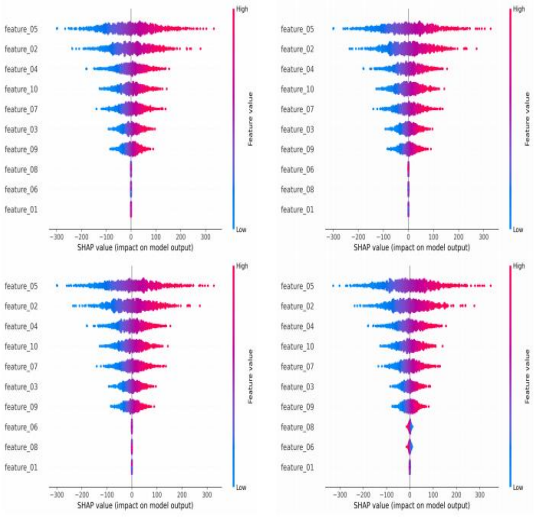


Fig. 17: SHAP summary plot for random dataset of ElasticNet model. From top left to bottom right is: Noise-free, Gaussian, Perlin and Simplex.

4.2.3 Model Trustworthiness

In the constructed random dataset (5,000 samples), five features were explicitly designed as informative. The introduction of varying noise levels significantly influences the resultant performance metrics of predictive models. An examination of both the PFI plots and SHAP summary diagrams reveals that each model adeptly identifies the inherent patterns of the dataset, preserving the original feature importance hierarchy across all noise conditions.

Based on the PFI plots and SHAP summary plots, it can be observed that all three models maintain consistent performance and interpretability under various noise conditions. ElasticNet shows minimal performance degradation from 1.0 to 0.90 even under the most severe noise, while Random Forest and Neural Network demonstrate robust stability with R^2 values consistently above 0.90. This enhanced stability, compared to the Elongation dataset, suggests that the larger dataset size contributes to improved model robustness against noise perturbations.

So it can be concluded that, for random dataset, all three models can be trusted. This conclusion is supported by their ability to maintain both performance metrics and interpretability under various noise conditions, while consistently identifying the designed important features regardless of noise level.

4.3 Chemical Dataset

4.3.1 Model Performance

In the analysis of the chemical dataset, for the noise-free scenario, ElasticNet achieved perfect performance with an R^2 of 1.0, outperforming both Random Forest and Neural Network. However, as noise types like Gaussian, Perlin and Simplex are introduced, all models exhibit a decrease in R^2, with ElasticNet still tending to outperform the other models. Notably, Neural Network consistently has the lowest R^2 across all noise types, as shown in Table 11.

Table 11: Chemical dataset model performances (R^2).

Algorithm	Noise-free	Gaussian	Perlin	Simplex
ElasticNet	1	0.96	0.93	0.94
Random Forest	0.94	0.92	0.91	0.91
Neural Network	0.86	0.82	0.84	0.83

4.3.2 Global Interpretability

In Figure 25, the SHAP summary plots pertaining to the Neural Network model applied to the chemical dataset are illustrated. It is noteworthy that the feature significance remains invariant for the Neural Network model, both pre and post the introduction of noise to the principal five features. Subsequent PFI visualizations and SHAP summary plots are shown from Figure 33 through Figure 37. Consistently across these representations, the hierarchy of feature relevance remains unaltered, underscoring that, for the chemical dataset, all models adeptly discern the intrinsic data patterns.

4.3.3 Model Trustworthiness

Upon juxtaposing the feature importance rankings across three distinct machine learning models, under the influence of varying noise intensities, the order remains invariant.

Pertaining to the chemical dataset (15,000 samples), all models adeptly discern the underlying data patterns and yield coherent interpretations.

When examining both the PFI plots and SHAP summary plots before and after noise introduction, we observe that: The hierarchy of feature importance remains consistent across all noise types. All three models maintain stable feature importance rankings. The top five features (I1, I2, I3, I4, I5) retain their relative positions regardless of noise level.

Within the context of the chemical dataset, each model exhibits reliability and robustness, demonstrating strong noise resilience in both performance metrics and interpretability measures. This enhanced stability, observed in this larger dataset, further supports the relationship between dataset size and model robustness to noise perturbations.

5.Conclusions

5.1 Summary

The comparative analysis across three datasets of varying sizes has revealed several key insights about machine learning model interpretability under noisy conditions. Our findings demonstrate a clear relationship between dataset size and model robustness to noise perturbations.

For the Elongation dataset (385 samples), all three models showed significant vulnerability to noise interference. The Neural Network proved particularly susceptible, with its R^2 dropping from 0.49 to 0.07 under Gaussian noise. More critically, both PFI and SHAP summary plots revealed that the models failed to maintain consistent feature importance rankings under noise conditions, indicating poor interpretability preservation.

In contrast, the Random dataset (5,000 samples) demonstrated markedly improved resilience. All models maintained high performance metrics even under noise conditions, with ElasticNet showing particularly strong results (R^2 = 0.97 under Gaussian noise). The feature importance hierarchies remained stable across all noise types, suggesting robust interpretability.

The Chemical dataset (15,000 samples) further confirmed this trend, with models showing strong resistance to noise perturbations. ElasticNet achieved perfect performance (R^2 = 1.0) in noise-free conditions and maintained high performance ($R^2 > 0.90$) even under complex noise patterns. Importantly, the feature importance rankings remained consistent across all noise conditions, demonstrating reliable interpretability.

Key findings from this research include:

Dataset Size Impact: Larger datasets consistently led to more robust model interpretability under noise conditions.

Model Behavior: ElasticNet showed superior performance in larger datasets. Random Forest demonstrated consistent stability across different dataset sizes. Neural Networks proved most sensitive to noise, particularly in smaller datasets

Noise Effects: Gaussian noise generally had t.he least impact on model interpretability. Complex noise patterns (Perlin and Simplex) showed stronger effects on smaller datasets. Larger datasets maintained interpretability even under high-level noise.

These findings have significant implications for practical applications of

machine learning, particularly in domains where both accuracy and interpretability are crucial. They suggest that when working with noisy data, emphasis should be placed on acquiring larger training datasets to ensure robust and trustworthy model interpretations.

5.2 Future Work

The findings from this project pave the way for several exciting research opportunities. One potential direction is the exploration of other ML algorithms, such as XGBoost and LightGBM beyond ElasticNet regression, Random Forest, and Neural Networks. This could broaden the understanding of interpretability and noise trustfulness across a more diverse range of ML models.

Furthermore, the influence of diverse noise patterns on model interpretability and efficacy warrants deeper exploration. This study scrutinized the effects of Gaussian, Perlin, and Simplex noise, yet future investigations could expand to encompass other noise varieties, such as value noise and worley noise , to provide a more comprehensive understanding.

The project could also be expanded to encompass larger datasets and more complex models. This would allow for a deeper exploration of the interplay between dataset size, model complexity, and interpretability. Furthermore, future research could focus on enhancing the noise trustfulness of ML models. This could involve the exploration of different training methods, regularization techniques, or model architectures that could improve a model's ability to handle noise.

References

[1]Hügle M, Omoumi P, van Laar J M, et al. Applied machine learning and artificial intelligence in rheumatology[J]. Rheumatology advances in practice, 2020, 4(1): rkaa005.

[2]Baduge S K, Thilakarathna S, Perera J S, et al. Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications[J]. Automation in Construction, 2022, 141: 104440.

[3]Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective[J]. Artificial Intelligence in medicine, 2001, 23(1): 89-109.

[4]Murdoch W J, Singh C, Kumbier K, et al. Definitions, methods, and applications in interpretable machine learning[J]. Proceedings of the National Academy of Sciences, 2019, 116(44): 22071-22080.

[5]Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models[J]. ACM computing surveys (CSUR), 2018, 51(5): 1-42.

[6]Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint arXiv:1702.08608, 2017.

[7]Gilpin L H, Bau D, Yuan B Z, et al. Explaining explanations: An overview of interpretability of machine learning[C]//2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018: 80-89.

[8]Lipton Z C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16(3): 31-57.

[9]Scott M, Su-In L. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30: 4765-4774.

[10]Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks[J]. arXiv preprint arXiv:1711.06104, 2017.

[11]Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning[J]. Advances in neural information processing systems, 2016, 29.

[12]Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon[J]. Information, communication & society, 2012, 15(5): 662-679.

[13]Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems[C]//2016 IEEE symposium on security and privacy (SP). IEEE, 2016: 598-617.

[14]Leo Breiman. Random forests. Machine learning, 45:5–32, 2001.

[15]Thomas G Dietterich et al. Ensemble learning. The handbook of brain theory and neural networks, 2(1):110–125, 2002.

[16]Ken Perlin. An image synthesizer. ACM Siggraph Computer Graphics, 19(3):287– 296, 1985

[17]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[18]Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599, 2010.

[19]Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[20]Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 2017.

[21]Fischer Black. Noise. The journal of finance, 41(3):528–543, 1986.

[22]Steven Worley. A cellular texture basis function. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 291–294, 1996.

Author information： Tong Zhen Hao Huazhong University of Science and Technology Education:Undergraduate study Research direction: Computer vision and big language mode