Efficient XAI: A Low-Cost Data Reduction Approach to SHAP Interpretability

SEVERIN BACHMANN*, Nuremberg Research Institute for Cooperative Studies, Germany

Explainable Artificial Intelligence (XAI) has become a critical area of research, particularly in ensuring transparency and trustworthiness in machine learning (ML) models. In this context SHAP (SHapley Additive exPlanations) is widely recognized as a robust method for feature attribution, yet its computational cost poses significant challenges, especially for large datasets. This study explores a novel approach to optimizing SHAP computations by leveraging Slovin's formula, a statistical sampling technique traditionally used in survey research. Unlike feature selection or dimensionality reduction methods, Slovin's formula requires minimal prior knowledge of the dataset's statistical properties while providing an efficient, heuristic-based alternative for data reduction. It offers a straightforward, low-cost sampling approach that can be applied without extensive preprocessing, making it accessible for computationally constrained environments. Through controlled experiments on synthetic datasets, we analyze the stability of SHAP values under Slovin-based subsampling across varying data characteristics, including feature and target types and distributions, and dataset sizes. Our findings reveal a U-shaped trade-off: SHAP values for mid-ranked features remain stable, whereas extreme values exhibit higher fluctuations. Additionally, categorical and non-skewed distributed features maintain greater robustness, while highly skewed target distributions introduce variability. Importantly, the effectiveness of Slovin's formula diminishes when the subsample-to-sample ratio falls below 5%. By integrating Slovin's formula into SHAP workflows, we demonstrate a practical solution for balancing interpretability and computational efficiency in machine learning. This method reduces processing costs while retaining key feature attributions, making it particularly valuable for researchers and practitioners working with resource-constrained environments. Our study contributes to the broader discourse on sustainable AI, offering a scalable and interpretable framework for advancing explainability in modern machine learning systems.

JAIR Associate Editor: Roberta Calegari

JAIR Reference Format:
Severin Bachmann. 2025. Efficient XAI: A Low-Cost Data Reduction Approach to SHAP Interpretability. *Journal of Artificial Intelligence Research* 83, Article 2 (June 2025), 21 pages. DOI: 10.1613/jair.1.18325

1 Introduction

Artificial intelligence (AI) has emerged as one of the most transformative forces in modern technology, revolutionizing industries ranging from healthcare and finance to autonomous systems and scientific research. However, as AI models grow in complexity and capability, they also demand ever-increasing amounts of computational power and energy. This rapid escalation of resource consumption has reached a critical juncture, as exemplified by Microsoft's recent decision to reactivate an entire nuclear power plant solely to sustain its AI operations [30]. The sheer scale of this investment highlights the immense energy requirements of state-of-the-art machine learning (ML) systems and underscores the urgent need for resource-efficient AI solutions [13, 60].

A critical aspect of this challenge is to ensure interpretability in machine learning models while maintaining computational feasibility. The increasing complexity of AI systems has made them more opaque, raising concerns

*Corresponding Author.

Author's Contact Information: Severin Bachmann, ORCID: 0000-0001-5996-4152, severin.bachmann@gmx.net, Nuremberg Research Institute for Cooperative Studies, Nuremberg, Bavaria, Germany.

DOI: 10.1613/jair.1.18325

about trust, fairness, and accountability [5]. The opacity of black-box models poses critical challenges to the safe and ethical deployment of AI systems, particularly in high-stakes environments such as healthcare, finance, and justice. Without the ability to trace how decisions are made, professionals cannot be held accountable nor can they confidently trust the outcomes. This has driven not only academic but also regulatory efforts to demand greater transparency and interpretability in AI systems. A compelling example is found in the European In Vitro Diagnostic Regulation (IVDR), which explicitly recognizes software including AI algorithms as medical devices, thus subjecting them to strict requirements for traceability and transparency. This regulation mandates that professionals must be able to understand and justify decisions made with the aid of AI systems, underscoring the need for explainability, interpretability, and causability in AI models. As Müller et al. (2022) argue, explainable AI is not only a scientific challenge, but increasingly a legal and ethical necessity in regulated domains.

Explainable AI (XAI) seeks to address this issue by developing methods that allow users to understand and interpret model decisions [51]. It has evolved into a dynamic and well-established field that addresses the pressing need for transparency, accountability, and trust in machine learning systems. A wide variety of approaches have emerged, including inherently interpretable models such as decision trees and rule-based systems, as well as post-hoc explanation techniques that can be applied to any model. Among the most prominent of the latter are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

While SHAP has gained considerable traction for its theoretical rigor and practical utility in feature attribution, it is just one of many viable tools within the explainability toolkit. For a comprehensive and current overview of XAI techniques including comparative discussions on their applicability across domains readers are referred to Bennetot et al. (2024), who offer an extensive survey of the field's development and practical challenges. Despite its effectiveness, SHAP is computationally expensive, particularly for large-scale datasets. Calculating them requires generating multiple model permutations, making it an impractical solution for many real-world applications [64].

This study explores an alternative approach to balancing computational efficiency and interpretability by applying Slovin's formula, a statistical sampling method traditionally used in survey research, to SHAP-based machine learning interpretability. Unlike feature selection or dimensionality reduction techniques that explicitly alter dataset structures, Slovin's formula provides a straightforward, low-cost sampling approach that can be applied without extensive preprocessing, making it accessible for computationally constrained environments [44]. By reducing the volume of data used in SHAP calculations, it becomes possible to alleviate computational burdens while preserving the interpretability of machine learning models under certain conditions.

The results of this study reveal that Slovin's formula offers a promising trade-off between computational efficiency and interpretability. Across various experimental conditions, the findings indicate that as the relative size of a feature's SHAP value increases, its stability after subsampling improves with an optimum for middle ranked SHAP values resulting in a U-shaped trade-off. Features with strong correlations to the target variable tend to exhibit greater fluctuations, whereas weak and moderate correlated features remain more stable, where feature correlation is a proxy for model performance. Additionally, datasets with normally or little-skewed distributed features and target variables benefit from Slovin's method, whereas highly skewed distributions introduce more variation. The same holds for feature and target types, as dummy types show less variation than continuous variables. The study also highlights the importance of dataset size in determining the effectiveness of Slovin's formula. While it proves to be a viable strategy for small and medium-sized datasets, its reliability diminishes when the subsample-to-sample ratio drops below 5%, particularly in large datasets exceeding 100,000 observations.

Despite variability in SHAP values, particularly for lower-ranked features, Slovin's formula emerges as a viable tool for computational optimization in machine learning interpretability under certain conditions. This study highlights the conditions to be met. While it does not eliminate all computational challenges associated with SHAP, it provides a structured and resource-friendly alternative that can significantly reduce processing costs.

These findings contribute to the broader discourse on sustainable AI, emphasizing that even in the absence of quantum computing, practical solutions exist to make AI more efficient and accessible. By integrating data reduction techniques such as Slovin's formula into explainability frameworks, AI researchers and practitioners can develop models that are not only powerful and interpretable but also computationally and environmentally sustainable.

The paper's structure is as follows. After reviewing relevant literature on machine learning, explainability, and computational constraints, the methodology section details the experimental framework. The data section then outlines the generation of synthetic datasets in detail. The results section systematically presents findings on SHAP stability across varying dataset characteristics after implementation of Slovin's formula.

2 Literature Review

The evolution of ML has been marked by rapid innovation and increasing complexity, necessitating a growing emphasis on interpretability, scalability, and computational feasibility. In response to concerns around the opaque nature of black-box models and the rising resource demands of advanced algorithms, the field of XAI has emerged as a critical area of inquiry. This chapter reviews the conceptual foundations and current research in machine learning, interpretability, computational constraints, and data reduction strategies culminating in a focused discussion on Slovin's formula as a low-cost alternative in the XAI space.

2.1 Machine Learning and Black Box Explaining

Machine Learning has undergone a transformative evolution, starting from its theoretical roots and advancing into a powerful toolkit for solving complex real-world problems. The journey of ML began with foundational work by researchers such as Alan Turing, who speculated about machines learning from data [63]. This was followed by the creation of early models like the Perceptron, developed by Frank Rosenblatt in 1958, which served as one of the first algorithms for binary classification tasks. Despite their advances, the field struggled during the "AI winter," a period marked by computational and theoretical challenges, including the inability of single-layer perceptrons to handle non-linear problems [40].

The 1980s and 1990s saw a revival of ML research with the development of the backpropagation algorithm, enabling effective training of multi-layer neural networks [55]. Alternative methods like support vector machines [19] and decision trees [41, 48] emerged, enhancing predictive capabilities across tasks. This era also introduced ensemble methods such as random forests [11] and boosting [57, 12], alongside advances in statistical learning theory [67].

The early 21st century witnessed a major leap with deep learning's rise, fueled by GPU's computational power [32]. Innovations like convolutional neural networks and long short-term memory networks transformed image classification and sequence data modeling, respectively [32, 26]. Recent years have seen natural language processing revolutionized by transformer-based architectures [68], leading to models like GPT and BERT that excel in text generation and comprehension. As ML has advanced, its applications have proliferated across diverse fields, such as financial forecasting, healthcare diagnostics, and environmental modeling [3, 66, 21]. Comparative studies across industries consistently reveal that ML methods outperform linear approaches in capturing nuanced patterns in data.

Despite their effectiveness, the "black box" nature of many ML models has sparked concerns about transparency and accountability. According to Athey and Imbens (2019) the emphasis in the ML literature has primarily been on evaluating out-of-sample performance, neglecting a traditional focus of the statistics and econometrics literatures, the capacity for inference. Addressing these concerns, researchers have shifted focus from solely optimizing predictive performance to ensuring models are interpretable, addressing the need to understand how input features contribute to predictions and improve decision-making [17, 56].

This has led to the evolution of Explainable Artificial Intelligence. A foundational distinction in the XAI field is between ante-hoc and post-hoc interpretability methods. Ante-hoc models such as rule-based expert systems and decision trees [20] are inherently interpretable by design, meaning their internal logic can be directly examined and understood. In contrast, post-hoc methods aim to explain complex, black-box models after training by approximating or attributing feature importance. This distinction is critical when evaluating transparency needs versus model performance trade-offs. A recent comprehensive study by Retzlaff et al. (2024) provides a comparative framework between these approaches and introduces design guidelines to help practitioners choose suitable explainability methods for their tasks.

The rapid growth of complex models like deep neural networks made traditional interpretability methods insufficient, spurring the development of novel post-hoc XAI techniques. Among these, model-agnostic methods such as LIME [53] and SHAP [39] emerged as powerful tools for generating post-hoc explanations [14].

SHAP values, based on cooperative game theory [58], offer a particularly strong approach to model interpretability. Unlike LIME that only provides heuristic or locally valid explanations, SHAP values attribute the contribution of each feature to the model's prediction in a theoretically sound manner, ensuring consistency and additivity [39, 61]. This makes SHAP highly advantageous for understanding complex models, as it offers a global view of feature importance while also providing individual-level explanations. This versatility has led to widespread adoption of SHAP in fields such as finance, healthcare, and automated decision systems, where understanding feature contributions is essential for regulatory compliance and user trust [33, 69, 4, 2, 46].

2.2 Computational Constraints

Despite advances, challenges remain. The SHAP method is computationally intensive, especially for high-dimensional datasets and complex models, creating barriers to its practical application [39]. The exponential scaling of SHAP calculations with the number of features leads to a "combinatorial explosion" that hinders scalability [16]. Van den Broeck et al. (2022) analyze the computational constraints of SHAP and propose tractability optimizations, yet these solutions do not fully address scalability concerns in large-scale datasets, which we tackle using Slovin's formula. This issue is particularly problematic for real-time systems or large datasets, where the computational overhead of SHAP values makes their use impractical. For example, tasks such as model training and inference in state-of-the-art models like GPT-3 require vast computational resources, posing significant challenges for smaller organizations [13]. Studies on the energy consumption of NLP models highlight these constraints, further emphasizing the need for efficient solutions [60].

To address these challenges, researchers have explored multiple strategies. Hybrid and approximation methods are techniques, which combine methods like Permutation Feature Importance (PFI) with SHAP, or use Accumulated Local Effects (ALE) plots, in order to reduce computational demands while preserving interpretability [31]. Dimensionality reduction and feature selection techniques such as Principal Component Analysis (PCA) and model pruning effectively streamline computational tasks [28, 24]. Distributed and scalable computing frameworks like Apache Spark parallelize tasks, enabling the handling of large datasets and complex models [22]. These frameworks, combined with energy-efficient hardware innovations, offer promising solutions to scalability challenges [34].

2.3 Data Reduction

Among these strategies, data reduction methods stand out as particularly promising for addressing computational constraints. By systematically reducing the volume of data, these approaches mitigate the resource demands of high-dimensional datasets, ensuring computational feasibility without compromising model interpretability [25, 65]. Data reduction techniques sometimes include dimensionality reduction methods, feature selection in

addition to data sampling, and data compression. In order to sharply distinguish our field of study from feature selection and dimensionality reduction, we exclude those two from the field of data reduction.

While these techniques can lead to significant benefits, applying them effectively is complex and time-consuming for several reasons.

Nature and Complexity of the Data: Data often contains many features and samples with potential noise, missing values, or redundancies. High-dimensional data is particularly challenging, as described by Bellman (1961) in the concept of the "curse of dimensionality," which indicates that the computational and analytical complexity of data increases exponentially with the number of features. Reducing dimensionality can lead to the loss of potentially valuable information if not done carefully. PCA and t-distributed Stochastic Neighbor Embedding (t-SNE) are popular dimensionality reduction techniques but selecting an appropriate number of components and interpreting results requires expertise and can lead to suboptimal reductions [65].

Selection and Optimization Challenges: Selecting the most appropriate data reduction technique is not straightforward and often depends on the specific characteristics of the dataset and the intended use case. Feature selection methods, such as wrapper, filter, and embedded approaches, need to be carefully tailored to the model and problem domain [25]. The selection process can involve exhaustive search, requiring significant computational resources, especially for large datasets. Even methods like PCA require fine-tuning, such as deciding on the number of principal components to retain [29].

Preservation of Data Integrity: Reducing the dimensionality or volume of data must be done without compromising data integrity and interpretability. Inadequate application of reduction methods may lead to biased or misleading insights, as highlighted by Chandrashekar and Sahin (2014). Moreover, data transformation techniques, such as scaling and normalization, often need to be performed before reduction to ensure meaningful results. Preserving relationships between features and the underlying data structure adds an additional layer of complexity to the process.

In summary, applying data reduction techniques requires an in-depth understanding of statistical and mathematical principles, familiarity with domain-specific knowledge, and skillful use of tools. This complexity often makes it challenging for non-experts to implement data reduction without making mistakes, thereby necessitating substantial training and experience [23].

2.4 Slovin Formula

Slovin's formula is a widely recognized method to estimate an appropriate sample size for surveys or research studies. It is particularly useful when the population size (N) is known, but other statistical parameters such as variance or standard deviation are unavailable or difficult to determine. The formula provides an approximation of the sample size required to achieve a specified level of precision (e) in estimates, especially when researchers aim to draw inferences about a population while operating under resource constraints. The formula is:

$$n = \frac{N}{1 + Ne^2} \tag{1}$$

where:

- n is the required sample size,
- N is the population size,
- e is the margin of error (expressed as a decimal)

Slovin's formula is frequently discussed in statistical and educational literature [44, 47], although its precise origins remain unclear. The formula is often cited as a heuristic tool rather than a rigorously derived statistical theorem. It is closely related to early discussions on sample size estimation techniques outlined by Cochran (1977)

and similar works, which emphasize simplified calculations for practitioners without advanced statistical training [35].

While the formula lacks a concrete attribution, its adoption in fields like business research, education, and social sciences can be attributed to its user-friendliness [50]. It is favored for its straightforward implementation, which eliminates the need for complex statistical computations. Designed for rapid calculation, Slovin's formula is especially useful in exploratory studies where precision is secondary [1]. It is also well suited for initial inquiries where detailed precision is unnecessary [36], researchers constrained by time, budget, or logistical factors, when variance or other population parameters are unavailable [27]. It is also appropriate for populations with minimal variability (homogeneous populations) and most effective when each individual has an equal probability of selection.

2.5 Why it is Reasonable to Engage with Slovin's Formula

The core critique against Slovin's formula is its reliance on simplifying assumptions. Tejada and Punzalan (2012) argue that Slovin's formula, derived under strict assumptions, is narrowly applicable to specific scenarios, such as estimating population proportions with a 95% confidence level and a maximum heterogeneity of 0.5. They suggest that these limitations make the formula unsuitable for inferential problems beyond estimating proportions. They further argue that its simplicity leads to widespread misuse, particularly in cases where researchers fail to justify its application or consider its underlying assumptions. While these criticisms are valid in scenarios where the formula is applied uncritically, they do not categorically invalidate its use. In fact, Slovin's formula can address specific research challenges effectively, as demonstrated by Bachmann (2025). In the study Slovin's formula is employed as a data reduction technique to tackle computational constraints in calculating SHAP values for over 0.5 million observations. By applying the formula, the researcher reduces the necessary test dataset from 100,000 to 2,946 observations, enabling them to conduct SHAP-based feature importance analysis without overburdening computational resources. Crucially, the reduction does not compromise the interpretability of the models, as the deviation in SHAP values between the reduced and full datasets is negligible. This innovative application demonstrates how Slovin's formula can be adapted to specific research contexts, effectively challenging the assertion that it is inherently flawed.

Moreover, the researcher takes care to justify his use of Slovin's formula. He explicitly defines the margin of error (e=0.02) and validates the representativeness of the reduced sample. This transparency stands in stark contrast to the uncritical applications criticized by Tejada and Punzalan (2012). By integrating Slovin's formula into SHAP computation, it would become possible to achieve a balance between two seemingly contradictory goals: computational efficiency and the preservation of interpretability, a trade-off that Radford et al. (2021) already engage in, and which addresses the scalability concerns cited by Linardatos et al. (2020)

This motivation sets the stage for the subsequent exploration of Slovin's formula's practical application in reducing SHAP's computational costs. The chapter to follow delves into implementation strategies, evaluating their effectiveness across diverse datasets and models. By situating Slovin's formula within the broader context of XAI challenges, we underscore its potential to redefine the balance between scalability and interpretability in modern machine learning systems.

3 Method

To explore the practical integration of Slovin's formula into SHAP-based machine learning interpretability, we develop a comprehensive experimental framework grounded in synthetic data generation and model evaluation. This chapter details the methodology, including model selection, data sampling strategies, and the analytical approach used to assess SHAP value stability. By systematically varying dataset characteristics, we aim to identify

the conditions under which Slovin's formula offers a viable trade-off between computational efficiency and interpretability.

3.1 Overall Approach

In response to the findings and assertions presented inTejada and Punzalan's (2012) paper, we aim to systematically challenge their conclusions by demonstrating a novel use-case of Slovin's formula in machine learning model training and evaluation. While the authors contend that Slovin's formula is often misapplied, our approach is designed to explore its utility and limitations in a controlled and scientifically rigorous manner using synthetic data and modern computational techniques. We will outline our methodology and steps as illustrated in Figure 1.

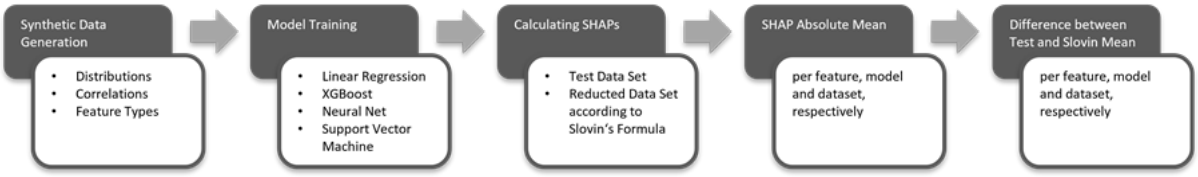

Fig. 1. Methodology overview

The process begins with **synthetic data generation**, where artificial datasets are constructed to simulate various data scenarios. This involves specifying the distributions of individual features, defining correlations between them, and considering different feature types, such as numerical and categorical variables. By carefully designing these characteristics, the generated data can reflect complex patterns that are representative of real-world phenomena.

Once the synthetic data is prepared, it is used to **train several machine learning models**. The selected models include linear regression (LR), extreme gradient boosting (XGB), neural networks (NN), and support vector machines (SVM). Each of these models offers unique advantages, ranging from the simplicity and interpretability of LR to the sophisticated capabilities of NN and XGB for capturing non-linear relationships and interactions. This range covers the common models applied in ML contexts.

Following model training, we **calculate SHAP values** to interpret the models. SHAP values are computed on two datasets: the full test dataset and a reduced dataset created using Slovin's formula. These SHAP values quantify the contribution of each feature to the model's predictions for every observation. To make all values practicable for comparison, the **absolute mean of SHAP** is computed for each feature on every model for every dataset (further detail in the result section). This step gives us the average magnitude of feature contributions, enabling comparisons across different scenarios. Finally, the **difference in SHAP absolute mean** values between the test dataset and the reduced dataset is calculated, again for each feature on every model for every dataset.

3.2 SHAP Values as Comparing Tool

To strengthen our proposed methodology for challenging the application of Slovin's formula in data reduction context, it is important to emphasize that the application of SHAP values as a tool of model comparison is well-grounded in existing literature.

Baptista et al. (2022) explore the correlation between SHAP values and classical metrics such as monotonicity, trendability, and prognosability in prognostic models. This demonstrates how SHAP values can track and reflect changes in model outputs, emphasizing the importance of SHAP-based explanations for understanding how features contribute to complex predictions between models. Yang et al. (2024) use SHAP values to analyze the contribution of different indicators in landslide susceptibility models, revealing that SHAP can effectively highlight feature importance across models of varying complexity and robustness.

Recent work by Nguyen et al. (2023) investigates the application of SHAP values in financial risk modeling, demonstrating their efficacy in identifying high-impact features that influence credit default predictions. By providing granular insights into feature contributions, SHAP values have allowed financial institutions to meet regulatory requirements for transparency while optimizing risk assessment models. Similarly, Liu et al. (2023) utilizes SHAP values to evaluate feature importance in climate prediction models, showcasing how local explanations align with global climate trends, ensuring both interpretability and predictive robustness.

In healthcare, Patel (2023) apply SHAP values to electronic health record data for early detection of sepsis. Their findings illustrate that SHAP explanations facilitate clinician trust by revealing feature contributions such as vital signs and lab results across various databases. Furthermore, an industrial application by Batouei (2024) demonstrates how SHAP can be used for sensitivity analysis in industrial autoclave curing processes, particularly in balancing trade-offs in cure uniformity.

These studies collectively support the validity and utility of using SHAP values to compare and interpret model behavior across different datasets and models.

3.3 Relative vs. Absolute Values

All papers mentioned in the previous section use absolute SHAP values in their comparison task. In scientific research and data analysis, the choice between relative and absolute values often shapes the interpretability and applicability of results. While absolute values provide raw magnitudes, relative values contextualize these magnitudes within a framework that often reflects real-world relationships more accurately. This section argues that the use of relative values is not only plausible but also scientifically grounded for our approach. To the best of our knowledge, we are the first to apply relative values in context of SHAP comparison.

Relative values adjust for scale and variance, enabling comparisons that are meaningful across datasets and contexts. For instance, the relative importance of SHAP values provides insights into proportional contributions rather than sheer magnitudes [39]. This approach aligns with the principle of interpretability emphasized in explainable AI research, where the goal is to understand the impact of a feature relative to others, irrespective of absolute numerical values [17]. Moreover, relative values are inherently robust in settings where absolute measures can be misleading due to heterogeneity in datasets. For example, in healthcare, relative risk ratios are favored over absolute counts to assess treatment efficacy because they offer standardized metrics, facilitating cross-study comparisons and meta-analyses [59].

Also, relative metrics excel in comparative studies. In environmental modeling, for instance, researchers often examine relative changes in pollution levels rather than absolute measurements to account for seasonal variations and differing baselines across regions [3]. Absolute values may inadvertently introduce biases, particularly in high-dimensional datasets with disparate feature distributions. By contrast, relative importance measures normalize these effects. Baptista et al. (2022) demonstrate that relative metrics mitigate overrepresentation of features with naturally high scales, thereby preserving fairness and ensuring interpretability in predictive maintenance systems.

The preference for relative values is not arbitrary but is underpinned by robust theoretical frameworks. Ratio analysis, a cornerstone in statistical and economic research, exemplifies how relative metrics provide deeper insights into proportional relationships [70]. Similarly, in dimensionality reduction, relative error metrics are used to evaluate the quality of transformations, ensuring that critical data patterns are preserved even when absolute values are altered [65].

4 Data

This chapter explores the process of generating synthetic datasets designed for our experiments. The methodologies follow a structured framework aiming to create data with specific statistical and relational properties,

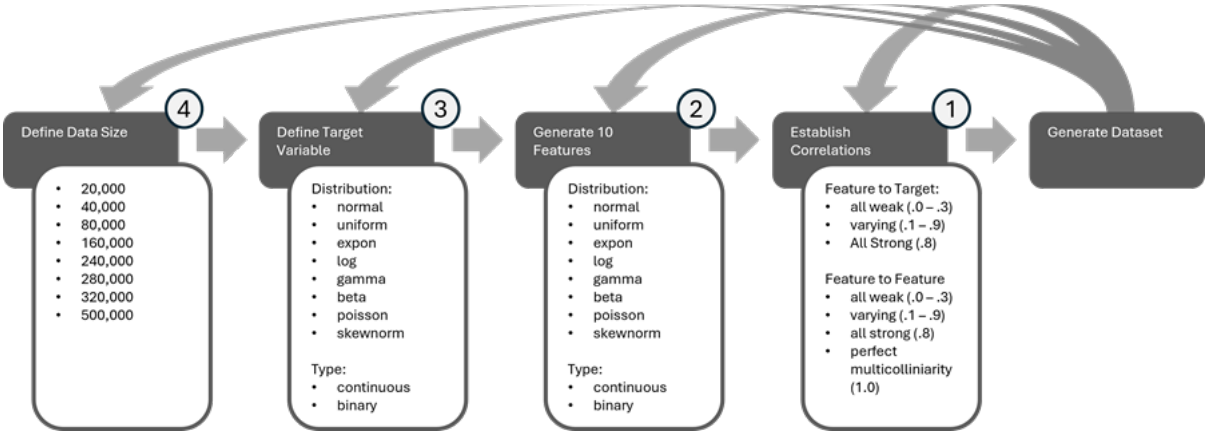

Fig. 2. Iterative data generation process and result presenting overview

ensuring flexibility and adaptability for a range of research scenarios. The primary focus is on developing datasets with controlled correlations between features and target variables, while also introducing diverse transformations and distributional variations. These datasets are designed to mimic real-world scenarios while maintaining full control over the data-generating process, allowing us to systematically explore the application of Slovin's formula on SHAP values under various conditions. Figure 2 illustrates the iterative creation process.

The methodology begins with the **definition of the dataset size**, which establishes the number of observations. The dataset size can be selected from a range of options, starting from 20,000 and extending to 500,000 observations[1]. This step is primarily driven by the fact that subsampling with Slovin results in diminishing subsample-sample ratio. While it is analytically evident from Slovin's formula that $\lim_{N\to\infty} n = \frac{1}{e^2}$, which equals 2,500 for a margin of error $e = 0.02$, the derivative $\frac{dn}{dN} = \frac{1}{(1+Ne^2)^2}$ confirms that n increases monotonically with decreasing marginal gains, and $\frac{d^2n}{dN^2} = \frac{2e^2}{(1+Ne^2)^3} < 0$ further confirms concavity, Table 1 illustrates the exact numbers for our datasets.

The first column (Sample Size) represents the total number of elements in the dataset, ranging from 4,000 to 100,000. The numbers represent the test dataset sizes of the total dataset sizes illustrated in Figure 2 and they correspond to them being the 20%-fraction. The second column (Subsample Size) indicates the number of elements extracted as a subsample from the larger dataset after applying Slovin under the assumption of error rate 0.02, which is a well-known value in literature (Harfitalia & Pujangkoro, 2022). The third column (Subsample-Sample Ratio) expresses the proportion of the subsample size to the total sample size, calculated as:

$$\text{Ratio (\%)} = \left(\frac{\text{Subsample Size}}{\text{Sample Size}}\right) \times 100 \tag{2}$$

The data concretely show that as the sample size increases, the corresponding subsample size grows at a diminishing rate. For instance, for a sample size of 4,000, the subsample size is 1,538, yielding a ratio of 38.46%. However, for the largest sample size of 100,000, the subsample size increases to only 2,439, with a significantly lower ratio of 2.44%. This trend indicates that the relative size of the subsample decreases as the dataset becomes

[1]The choice of sample sizes is mainly driven by practical forces. We start of with 20,000 because of the manageable time consumption of SHAP value calculations. From there we make a large step to 500,000 to check the representativeness of the 20,000-results. Seeing the results drastically drifting off we follow a step-by-step approach to find the trade-off data size, where SHAP value variation exceeds beyond senseful results.

Table 1. Overview of how the subsample size and its ratio to the total sample change with increasing dataset size.

Sample Size	Subsample Size	Subsample-Sample Ratio
4000	1538	38,46%
8000	1905	23,81%
16000	2162	13,51%
32000	2319	7,25%
48000	2376	4,95%
56000	2393	4,27%
64000	2406	3,76%
100000	2439	2,44%

larger and asymptotically merges to the value of 2,500 (in case of constant error rate of 0.02). While such behavior is helpful for securing computational feasibility, it increases the risk of resulting in non-representative subsamples.

In the second step, we define the **target variable**, serving as the dependent variable in the dataset. This variable can be generated from a wide range of statistical distributions, including normal, uniform, exponential, logarithmic, gamma, beta, poisson, and skewnormal distributions. Each distribution provides unique characteristics, allowing us to model various real-world phenomena. Furthermore, the target variable can be specified as either continuous or binary, in order to simulate datasets for regression tasks and classification problems. The third step involves the **generation of ten**[2] **features**. These features, like the target variable, can be sampled from a broad selection of statistical distributions. This mimics a variety of data characteristics, ranging from symmetric to skewed distributions and from continuous to categorical data. Another set of datasets incorporated features with mixed distributions, such as bimodal and skewed patterns. These datasets tested the adaptability of machine learning models to non-standard data distributions. The fourth step focuses on establishing correlations between the features and the target variable, as well as among the features themselves. This step is critical for creating datasets with realistic interdependencies and for testing under specific conditions. We define the correlation *strength* between the characteristics and the targets at three levels: all weak ($|r| \leq 0.3$), varying ($|r| \in [0.0, 0.9]$), and all strong ($|r| \geq 0.8$). While feature–target correlations were generally constructed to be positive for consistency, inter-feature correlations were not constrained in sign, allowing for both positive and negative relationships to emerge.

Datasets with strongly correlated features and targets serve as a foundation for studying regression and feature importance techniques. These datasets provide clear predictive relationships, enabling robust model training and evaluation. Datasets with weak or negligible correlations challenge models to handle noise effectively. These scenarios highlight the robustness results in the absence of strong signals. Similarly, the correlations among features can be set to several configurations, including **all weak, varying, all strong, or perfect multicollinearity** (a inter feature correlation of 1.0). By inducing multicollinearity—where some features are linear combinations or transformations of others—these datasets provide a testbed for studying the limitations of regression models and the efficacy of feature selection methods.

The datasets are created by systematically altering one parameter at a time while keeping others constant. This methodical variation allows for the isolation and examination of the impact of each parameter on the robustness of SHAP values across dataset sizes.

[2]Clearly choosing ten as the number of features is arbitrary. Real life datasets can reach up to hundreds of features. Since this is a first-time approach of applying Slovin formula we don't want to overcomplicate. The number of ten is a good first shot as it allows for modelling necessary interdependencies between features and for supplying various feature characteristics within one dataset, but on the other hand maintains clarity. Examining SHAP values changes for varying number of features is an interesting topic of further research.

5 Results

We will present the results of the research step-by-step, following the well-known sequence depicted in Figure 2, this time focusing on the stages marked by number 1 to 4. Each stage corresponds to one of the specific aspects of the data generated. We start by looking at conspicuities with SHAP value differences emerging from variation of correlations (1) going on to eminences based on variation in feature type and distribution (2), to target type and distribution (3) and, finally, to data size (4).

The average absolute SHAP values are computed and compared across full and Slovin-reduced datasets, revealing how different dataset attributes influence feature importance measures. As argued in the method section we choose to compare relative measures with each other. Hence, we put absolute mean SHAP values for each feature within each iteration into relation of all remaining SHAP means resulting in a value between 0 and 1 for each feature in every model for both, full dataset and reduced as illustrated in formula 3.

$$\text{Relative SHAP}_{i,j,k} = \frac{\text{SHAP}_{i,j,k}}{\text{SHAP}_{1,j,k} + (\cdots) + \text{SHAP}_{n,j,k}} \tag{3}$$

$\text{SHAP}_{i,j,k}$ is the absolute SHAP mean for feature i in model j trained on dataset k. The formula basically tells us that we calculate a relative SHAP mean for each feature by dividing its absolute SHAP mean by the sum of all absolute SHAP means within the same model j and dataset k.

Additionally, we calculate the relative difference between the relative SHAP mean based on the original test dataset ($\text{Relative SHAP}_{i,j,k}$) and the reduced dataset ($\text{Reduced Relative SHAP}_{i,j,k}$), as depicted in formula 4.

$$\text{Difference}_{i,j,k} = \frac{\text{abs}(\text{Reduced Relative SHAP}_{i,j,k} - \text{Relative SHAP}_{i,j,k})}{\text{Relative SHAP}_{i,j,k}} \tag{4}$$

To ensure that all values are positive, we calculate the absolute value of the difference between *Reduced Relative* $SHAP_{i,j,k}$ and *Relative* $SHAP_{i,j,k}$.[3] As a result, we receive a percentage value of difference for each feature i in model j trained on dataset k.

Using these two key measures, the following paragraphs ties together the results, offering a comprehensive evaluation of how changes in characteristics affect SHAP values across sample sizes.

5.1 Correlation

Figure 3 illustrates the relationship between relative size according to formula 3 (x-axis) and the percental difference according to formula 4 (y-axis) with varying correlation structures leaving other characteristics stable.

The data points in the figure are marker-coded to reflect four correlation configurations. Circles represent datasets with poor or negligible correlations between features and the target variable, squares represent datasets with a wide range of correlations, triangles represent datasets with consistently strong correlations, and diamonds denote datasets where some features exhibit perfect linear dependency (perfect multicollinearity). This distinction follows the one we present in the data generation process (Figure 2).

A clear trend emerges across all correlation configurations. As the relative size increases, the percentage difference in SHAP values decreases. This trend indicates that the more relevant a feature is marked by SHAP, the more stable the value across Slovin's data reduction, and that this connection is independent of features correlation or, in other words, model performance.

[3]Although no explicit restrictions were imposed on the data generation process regarding SHAP value distributions or correlations, we acknowledge that the absence of observed values exceeding 1 in the relative differences is possibly due to emergent properties of the synthetic data design. While this behavior held consistently across all simulation runs, its generalizability warrants further investigation in future research, particularly under more adversarial or unbalanced conditions.

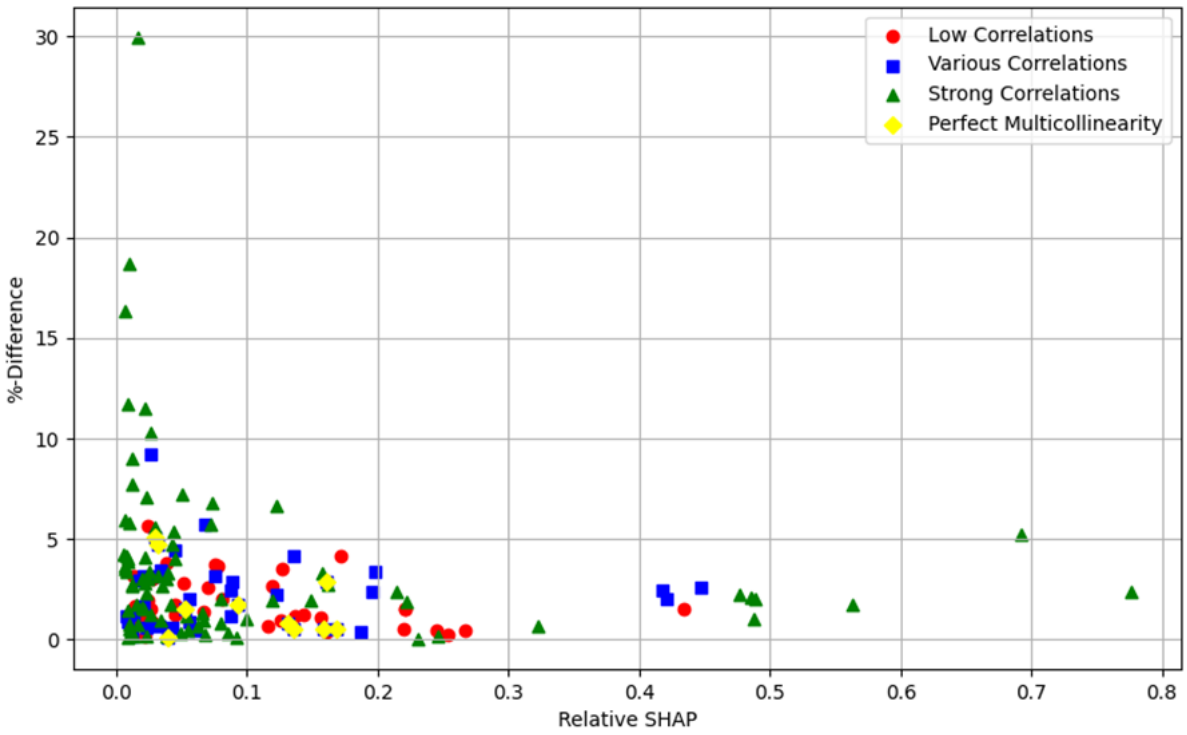

Fig. 3. Results of correlation variation

For datasets with poor correlations, the variability in SHAP values is notably low, mostly below 5%. This is surprising because of the inherent instability in datasets with weak relationships between features and the target variable, where noise should make SHAP computations more sensitive to sampling variations. Datasets with variable correlations also show moderate levels of variability (< 10%). Similarly, datasets with perfect multicollinearity display low variability (<= 5%). In contrast, datasets with high correlations exhibit high differences especially for relatively small values, but less for very high values, demonstrating that strong feature-target relationships result in inconsistent SHAP computations for low SHAP values.

At smaller relative sizes (< 0.1), the figure reveals a wide scatter of points across all correlation types, with percentage differences reaching up to 30%. This pattern highlights the general challenge of delivering stable SHAP values for very small SHAP values, particularly for datasets with strong correlations, where SHAP values are highly sensitive. As relative size increases, the points for all correlation configurations converge near the lower range of percentage differences. This convergence reflects the growing stability of SHAP computations as relative relevance increases.

5.2 Feature Type and Distribution

Figure 4 presents the same relationship as Figure 3. This time we distinguish between four categories of features: continuous features (circle), binary dummy features (square), features sampled from normal distributions (triangle), and features drawn from a mix of various distributions (diamond) that we establish in the data section leaving all other characteristics constant.

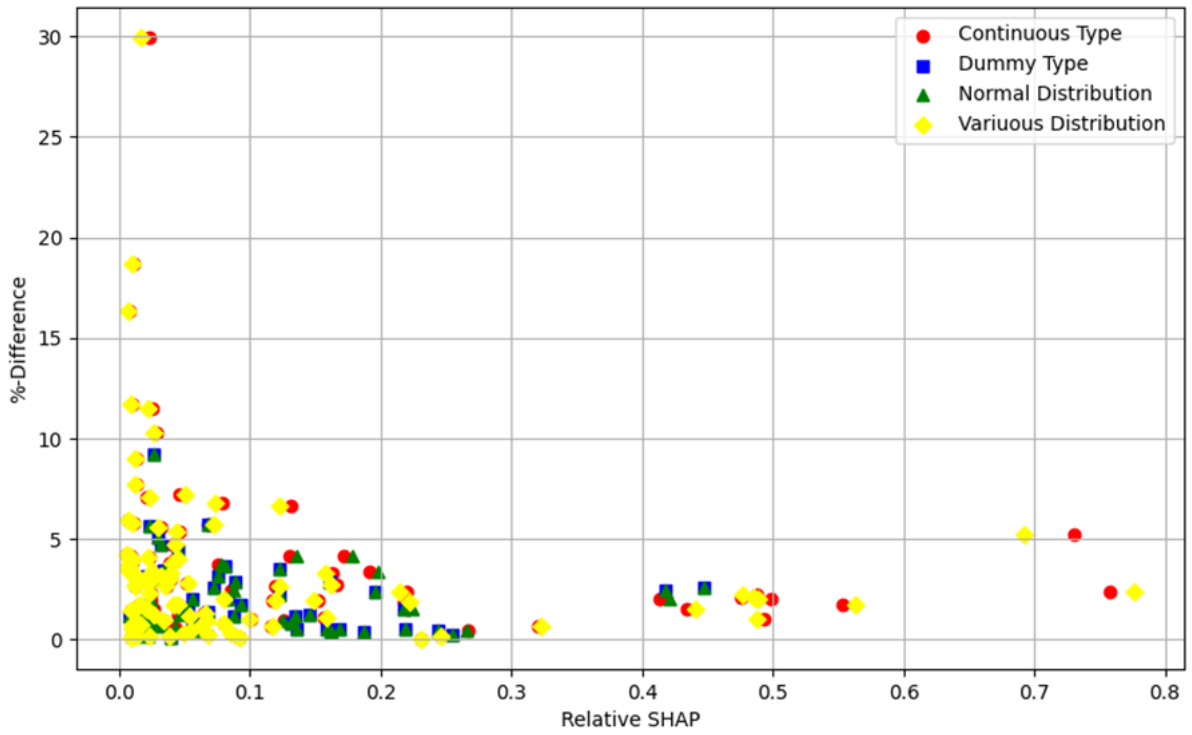

Fig. 4. Results of dependent feature variation

Again, we see the general trend showing that as the relative size increases, the percentage difference in SHAP values decreases across almost all feature types and distributions. Notably, this does not hold for dummy-type features, as they do not appear under the strong outliers. Continuous features and datasets with non-normally distributed features have higher deviation especially for very low SHAP values. Continuous features exhibit relatively high variability in SHAP values, particularly for smaller relative sizes (< 0.1). This heightened sensitivity is likely due to the dominance of continuous features in influencing model predictions, particularly in regression tasks, making their SHAP values more susceptible to changes in subsampling. By contrast, dummy features demonstrate much lower variability. Their binary or categorical nature limits the range of possible SHAP values, resulting in greater stability even at smaller SHAP values.

Features sampled from normal distributions exhibit moderate variability. The symmetry and consistency inherent in normal distributions supposably contribute to relatively stable SHAP outputs across samples. However, the variability is still slightly higher than that observed for dummy features, suggesting that the numerical nature of normally distributed features introduces some sensitivity to subsampling.

Features derived from mixed distributions also exhibit very high variability, particularly at smaller relative sizes. These features, drawn from a combination of skewed, bimodal, and other complex patterns, increase complexity to the models, which amplifies the challenges of maintaining representativeness in subsamples, obviously leading to greater variability in SHAP values. As the relative size increases, the variability for mixed distributions diminishes, but their inherent complexity continues to introduce subtle inconsistencies compared to more uniform feature types.

Again, for smaller relative dataset sizes (< 0.1), the scatter of points for all feature types and distributions, except dummy types, is substantial. This reflects the inherent difficulty of maintaining consistent SHAP values drawn from reduced subsamples, especially for small SHAP values and among those, particularly for features exhibiting diverse or complex distributions. As relative size increases, the points converge toward lower percentage differences, reflecting the improved stability of SHAP computations for larger SHAP values.

5.3 Target Type and Distribution

Looking at the target feature we distinguish distributional differences in further detail, so that we present SHAP values in two separate plots (Figure 5).

The first figure provides a comparison of datasets with continuous target variables (circle) and dummy target variables (square). Again, we see the trend that the relative dataset size increases, the percentage difference in SHAP values decreases, but this time with a clear difference in feature type. For continuous target variables, SHAP values exhibit much greater variability at smaller relative sizes. The pattern corresponds to the one we already see in dependent feature types. This is again likely because continuous targets are sensitive to subtle variations in feature relationships, making SHAP values more susceptible to subsampling effects. In contrast, dummy target variables display much lower variability, as their binary nature limits the complexity of feature-target interactions, leading to more consistent SHAP outputs across the whole branch of SHAP sizes. Although we can also recognize the negative correlation between relative size and difference here, it should be noted that the SHAP values in data samples with dummy target variables are rather stable.

The second figure examines how different target distributions affect SHAP value stability during Slovin subsampling. The distributions analyzed include normal (circle), uniform (square), gamma (triangle up), beta (diamond), skewnormal (star), logarithmic (triangle left), exponential (triangle right), and poisson (triangle down).

Some distributions introduce more variability than others, particularly at smaller relative sizes. Poisson and exponential display the highest variability at smaller relative sizes. These distributions apparently are more challenging for SHAP computations to stabilize after application of Slovin. Logarithmic targets also express higher differences at a small SHAP level, but the deflections are significantly smaller than the ones of poisson and exponential. All remaining target distributions, although showing more variance with small SHAPs, generally give out deviations below 10%. The higher variability in SHAP values for Poisson, exponential, and logarithmic target distributions can be attributed to their inherent skewness and concentration of values in specific regions. Exponential and Poisson distributions are heavily right-skewed, meaning that a large proportion of their values are clustered near zero, while a few extreme values extend far into the positive range. When applying Slovin's subsampling, the reduced dataset may disproportionately exclude these extreme values, leading to instability in SHAP value computations, especially at small SHAP levels where model sensitivities to small perturbations are highest. Logarithmic distributions, on the other hand, exhibit a different but related behavior: their values tend to spread unevenly, with large differences between consecutive observations at lower ranges. This characteristic can amplify small fluctuations when recalculating SHAP values after subsampling, making them more sensitive to representational distortions. By contrast, distributions such as normal and uniform maintain relatively stable variance across their entire range, reducing the impact of subsampling and leading to lower deviations in SHAP computations.

5.4 Data Size

As shown in the data section the subsample produced by Slovin's formula becomes relatively smaller compared to the original one, as the test dataset size increases (cp. Table 1). As this might logically influence the subsample's representativeness, we plot the relative SHAP values size against the percental difference across test data sizes (Figure 6). Each curve in the plot corresponds to a different sample size, ranging from 4,000 to 100,000 observations.

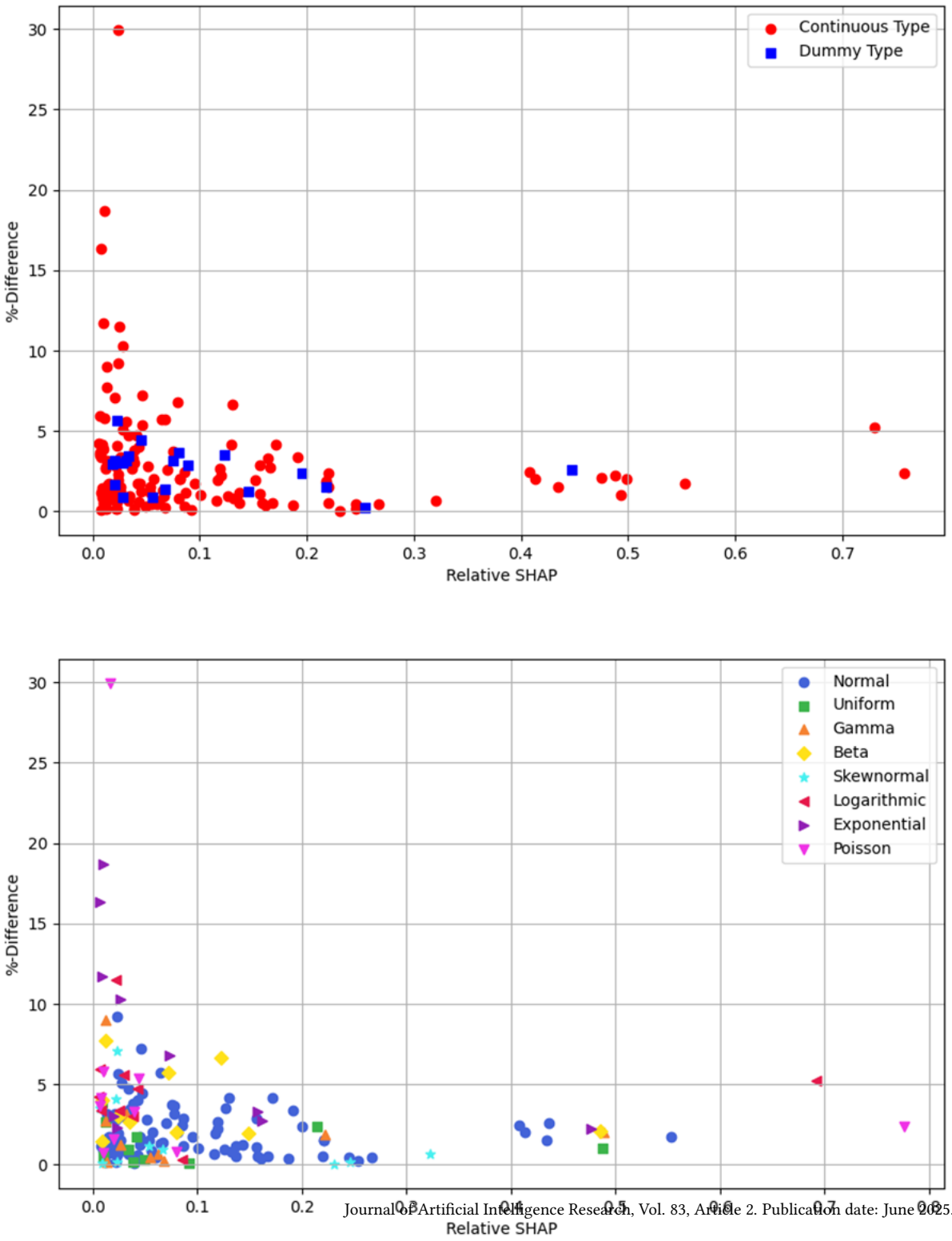

Fig. 5. Results of target feature variation

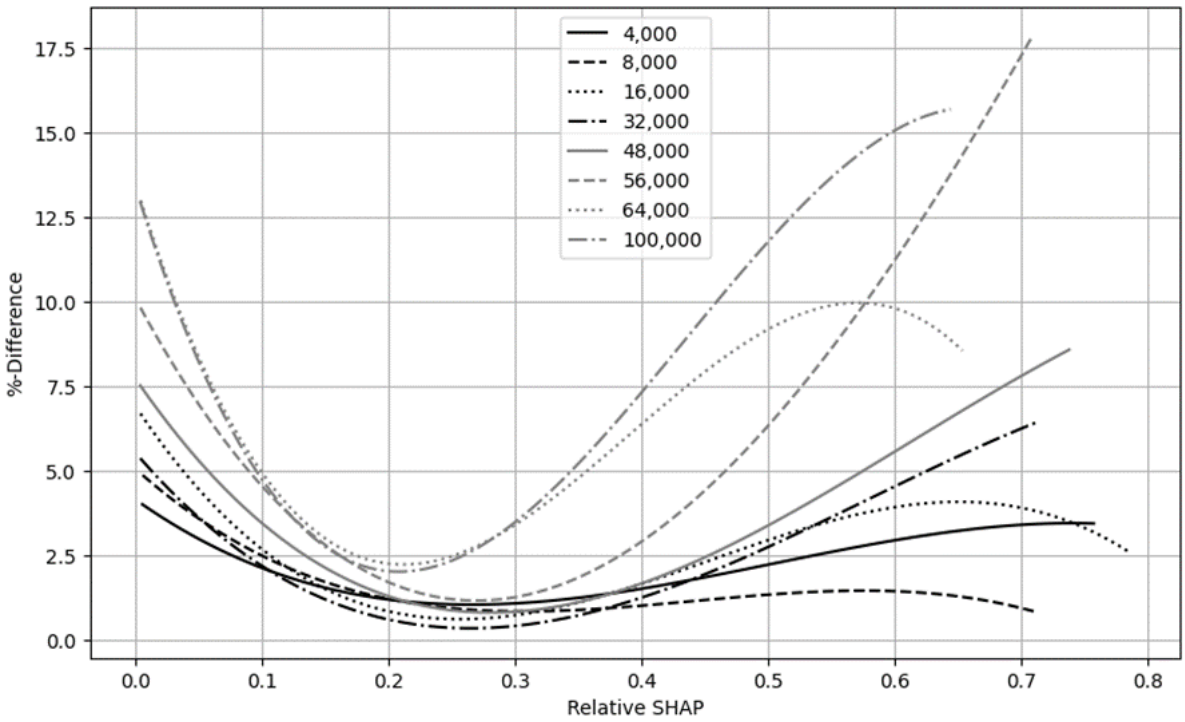

Fig. 6. Results of data size variation

As it would confuse the illustration, if we draw all single points as in previous figures, we calculate a fitted line for each data size. We choose the third-degree polynomial, as it allows for necessary non-linearity and at the same time is robust against outliers.

Across all sample sizes, the relationship between relative size and percentage difference follows a rather characteristic U-shaped curve. As the relative SHAP size increases from very small sizes, the percentage difference in SHAP values initially decreases, reaching a minimum between 0.2 and 0.4 before increasing again as relative size approaches 0.8. These curves very well reflect the interplay between representativeness and variability. For very small relative SHAPs, the subsample is not representative of the full dataset, resulting in high variability in SHAP values. As the relative size grows, the subsample becomes more representative, reducing variability and stabilizing SHAP computations. Beyond a certain point, the effect of relative size diminishes, and the increasing size leads to a slight uptick in percentage difference. For some datasets (4,000; 8,000; 64,000; 100,000) the fitted curve drops again in the end, but it is not representative of all datasets.

The curves reveal that larger sample sizes generally lead to higher percentage differences in SHAP values across all relative sizes. For instance, the dataset with 100,000 observations (dash-dotted grey) shows merely a steep drop of difference around the point of 0.2 to then drastically take off again demonstrating its insufficient stability and robustness. In contrast, the dataset with 4,000 observations (solid black) exhibits the lowest percentage differences, particularly at small and high relative sizes, reflecting a rather consistent application of Slovin subsampling. Intermediate sample sizes, such as 16,000 (dotted black) and 48,000 (solid grey), strike a balance, with moderate percentage differences that decrease significantly as the relative size increases and increases again

after hitting a low between 0.2 and 0.4. Altogether they still show robustness being below 7.5% throughout all SHAP sizes. Sample sizes from 56,000 and above show a significant shift upwards with steeper drops in the beginning, but also steeper takeoffs after hitting their lower bound. Taking together with Table 1 we note that with the subsample-sample ration of 5% and below the robustness significantly loses stability.

Each curve has a distinct minimum point where the percentage difference in SHAP values is at its lowest. For smaller datasets (e.g., 4,000), this minimum occurs at a smaller relative size, while for larger datasets (e.g., 100,000), the minimum shifts to a higher relative size. Moreover, the curvature of the lines becomes more pronounced as the sample size increases. Larger datasets, such as those with 100,000 observations, exhibit steeper curves, reflecting their lack of resilience to subsampling effects. Smaller datasets, by contrast, display flatter curves, emphasizing the diminishing variability introduced by small subsamples. This pattern suggests that Slovin's formula subsampling delivers more stable results for small datasets, whereas it loses its robustness with larger datasets.

5.5 Summary

The findings demonstrate that while Slovin's formula presents a promising trade-off between computational efficiency and interpretability, its effectiveness varies significantly depending on dataset characteristics.

Concerning correlations, a clear trend emerges, indicating that features with higher SHAP values exhibit greater stability after Slovin's subsampling, independent of correlation configurations. Datasets with weak correlations show the lowest variability in SHAP values, whereas those with strong feature-target relationships demonstrate substantial deviations, particularly for features with smaller SHAP values. Notably, datasets exhibiting perfect multicollinearity display a reduction in SHAP variability, suggesting that high redundancy among features mitigates the destabilizing effects of subsampling.

Regarding feature type and distribution, the results reveal that categorical features maintain relatively stable SHAP values, whereas continuous features display higher deviations, particularly at lower feature importance levels. Features derived from normal distributions exhibit moderate variability, while those sampled from mixed distributions tend to show greater fluctuations. These findings underscore the influence of data homogeneity in ensuring SHAP robustness after subsampling, suggesting that the stability of feature importance measures is closely tied to the distributional properties of input variables.

Observing target variable types and distributions, similarly to the findings on feature types, dummy target variables contribute to greater consistency in SHAP values, reinforcing the notion that simpler categorical structures enhance model interpretability under subsampling conditions. However, target distribution plays a crucial role in shaping SHAP variability. Highly skewed distributions such as Poisson and exponential exhibit greater fluctuations in SHAP values, while normal and non-skewed distributions yield more stable results. This pattern suggests that the inherent properties of target distributions influence the reliability of SHAP-based interpretability after applying Slovin's formula.

Probably the most significant findings emerge from the investigation of dataset size and the subsample-to-sample ratio. Across all dataset sizes, the relationship between SHAP stability and feature importance follows a U-shaped pattern, where stability is highest for mid-range SHAP values and declines at both extremes. Larger datasets generally exhibit higher variability in SHAP values, particularly when the subsample-to-sample ratio falls below 5%. Smaller datasets, by contrast, retain a higher degree of stability across all feature importance levels. These results indicate that Slovin's formula is most effective for small to medium-sized datasets, where it successfully reduces computational costs while preserving interpretability. For large datasets, one possible mitigation strategy is to employ adaptive sampling techniques that prioritize representativeness. For instance, stratified sampling based on feature distributions or importance-based sampling that considers feature variance or preliminary model influence scores can help maintain the interpretative quality of SHAP values. Additionally,

hybrid approaches that combine Slovin's formula as a baseline estimator with more dynamic sampling adjustments either informed by model diagnostics or data clustering may help preserve both computational efficiency and explanation fidelity. Future work should further explore these integrative techniques to extend the benefits of low-cost sampling into large-scale ML contexts while safeguarding SHAP robustness.

In conclusion, the findings confirm that Slovin's formula can serve as a viable computational optimization tool for SHAP value estimation, provided that specific dataset characteristics are considered. The method proves beneficial in reducing computational burdens for small and medium-sized datasets while maintaining the integrity of feature importance measures. However, its reliability diminishes in large datasets with highly skewed distributions, emphasizing the importance of carefully assessing dataset attributes before applying Slovin's formula. By identifying the conditions under which this approach preserves SHAP stability, this study contributes to the broader discourse on resource-efficient AI, offering a structured methodology for balancing computational feasibility with explainability in machine learning interpretability.

6 Conclusion

Our study introduces a resource-efficient method to balance computational efficiency and interpretability in machine learning by incorporating Slovin's formula into SHAP-based feature importance estimation. The findings offer practical solutions for scientists and researchers who face increasing computational challenges in XAI, particularly when working with large datasets or limited computational resources.

The research demonstrates that Slovin's formula can significantly alleviate the computational burden while preserving interpretability in most scenarios. Features with higher importance scores retain greater stability after subsampling, regardless of dataset correlation structures. Additionally, categorical and normally distributed features tend to produce more reliable SHAP estimates, while continuous and mixed-distribution features present higher fluctuations. Similarly, target variables with non-skewed distributions exhibit better stability, while highly skewed distributions, such as Poisson and exponential, present greater deviations. Importantly, the results establish that mid-ranked SHAP values offer the best balance between computational feasibility and stability, presenting a clear optimization strategy for researchers.

Perhaps the most significant contribution of this study is its applicability to small and medium-sized datasets, where Slovin's formula maintains high stability even under tight computational constraints. For scientists dealing with mid-sized data in disciplines such as healthcare, finance, or environmental sciences, this method provides a scalable solution to accelerate SHAP-based interpretability without compromising on accuracy. However, the research emphasizes caution for large datasets exceeding 100,000 observations, where subsample-to-sample ratios below 5% reduce reliability and require supplementary strategies.

The study not only addresses a critical gap in the discourse on resource-efficient AI but also equips the scientific community with a systematic approach to reduce processing costs while maintaining transparency and trust in machine learning predictions. By optimizing SHAP computations, researchers can better utilize their resources, minimize energy consumption, and accelerate experimentation, making AI research both more sustainable and accessible.

One notable limitation of this study is the exclusive use of synthetic datasets. This approach offers strong experimental control, enabling systematic manipulation of feature distributions, correlation structures, and data sizes. However, such datasets may not fully capture the complexities, irregularities, and noise characteristics inherent in real-world domains. In practice, real-world data often contain missing values, unbalanced class distributions, hidden confounders, and domain-specific patterns that could affect both SHAP values and the performance of Slovin-based subsampling. Consequently, while our results provide valuable insights into the behavior of SHAP under controlled conditions, caution should be exercised when generalizing these findings to operational datasets.

Future research should extend this work by validating Slovin's formula on real-world benchmarks, ideally from regulated domains where explainability is not only a technical goal but a legal requirement.

Furthermore, for future work it is reasonable to extend our analysis by exploring the effects of increasing the number of features beyond the current ten investigating configurations with 10, 20, 50, or even 100 features in order to better reflect the complexity of real-world datasets and examine how SHAP subsampling behaves under higher-dimensional settings. Additionally, we recognize the importance of assessing computational performance in greater detail. A systematic evaluation comparing execution times as the data size and feature characteristics vary will be invaluable. Such an analysis will help clarify the trade-offs between interpretability and computational efficiency, especially in large-scale applications. These directions not only promise to refine the current approach but also to contribute toward a more comprehensive understanding of scalable XAI techniques.

Furthermore, by integrating Slovin's formula into machine learning workflows, researchers can achieve a high degree of transparency, meet regulatory demands for interpretability, and ensure that models remain both resource-efficient and accessible across academic and industrial contexts. The study encourages the adoption of hybrid techniques that combine Slovin's subsampling with other optimization methods to further improve scalability and robustness. This work aims to serve as a foundation for future advancements, promoting more sustainable and responsible AI practices for scientists around the world.

References

[1] A. M. Abdullahi. 2023. The challenges of advancing inclusive education: the case of somalia's higher education. *Journal of Law and Sustainable Development*, 11, 2, e422–e422.

[2] A. A. Adeniran, A. P. Onebunne, and P. William. 2024. Explainable ai (xai) in healthcare: enhancing trust and transparency in critical decision-making. *World Journal of Advanced Research and Reviews*, 23, 2647–2658.

[3] Q. An, S. Rahman, J. Zhou, and J. J. Kang. 2023. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*, 23, 9, 4178.

[4] Z. Asimiyu. 2024. Balancing explainable ai and security: machine learning for iot, finance, and real estate. Preprint. (2024).

[5] S. Athey and G. W. Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 1, 685–725.

[6] S. Bachmann. 2025. Interpretable machine learning for the german residential rental market – shedding light into model mechanics. *Aestimum*. Just Accepted.

[7] M. L. Baptista, K. Goebel, and E. M. Henriques. 2022. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artificial Intelligence*, 306, 103667.

[8] M. A. Batouei. 2024. *A Feasibility Study On Artificial Neural Network-Based Prediction And Optimization Of Autoclave Curing Process utcomes Via Simulation-Based Thermal Images And Haralick Texture Features*. Ph.D. Dissertation. University Of British Columbia, Okanagan.

[9] R. Bellman. 1961. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4, 6, 284.

[10] A. Bennetot et al. 2024. A practical tutorial on explainable ai techniques. *ACM Computing Surveys*, 57, 2, 1–44.

[11] L. Breiman. 2001. Random forests. *Machine Learning*, 45, 5–32.

[12] L. Breiman and J. H. Friedman. 1997. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 1, 3–54.

[13] T. Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. Vol. 33, 1877–1901.

[14] N. Burkart and M. F. Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245–317.

[15] G. Chandrashekar and F. Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40, 1, 16–28.

[16] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan. 2018. Learning to explain: an information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 883–892.

[17] M. Christoph. 2020. *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

[18] W. G. Cochran. 1977. *Sampling Techniques*. (3rd ed.). John Wiley and Sons, New York.

[19] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine learning*, 20, 273–297.

[20] R. Davis, B. Buchanan, and E. Shortliffe. 1977. Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence*, 8, 1, 15–45.

[21] C. Davoli. 2024. *Data-Driven Approaches for the design of Traction Electrical Motors*. Ph.D. Dissertation. Politecnico di Torino.
[22] J. Dean and S. Ghemawat. 2004. Mapreduce: simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation*.
[23] I. K. Fodor. 2002. A survey of dimension reduction techniques. Tech. rep. UCRL-ID-148494. Lawrence Livermore National Laboratory.
[24] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. 2019. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*.
[25] I. Guyon and A. Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
[26] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9, 8, 1735–1780.
[27] G. D. Israel. 1992. Determining sample size. (1992).
[28] D. Janzing, L. Minorics, and P. Blöbaum. 2020. Feature relevance quantification in explainable ai: a causal problem. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 2907–2916.
[29] I. T. Jolliffe and J. Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 2065, 20150202.
[30] M. Kirchner. 2024. Nuclear power for ai data centers: microsoft has three mile island reactivated. Accessed: 2024-02-05. https://www.heise.de/en/news/Nuclear-power-for-AI-data-centers-Microsoft-has-Three-Mile-Island-reactivated-9939253.html.
[31] B. Krämer, C. Nagl, M. Stang, and W. Schäfers. 2023. Explainable ai in a real estate context – exploring the determinants of residential real estate values. *Journal of Housing Research*, 32, 2, 204–245.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Vol. 25.
[33] P. Kumar. 2023. Explainable ai/ml testing: ensuring transparency, accountability, and compliance. *Journal of Artificial Intelligence, Machine Learning & Data Science*, 1, 4, 476–482.
[34] V. Kumar, K. Joshi, R. Kumar, M. Memoria, A. Gupta, and F. Ajesh. 2025. Future prospective of neuromorphic computing in artificial intelligence: a review, methods, and challenges. In *Primer to Neuromorphic Computing*, 185–197.
[35] D. Lanin and N. Hermanto. 2019. The effect of service quality toward public satisfaction and public trust on local government in indonesia. *International Journal of Social Economics*, 46, 3, 377–392.
[36] H. M. Levitt, M. Bamberg, J. W. Creswell, D. M. Frost, R. Josselson, and C. Suárez-Orozco. 2018. Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the apa publications and communications board task force report. *American Psychologist*, 73, 1, 26.
[37] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. 2020. Explainable ai: a review of machine learning interpretability methods. *Entropy*, 23, 1, 18.
[38] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding. 2023. Multi-modal fusion network with complementarity and importance for emotion recognition. *Information Sciences*, 619, 679–694.
[39] S. M. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. Vol. 30.
[40] M. Minsky and S. Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA.
[41] J. N. Morgan and J. A. Sonquist. 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58, 302, 415–434.
[42] H. Müller, A. Holzinger, M. Plass, L. Brcic, C. Stumptner, and K. Zatloukal. 2022. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the european in vitro diagnostic regulation. *New Biotechnology*, 70, 67–72.
[43] D. K. Nguyen, G. Sermpinis, and C. Stasinakis. 2023. Big data, artificial intelligence and machine learning: a transformative symbiosis in favour of financial technology. *European Financial Management*, 29, 2, 517–548.
[44] D. Normelindasari and A. Solichin. 2020. Effect of system quality, information quality, and perceived usefulness on user satisfaction of webstudent applications to improve service quality for budi luhur university students. In *Proceedings of the 4th International Conference on Management, Economics and Business (ICMEB 2019)*. Atlantis Press, 77–82.
[45] S. S. Patel. 2023. Explainable machine learning models to analyse maternal health. *Data & Knowledge Engineering*, 146, 102198.
[46] N. Patidar, S. Mishra, R. Jain, D. Prajapati, A. Solanki, R. Suthar, K. Patel, and H. Patel. 2024. Transparency in ai decision making: a survey of explainable ai methods and applications. *Advances of Robotic Technology*, 2, 1.
[47] S. P. Putri, Y. Nakayama, F. Matsuda, T. Uchikata, S. Kobayashi, A. Matsubara, and E. Fukusaki. 2013. Current metabolomics: practical applications. *Journal of Bioscience and Bioengineering*, 115, 6, 579–589.
[48] J. R. Quinlan. 1993. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, 236–243.
[49] A. Radford et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8748–8763.
[50] H. K. Ramadhani and D. Aldyandi. 2024. The relationship between the level of knowledge of kiasu culture and the way of view of high school/vocational school students in the city of surabaya to achieve golden indonesia. *Medical Technology and Public Health Journal*, 8, 1, 55–61.

[51] G. Ras, N. Xie, M. Van Gerven, and D. Doran. 2022. Explainable deep learning: a field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73, 329–396.
[52] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Roettger, H. Mueller, and A. Holzinger. 2024. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86, 101243.
[53] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
[54] F. Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 6, 386–408.
[55] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323, 6088, 533–536.
[56] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. 2021. Explaining deep neural networks and beyond: a review of methods and applications. *Proceedings of the IEEE*, 109, 3, 247–278.
[57] R. E. Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5, 197–227.
[58] L. S. Shapley. 1953. Stochastic games. *Proceedings of the National Academy of Sciences*, 39, 10, 1095–1100.
[59] V. W. Skrivankova et al. 2021. Strengthening the reporting of observational studies in epidemiology using mendelian randomization: the STROBE-MR statement. *JAMA*, 326, 16, 1614–1621.
[60] E. Strubell, A. Ganesh, and A. McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 9. Vol. 34, 13693–13696.
[61] M. Sundararajan and A. Najmi. 2020. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 9269–9278.
[62] J. J. Tejada and J. R. B. Punzalan. 2012. On the misuse of slovin's formula. *The Philippine Statistician*, 61, 1, 129–136.
[63] A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59, 236, 433–460.
[64] G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu. 2022. On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, 74, 851–886.
[65] L. Van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 11, 2579–2605.
[66] E. N. Vanegas Herrera. 2024. *Three essays on machine learning and time series applications on finance: Skew index and return predictability*. Ph.D. Dissertation. Unknown.
[67] V. N. Vapnik. 1999. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10, 5, 988–999.
[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. Vol. 30.
[69] T. Wanga et al. 2024. Explainable ai across domains: techniques, domain-specific applications, and future directions. (2024).
[70] J. M. Wooldridge. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.
[71] Z. Yang et al. 2024. Cogvideox: text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Received 11 February 2025; revised 25 March 2025; accepted 1 May 2025