EUROPEAN JOURNAL OF EMERGING ARTIFICIAL INTELLIGENCE

# OPTIMIZING SHAP EXPLANATIONS: A COST-EFFECTIVE DATA SAMPLING METHOD FOR ENHANCED INTERPRETABILITY

**Prof. Li Wei**
**School of Data Science, The Chinese University of Hong Kong, Hong Kong**

**Alessia Romano**
**Department of Computer Engineering, Politecnico di Milano, Italy**

**V0LUME01 ISSUE01 (2024)**
**Published Date: 25 December 2024 // Page no.: - 71-89**

## ABSTRACT

The proliferation of complex machine learning (ML) models in critical domains such as healthcare, finance, and real estate has underscored the urgent need for Explainable Artificial Intelligence (XAI) [2, 3, 4, 31, 43, 45]. While these "black-box" models often achieve superior predictive performance, their inherent lack of transparency hinders trust, accountability, and the ability to effectively debug and refine them. SHapley Additive exPlanations (SHAP) is a widely recognized model-agnostic XAI method that provides detailed insights into individual feature contributions to model predictions, robustly grounded in cooperative game theory [39, 58, 61]. However, a significant drawback of SHAP, particularly when applied to large datasets or computationally intensive models, is its substantial computational overhead, often rendering its application impractical in resource-constrained or real-time operational environments [60, 64]. This article proposes and rigorously investigates a data-efficient strategy for achieving SHAP interpretability by leveraging intelligent data reduction techniques. Specifically, we explore the application of Slovin's formula, a statistical sampling technique traditionally employed in survey research, as a low-cost heuristic for data reduction. Unlike more complex feature selection or dimensionality reduction methods, Slovin's formula requires minimal prior statistical knowledge of the dataset's properties, offering a straightforward, accessible, and efficient alternative for subsampling without extensive preprocessing. Through controlled experiments on synthetically generated datasets, we demonstrate that by judiciously sampling a representative subset of the original data, SHAP explanations can be generated with significantly reduced computational cost while maintaining a high degree of fidelity to the explanations derived from the full dataset. Our findings highlight a U-shaped trade-off in SHAP value stability: mid-ranked features tend to remain more stable under subsampling, whereas features with extreme (very low or very high) importance exhibit higher fluctuations. Furthermore, we observe that categorical and non-skewed distributed features generally maintain greater robustness, while highly skewed target distributions introduce increased variability. Crucially, the effectiveness and reliability of Slovin's formula diminish when the subsample-to-sample ratio falls below a critical threshold of approximately 5%. This empirical evaluation underscores the potential of our cost-effective approach to democratize access to advanced interpretability, enabling faster model insights, improved debugging, and broader, more sustainable deployment of transparent AI systems in various domains.

**Keywords:** Cryptographic hash functions, Preimage attacks, Cube-and-Conquer, SAT solvers, MD4, MD5, Cryptanalysis, Boolean Satisfiability, Parallel SAT solving, Dobbertin's constraints.

## INTRODUCTION

The rapid advancement and widespread adoption of machine learning (ML) and artificial intelligence (AI) have fundamentally reshaped numerous industries, offering unprecedented capabilities in prediction, automation, and complex decision-making [3, 43, 66]. From accurately diagnosing diseases in healthcare to meticulously detecting fraudulent financial transactions and optimizing real estate valuations, AI systems are increasingly being deployed in high-stakes environments where their decisions wield significant societal and economic impact [2, 3, 4, 6, 31, 45]. This pervasive integration, however, comes with a growing recognition of a critical challenge: many state-of-the-art ML models, especially intricate architectures like deep neural networks and powerful ensemble methods such as Random Forests or Gradient Boosting Machines, frequently operate as "black boxes" [11, 19, 32, 55, 56, 67]. Their complex internal workings make it exceedingly difficult to ascertain why a specific decision was rendered or how individual input features contributed to a particular outcome [14, 17, 37]. This inherent opacity raises profound concerns regarding public trust, fairness, accountability, and the capacity to effectively debug and refine these models [2, 10, 33, 46, 51, 69].

The need for transparency is not merely an academic pursuit; it is increasingly a regulatory imperative. For instance, the European In Vitro Diagnostic Regulation (IVDR) explicitly classifies software, including AI algorithms, as medical devices, thereby subjecting them to stringent requirements for traceability and transparency in their decision-making processes. This regulatory landscape mandates that professionals must possess the ability to understand and justify decisions supported by AI systems, thereby underscoring the critical necessity for explainability, interpretability, and causability in AI models [42]. Without the capacity to trace the reasoning behind an AI's output, human professionals cannot be held fully accountable, nor can they confidently place trust in the system's recommendations [5]. This shift has driven research beyond solely optimizing predictive performance towards ensuring models are interpretable, focusing on understanding how input features contribute to predictions and how these insights can improve overall decision-making [17, 56].

In response to these multifaceted challenges, the burgeoning field of Explainable Artificial Intelligence (XAI) has emerged, dedicated to developing methodologies that render AI systems more transparent, interpretable, and ultimately understandable to human users [10, 14, 37, 51]. XAI techniques strive to illuminate the intricate behaviors of models, providing invaluable insights into feature importance, decision boundaries, and explanations for individual predictions [17, 37, 52]. This not only fosters greater user confidence and trust but also significantly aids developers in identifying and mitigating biases, enhancing model performance, and ensuring strict compliance with evolving regulatory frameworks [2, 33, 42].

Among the diverse array of XAI techniques, SHapley Additive exPlanations (SHAP) has garnered considerable prominence due to its robust theoretical foundations [39, 58, 61]. Rooted in cooperative game theory, the core principle of SHAP is to attribute the contribution of each feature to a model's prediction by calculating the average marginal contribution of that feature across all possible permutations (coalitions) of features. This approach ensures a "fair" allocation of the prediction difference among features, establishing SHAP as a powerful tool for both global (overall feature importance) and local (individual prediction explanation) interpretability [39, 61]. The model-agnostic nature of SHAP further broadens its applicability across a vast spectrum of ML architectures, ranging from traditional linear models to cutting-edge deep learning networks [13, 39, 49, 68, 71].

Despite its compelling theoretical rigor and undeniable practical utility, SHAP is confronted by a significant practical limitation: its inherent computational intensity [64]. The process of calculating SHAP values typically involves iterating through permutations of features and executing numerous predictions with the target ML model. This often leads to a "combinatorial explosion" in computational complexity [16], which can become prohibitively expensive and time-consuming for large datasets or computationally demanding models [60]. As data volumes continue their exponential growth and AI models become increasingly sophisticated, the sheer computational burden associated with generating comprehensive SHAP explanations poses a considerable barrier to their routine application in many real-world scenarios [22, 30, 49, 68, 71]. This challenge is particularly acute in situations demanding real-time explanations, or in resource-constrained environments such as edge devices or specific Internet of Things (IoT) applications [4]. The environmental impact and energy consumption of large-scale AI operations, exemplified by the reactivation of nuclear power plants to support data centers, further underscore the urgent need for resource-efficient AI solutions [30, 60].

This article directly addresses this critical computational bottleneck by proposing and empirically evaluating a cost-effective data reduction approach for generating SHAP explanations. Our central hypothesis is that a carefully selected, smaller, yet statistically representative subset of the original data can yield SHAP values that are highly correlated with those obtained from the full dataset. This strategy aims to drastically reduce the computational time and resources required without significantly compromising the quality or fidelity of the resulting interpretability. We delve into the efficacy of simple random sampling as a primary data reduction strategy, and specifically investigate the applicability of Slovin's formula, a statistical sampling technique, to determine appropriate subsample sizes. By demonstrating its potential to unlock efficient SHAP-based interpretability, this research aims to facilitate the broader and more practical adoption of explainable AI for even the most demanding and resource-sensitive applications.

## 2. METHODS

This section provides a detailed exposition of the theoretical underpinnings of SHAP, the proposed data reduction methodology focusing on Slovin's formula, the comprehensive experimental setup, and the quantitative metrics employed to evaluate both the computational efficiency and the interpretability quality (fidelity) of the achieved explanations. Our methodology is designed to systematically explore the trade-offs involved and identify the conditions under which the proposed approach offers a viable solution to the computational challenges of SHAP.

### 2.1 Background on SHAP (SHapley Additive exPlanations)

SHAP is a unified and widely adopted framework for interpreting model predictions, building directly upon the rigorous concept of Shapley values derived from cooperative game theory [39, 58, 61]. In this framework, the input features of a machine learning model are conceptualized as "players" in a collaborative game, where the "gain" or "payout" of the game is the prediction output

by the model for a specific instance. SHAP values, denoted as ϕi, quantify the contribution of each feature i to a prediction by calculating the average marginal contribution of that feature across all possible unique permutations (or "coalitions") of features. This averaging process ensures that the contribution is fairly distributed, irrespective of the order in which features are added to a coalition.

Mathematically, the SHAP value ϕi for a feature i given a model f and an input instance x is defined as:

$$\phi i(f,x)=S\subseteq N\setminus\{i\}\sum|N|!|S|!(|N|-|S|-1)![fx(S\cup\{i\})-fx(S)]$$

where:

- N represents the complete set of all features considered by the model.
- S denotes a subset of features that does not include feature i.
- fx(S) is the prediction of the model when using only the features present in the set S. Features not included in S are handled by marginalizing out their effects. This is typically achieved by averaging predictions over a background dataset (e.g., the training data or a representative sample thereof) or by setting them to a baseline value.
- The term [fx(S∪{i})−fx(S)] represents the marginal contribution of adding feature i to the existing coalition S. This difference captures how much the prediction changes when feature i is introduced, given the context of the features already in S.

SHAP possesses several highly desirable properties that contribute to its widespread adoption and theoretical soundness. These include:

- Local Accuracy (or Local Fidelity): The sum of the SHAP values for all input features approximately equals the difference between the model's output for the current input and the expected output based on the background dataset.
- Consistency: If a feature's marginal contribution always increases (or stays the same) whenever it's added to a coalition, its SHAP value will not decrease.
- Missingness: Features that are missing (or effectively have no impact) for a given prediction are assigned a SHAP value of zero.

While the exact computation of Shapley values is combinatorially explosive, requiring 2|N| model evaluations, making it intractable for many real-world problems with even a moderate number of features, SHAP offers various practical approximation algorithms [39]. Notable among these are KernelSHAP and TreeSHAP. KernelSHAP is a model-agnostic approximation that uses a linear LIME-like model weighted by Shapley kernel to estimate the values [39, 53]. TreeSHAP is an optimized algorithm specifically designed for tree-based models (like Random Forests and Gradient Boosting Machines) that can compute exact SHAP values much more efficiently for these specific architectures [39]. Despite these algorithmic advancements, the computational burden remains a significant challenge for large datasets, as each SHAP value computation still necessitates numerous model evaluations, often against a substantial background dataset [64]. This inherent intensity is the primary motivation for exploring data reduction strategies.

2.2 Data Reduction Strategy

The core objective of this research is to substantially reduce the computational overhead associated with SHAP value calculation without unduly compromising the quality or fidelity of the resulting explanations. Our proposed strategy centers on the judicious application of data reduction techniques, specifically data sampling, to the dataset that serves as the "background" for SHAP calculations or as the collection of instances for which explanations are desired. Instead of performing SHAP computations against the entirety of a large dataset, we advocate for the use of a carefully selected, smaller, yet representative subset.

Rationale for Data Sampling:

The underlying assumption guiding our approach is that for a significant proportion of real-world datasets, a smaller, statistically representative sample can effectively capture the essential relationships, patterns, and feature distributions inherent in the full dataset. Consequently, this representative subset can then serve as an efficient "background dataset" for SHAP computations or as the specific instances for which interpretability insights are sought, leading to substantial computational and memory savings. The general concept of reducing data volume or dimensionality to enhance model efficiency and interpretability has been explored in various contexts, including feature selection and dimensionality reduction techniques [15, 23, 25]. However, our approach using Slovin's formula distinguishes itself by offering a simple, low-cost sampling mechanism that does not explicitly alter the feature space or require complex preprocessing steps beyond simple subsampling.

Sampling Methodology: The Role of Slovin's Formula

For this study, we primarily employ simple random sampling without replacement [18, 27, 62]. This foundational method involves selecting a fixed number of instances (k) or a fixed percentage of instances (P%) from the original dataset entirely at random. To guide the determination of an appropriate sample size, particularly in contexts where detailed statistical parameters like population variance are unknown or difficult to ascertain, we leverage Slovin's formula.

Slovin's Formula in Detail:

Slovin's formula is a well-recognized heuristic method used to estimate an appropriate sample size (n) for surveys or research studies when the total population size

(N) is known but other statistical parameters (like population standard deviation) are not readily available. It provides an approximation of the sample size needed to achieve a specified margin of error (e) in estimates. The formula is expressed as:

n=1+Ne2N

where:

- n is the calculated required sample size.
- N is the total population size (in our context, the size of the original dataset).
- e is the desired margin of error (expressed as a decimal, e.g., 0.05 for 5% margin of error). A commonly used margin of error in the literature is e=0.02 [62].

While Slovin's formula is widely cited for its simplicity and practicality in fields like business research, education, and social sciences [44, 47, 50], its precise origins are somewhat debated, and it is often considered a heuristic rather than a rigorously derived statistical theorem [62]. It is closely related to early discussions on sample size estimation techniques that sought to provide simplified calculations for practitioners without advanced statistical training [18, 35]. Its user-friendliness and rapid calculation capabilities make it especially useful in exploratory studies or when researchers face time, budget, or logistical constraints, or when detailed population parameters are unavailable [1, 27, 36].

Justification for Engaging with Slovin's Formula:

The primary critique against Slovin's formula is its reliance on simplifying assumptions, which can limit its applicability to specific scenarios, particularly for inferential problems beyond estimating proportions [62]. However, as demonstrated by Bachmann (2025) [6], Slovin's formula can be effectively adapted as a data reduction technique to tackle computational constraints in complex machine learning contexts, such as calculating SHAP values for very large datasets (e.g., over 0.5 million observations). In such applications, the formula helps reduce the necessary test dataset from a large number of observations (e.g., 100,000) to a significantly smaller, manageable size (e.g., 2,946 observations), thereby enabling feature importance analysis without overburdening computational resources. Crucially, as empirical results have suggested, this reduction can be achieved without significantly compromising the interpretability of the models, with only negligible deviation in SHAP values between the reduced and full datasets. This innovative application directly challenges the assertion that Slovin's formula is inherently flawed for all purposes, highlighting its utility when carefully justified and applied within specific research contexts. By integrating Slovin's formula into SHAP computations, it becomes possible to achieve a critical balance between computational efficiency and the preservation of interpretability, addressing scalability concerns cited in XAI literature [37, 49].

Process Flow for Data-Efficient SHAP Explanation Generation:

The systematic process flow for generating SHAP explanations using our data reduction approach is as follows:

1. Original Dataset and Model: Begin with the complete, large-scale dataset D and a pre-trained black-box machine learning model M.
2. Sample Size Determination: Determine the target sample size (n) for the reduced dataset using Slovin's formula, based on the total population size (N) of the original dataset and a chosen margin of error (e).
3. Data Sampling: Randomly sample n instances from the original dataset D to create a reduced dataset, denoted as Dreduced. This Dreduced will serve as the background dataset for SHAP computations, or as the set of instances for which local SHAP explanations are to be generated.
4. SHAP Calculation: Compute the SHAP values for the instances of interest (e.g., a specific prediction, a test set, or the Dreduced itself) by referencing Dreduced as the background dataset.
5. Evaluation and Comparison: Compare the SHAP explanations generated using the reduced dataset Dreduced against those generated using the full original dataset D. This comparison forms the basis of our fidelity assessment.

2.3 Experimental Setup

To rigorously evaluate the effectiveness and validity of the proposed data reduction approach, we designed and conducted a series of controlled experiments. These experiments utilized synthetically generated datasets, allowing for precise manipulation of various data characteristics, and involved a diverse selection of machine learning models.

2.3.1 Synthetic Data Generation

A critical component of our methodology involved the generation of synthetic datasets. This approach offers significant advantages, primarily providing absolute control over data characteristics, correlations, and distributions, thereby enabling systematic and isolated analysis of their impact on SHAP value stability under subsampling. The data generation process follows a structured, iterative framework as conceptually illustrated in Figure 2 (refer to the Results section for visual representation and detailed discussion).

The iterative data generation process encompasses four main steps:

1. Define Data Size (Step 4 in Figure 2): This initial step establishes the total number of observations (N) for the synthetic datasets. We generated datasets with varying sizes, ranging from 20,000 to a large-scale 500,000

observations. The choice of these sizes is pragmatic, balancing the computational feasibility of initial SHAP calculations with the need to explore scenarios where subsampling becomes crucial due to large data volumes. Notably, as demonstrated by Slovin's formula, the calculated subsample size (n) increases monotonically with N but at a diminishing marginal rate. Analytically, as $N \to \infty$, $n \to e21$, which for a margin of error e=0.02, asymptotically approaches 2,500. The first derivative, dNdn=(1+Ne2)21, confirms the decreasing marginal gains, and the second derivative, dN2d2n=(1+Ne2)32e2<0, further confirms the concavity of the relationship. This asymptotic behavior means that for very large datasets, the subsample becomes a progressively smaller proportion of the total data, increasing the risk of non-representativeness if the total size becomes excessively large. Table 1, provided in the Results section, concretely illustrates how the subsample size and its ratio to the total sample change with increasing dataset size, based on an error rate e=0.02.

2. Define Target Variable (Step 3 in Figure 2): The target variable, serving as the dependent variable for our ML models, was generated with diverse statistical properties. It could be sampled from a wide range of distributions, including normal, uniform, exponential, logarithmic, gamma, beta, Poisson, and skew-normal distributions. This variety allowed us to simulate different real-world phenomena, from symmetrically distributed outcomes to highly skewed or discrete counts. Furthermore, the target variable could be specified as either continuous (for regression tasks) or binary (for classification problems), ensuring a comprehensive evaluation across different machine learning problem types.

3. Generate 10 Features (Step 2 in Figure 2): For each dataset, ten input features were generated. This number, while somewhat arbitrary, was chosen to be sufficient for modeling necessary interdependencies and supplying various feature characteristics within one dataset, while maintaining clarity for analysis. Future research could explore the impact of varying the number of features. Similar to the target variable, these features were sampled from a broad selection of statistical distributions (normal, uniform, exponential, logarithmic, gamma, beta, Poisson, skew-normal), mimicking diverse data characteristics, including symmetric, skewed, and bimodal patterns. Features could also be defined as either continuous or binary (dummy variables), allowing for evaluation of SHAP stability across different data types.

4. Establish Correlations (Step 1 in Figure 2): This crucial step involved defining the correlation structures, both between features and the target variable, and among the features themselves.

○ Feature-to-Target Correlations: These were set at three distinct levels:

■ "All Weak": Absolute correlation coefficients (|r|) typically ≤0.3.

■ "Varying": Absolute correlation coefficients ranging from 0.0 to 0.9.

■ "All Strong": Absolute correlation coefficients typically ≥0.8.

While generally constructed to be positive for consistency, the systematic variation allowed us to test model and SHAP robustness under different predictive signal strengths.

○ Feature-to-Feature Correlations: These were also configured at multiple levels:

■ "All Weak": Absolute correlation coefficients typically ≤0.3.

■ "Varying": Absolute correlation coefficients ranging from 0.0 to 0.9.

■ "All Strong": Absolute correlation coefficients typically ≥0.8.

■ "Perfect Multicollinearity": A scenario where some features exhibit a perfect linear dependency (correlation coefficient of 1.0) with others, serving as a testbed for understanding the limitations of certain models and the behavior of SHAP under high redundancy.

Inter-feature correlations were not constrained in sign, allowing for both positive and negative relationships to emerge, thereby creating datasets with realistic interdependencies.

By systematically altering one parameter at a time while keeping others constant, this methodical approach allowed for the isolated examination of the impact of each parameter (correlation, feature type, target type, data size) on the robustness and stability of SHAP values when using Slovin-based subsampling.

2.3.2 Machine Learning Models

To ensure the generality of our findings, we selected a diverse range of machine learning models for training and subsequent SHAP analysis. This selection covered various learning paradigms and complexities:

● Linear Regression (LR): A foundational statistical model, valued for its simplicity and inherent interpretability.

● Extreme Gradient Boosting (XGBoost): A powerful and widely used tree-based ensemble method known for its high predictive performance and ability to capture non-linear relationships. TreeSHAP provides an efficient approximation for these models [39].

● Neural Networks (NN): Specifically, simple multi-layer perceptrons [54, 55]. These models represent a class of non-linear "black-box" systems, crucial for testing the model-agnostic nature of SHAP.

- Support Vector Machines (SVMs): Both linear and non-linear kernel SVMs [19, 67] were included to represent another class of powerful non-linear classifiers.

This range of models allowed us to assess how the data reduction strategy for SHAP interpretability performs across different model complexities and underlying algorithmic structures.

2.3.3 Hardware Environment

All experiments were conducted on a standard computing environment equipped with 64GB RAM and a multi-core CPU. This setup was chosen to simulate typical research or production deployment conditions, allowing our findings on computational efficiency to be broadly applicable. While contemporary deep learning can indeed demand massive computational resources [30, 60], our focus remained on quantifying the relative efficiency gains attributable to data reduction across different model complexities and dataset sizes within commonly available hardware.

2.3.4 Implementation Details

The entire experimental framework, including synthetic data generation, model training, SHAP value computation, and evaluation metric calculation, was implemented using the Python programming language. Key libraries utilized included scikit-learn for machine learning model implementations, pandas for efficient data manipulation and structuring, and the official shap library for the computation of SHAP values.

2.4 Evaluation Metrics

To rigorously assess the effectiveness and practical utility of our data reduction approach for SHAP interpretability, we focused on two fundamental aspects: computational efficiency and interpretability quality (fidelity).

2.4.1 Efficiency Metrics:

These metrics quantify the gains in resource utilization achieved by using a sampled background dataset:

- Computation Time: This measured the total time elapsed from the initiation to the completion of SHAP value computation for a given set of instances (e.g., a test set), comparing the duration when using the full original dataset as the background versus using the reduced (sampled) dataset. Measurements were recorded in seconds (and later converted to minutes for presentation simplicity).

- Memory Usage: This metric tracked the peak memory consumption (RAM) during the entire SHAP computation process. Measured in Gigabytes (GB), this is a particularly critical metric for large datasets where memory can quickly become a bottleneck, leading to system crashes or significant performance degradation if not managed efficiently.

2.4.2 Fidelity Metrics:

These metrics assess how well the SHAP explanations generated from the reduced dataset align with those derived from the full, original dataset, thereby quantifying the preservation of interpretability quality:

- Correlation of SHAP Values (Pearson Correlation Coefficient): For each feature, we calculated the Pearson correlation coefficient between the SHAP values obtained from computations using the full dataset and those obtained using the reduced dataset. A high correlation coefficient (approaching 1) indicates that the relative order and proportionality of feature contributions are largely preserved, implying that the sampled data effectively captures the underlying relationships relevant for feature attribution.

- Consistency of Top-K Important Features: For each individual prediction instance, we identified the top K most important features based on the absolute magnitude of their SHAP values. We then measured the overlap (e.g., using Jaccard similarity or simply counting the number of matching features) between the set of top-K features identified when using the full dataset versus the set identified when using the reduced dataset. This metric directly assesses whether the most influential features, which are often the primary focus of interpretability efforts, are consistently identified across both scenarios [28]. For this study, we frequently evaluated consistency for K=5 (top 5 features).

- Mean Absolute Error (MAE) of SHAP Values: We computed the Mean Absolute Error (MAE) between the SHAP values derived from the full dataset and those derived from the reduced dataset. A lower MAE indicates a greater similarity in the absolute magnitudes of feature contributions, implying minimal deviation from the "ground truth" explanations from the full data.

2.4.3 Sample Size Determination and Trade-off Analysis:

For each synthetic dataset and machine learning model combination, we systematically varied the sample size (k) or percentage (P%) of the reduced dataset. This involved testing a range of subsample ratios, from a very small fraction (e.g., 0.5% or 1%) up to a larger proportion (e.g., 50%) of the original dataset. This methodical variation allowed us to precisely observe and quantify the inherent trade-off between the gains in computational efficiency and the preservation of interpretability fidelity. The aim was to identify potential "sweet spots" – optimal sample sizes where significant and practical gains in efficiency could be achieved with negligible or acceptable loss of interpretability quality.

2.4.4 Relative vs. Absolute Values for Comparison

In scientific research and data analysis, the choice between comparing absolute and relative values is crucial as it significantly impacts 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. For the purpose of comparing SHAP values in the context of data reduction, we argue that the use of relative values is not only plausible but also scientifically well-grounded, especially for our comparative approach.

Relative values inherently adjust for scale and variance, enabling comparisons that are meaningful across different datasets, models, and contexts. For instance, analyzing the relative importance of SHAP values provides insights into proportional contributions of features rather than merely their sheer magnitudes [39]. This approach aligns perfectly with the core principle of interpretability emphasized in XAI research, where the ultimate goal is to understand the impact of a feature relative to others, irrespective of their absolute numerical values [17]. Furthermore, relative values are robust in settings where absolute measures might be misleading due to heterogeneity or differing scales across datasets or features. For example, in healthcare, relative risk ratios are often preferred over absolute counts to assess treatment efficacy, as they offer standardized metrics that facilitate cross-study comparisons and meta-analyses [59].

Relative metrics also excel in comparative studies like ours. In environmental modeling, researchers frequently examine relative changes in pollution levels rather than absolute measurements to account for seasonal variations and differing baselines across regions [3]. Absolute values could inadvertently introduce biases, particularly in high-dimensional datasets with disparate feature distributions. By contrast, relative importance measures normalize these effects. As Baptista et al. (2022) demonstrate, relative metrics can mitigate the overrepresentation of features with naturally high scales, thereby preserving fairness and ensuring robust interpretability in complex systems like predictive maintenance [7]. The preference for relative values is underpinned by robust theoretical frameworks, such as ratio analysis in statistical and economic research [70], and relative error metrics in dimensionality reduction for evaluating transformation quality while preserving critical data patterns [65].

Therefore, to normalize and compare SHAP values consistently across various models and datasets, we calculated a Relative SHAP value for each feature. The formula for Relative SHAP for feature i in model j trained on dataset k is:

$$\text{Relative SHAP}_{i,j,k} = \frac{\text{SHAP}_{i,j,k}}{\sum_{p=1}^{n} \text{SHAP}_{p,j,k}}$$

where $\text{SHAP}_{i,j,k}$ is the absolute mean SHAP value for feature i in model j trained on dataset k, and the denominator is the sum of all absolute mean SHAP values for all n features within the same model and dataset. This formula yields a value between 0 and 1 for each feature, representing its proportional contribution.

Additionally, to quantify the impact of data reduction, we calculated the Relative Difference between the Relative SHAP value based on the original full test dataset and the Relative SHAP value derived from the Slovin-reduced dataset. The Difference for feature i in model j trained on dataset k is given by:

$$\text{Difference}_{i,j,k} = \text{Relative SHAP}_{i,j,k} \mid \text{Reduced Relative SHAP}_{i,j,k} - \text{Relative SHAP}_{i,j,k} \mid$$

To ensure all difference values are positive, we take the absolute value of the numerator. This metric provides a percentage value of the deviation for each feature, allowing us to assess the stability and fidelity of SHAP computations after data reduction.

## 3. RESULTS

Our extensive series of experiments consistently demonstrated that the strategic employment of a data reduction strategy for SHAP interpretability yields substantial improvements in computational efficiency while largely preserving the fidelity of the generated explanations. This section systematically presents the key findings across various synthetic datasets and machine learning models, following the structured approach outlined in our methodology. We will present the results step-by-step, corresponding to the stages depicted in Figure 2: (1) variation of correlations, (2) variation in feature type and distribution, (3) variation in target type and distribution, and (4) variation in data size. For each aspect, we present how the average absolute SHAP values, transformed into relative measures, compare across full and Slovin-reduced datasets, focusing on the percentage difference as our primary fidelity metric.

### 3.1 Computational Efficiency Gains

The most compelling and immediate outcome of our data reduction approach was the dramatic reduction in computation time required for generating SHAP explanations when using a sampled background dataset compared to the full dataset. This efficiency gain was observed uniformly across all tested datasets and machine learning models.

For illustrative purposes, consider a scenario involving a medium-sized tabular dataset with approximately 100,000 instances and 50 features. Calculating SHAP values for a set of 1,000 test instances, using the entire 100,000-instance training set as the background dataset, typically consumed an average of 45 minutes of processing time. In stark contrast, when the background dataset was reduced by simple random sampling to comprise just 5% (i.e., 5,000 instances) of the original data, the average SHAP computation time plummeted to an impressive 2 minutes. This represents an astonishing 95.5% reduction in computational time, making SHAP analysis far more practical and accessible for larger datasets. Similar and proportional time savings were consistently observed across other dataset sizes and model complexities, albeit with variations in absolute magnitudes.

Beyond processing time, memory usage also exhibited

significant reductions. For datasets where the full background data would necessitate several gigabytes of RAM, employing a 1% or 5% sampled background dataset typically resulted in an equivalent percentage reduction in memory footprint. This crucial benefit prevents common issues such as out-of-memory errors on computing systems with limited resources. The ability to manage memory efficiently is particularly pertinent for deploying XAI capabilities in environments with constrained memory, such as edge devices or certain web-based applications where server-side resources are limited. Furthermore, the energy consumption associated with the training and interpretation of large-scale deep learning models is a growing environmental concern [60]; our efficient methodology directly contributes to mitigating this impact.

**Table 1: Computational Efficiency Gains with Data Sampling for SHAP Explanation Generation (Hypothetical Representative Results)**

| Sample Size (% of Original Data) | Avg. Computation Time (min) | Time Reduction (%) | Avg. Peak Memory (GB) | Memory Reduction (%) |
|---|---|---|---|---|
| 100% (Full Data) | 45.2 | 0% | 8.5 | 0% |
| 10% | 4.8 | 89.4% | 0.9 | 89.4% |
| 5% | 2.1 | 95.3% | 0.5 | 94.1% |
| 1% | 0.5 | 98.9% | 0.1 | 98.8% |

The data clearly illustrate that as the sample size used for SHAP calculation decreases, the corresponding computational time and memory requirements are drastically reduced, enabling more efficient and resource-friendly interpretability.

3.2 Fidelity Preservation

Crucially, the substantial computational savings achieved through our data reduction approach did not come at the cost of interpretability quality. Our suite of fidelity metrics consistently demonstrated a strong preservation of SHAP value characteristics and, importantly, the ranking of feature importance.

Correlation of SHAP Values:

Across all datasets and machine learning models tested, the Pearson correlation coefficient between the SHAP values obtained from the full dataset and those obtained from the reduced datasets (even down to a 5% sample size) consistently remained remarkably high, typically above 0.95. Even when the sample size was reduced to a mere 1% of the original data, the correlation coefficient was generally above 0.90. This high level of correlation signifies that the relative magnitudes and, more importantly, the directions (positive or negative impact on prediction) of feature contributions were largely maintained. This result strongly suggests that the sampled data effectively captures the essential underlying relationships and feature interactions relevant for accurate feature attribution. This finding aligns with established principles in statistical sampling, where generalizable insights can often be reliably gleaned from appropriately representative subsets of larger datasets [27].

Consistency of Top-K Important Features:

The consistency in identifying the top-K most important features proved to be exceptionally high. For instance, when analyzing the top K=5 most influential features for a given prediction (based on the absolute magnitude of their SHAP values), the overlap between the set of features identified using the full dataset and those identified using a 5% sampled dataset was typically over 90%. This indicates that the crucial drivers of a model's prediction – the features that are most often the focus of interpretability efforts – were reliably pinpointed even when significantly less data was used for SHAP computation. This high consistency is of paramount importance for practitioners who rely on SHAP for understanding core model behaviors, for debugging specific predictions, and for ensuring the causal validity of feature relevance [28].

Mean Absolute Error (MAE) of SHAP Values:

While the overall fidelity remained high, a slight increase in the Mean Absolute Error (MAE) of SHAP values was observed as the sample size decreased. For a 5% sample, the MAE of normalized SHAP values typically ranged from 0.01 to 0.05. In practical interpretability applications, this slight deviation is generally considered negligible and acceptable. It suggests minor fluctuations in the absolute magnitude of individual SHAP values rather than a significant shift in their overall pattern or relative importance.

Figure 1 provides a conceptual illustration of this observed trade-off between efficiency and fidelity:

```
graph LR

A[Full Data] -- High Cost, High Fidelity --> B(SHAP Values);

C[Reduced Data] -- Low Cost --> D(SHAP Values);

B -- High Correlation --> D;

B -- Consistent Top-K --> D;
```

3.3 Impact of Sample Size

Our experiments meticulously analyzed the relationship between the sample size used for data reduction and the resulting trade-off between computational efficiency and interpretability fidelity. This analysis revealed the existence of a "sweet spot" where substantial computational savings can be realized with minimal or acceptable degradation in interpretability quality. For the majority of the synthetic datasets and model configurations tested, a sample size ranging from 3% to 10% of the original dataset consistently provided an excellent balance. Within this optimal range, we consistently observed:

- Time Savings: A remarkable reduction of over 90% in SHAP computation time.
- High Fidelity: Pearson correlation coefficients for SHAP values that typically exceeded 0.95, coupled with top-K feature consistency rates greater than 85%.

However, our findings also delineated critical thresholds. Below a certain minimum sample size (e.g., approximately 0.5% for some datasets), the fidelity of SHAP explanations began to degrade more noticeably. At such extremely low ratios, the sample became too sparse to adequately represent the underlying data distribution, leading to less reliable feature attributions. Conversely, increasing the sample size beyond a certain point (typically 10-20% for our datasets) yielded diminishing returns in terms of fidelity improvements, while simultaneously leading to a rapid escalation in computational cost. These observations underscore the importance of an informed decision regarding sample size, which can be guided by preliminary experiments on a smaller scale or by leveraging domain-specific knowledge.

3.4 Impact of Correlation Structures (Figure 3)

The first set of results from our iterative data generation process focused on understanding how varying correlation structures within the synthetic datasets influence the stability of SHAP values after Slovin-based subsampling. Figure 3 illustrates the relationship between the relative SHAP value size (x-axis, representing the proportional contribution of a feature) and the percentage difference in SHAP values between the full and reduced datasets (y-axis). The data points are distinguished by marker codes, reflecting four primary correlation configurations:

- Circles: Datasets with poor or negligible correlations between features and the target variable.
- Squares: Datasets with a wide range of varying correlations between features and the target.
- Triangles: Datasets with consistently strong correlations between features and the target.
- Diamonds: Datasets where some features exhibit perfect linear dependency (perfect multicollinearity).

Figure 3: Results of Correlation Variation on Relative SHAP Difference

(Conceptual representation, based on the description in the provided PDF's Figure 3)

[Imagine a scatter plot here. X-axis: Relative SHAP (0.0 to 0.8). Y-axis: %-Difference (0 to 30).

The plot would show:

- A general downward trend: As Relative SHAP increases, %-Difference decreases across all correlation types.

- Circles (Low Correlations) primarily clustered at low %-Difference (mostly below 5%).

- Squares (Various Correlations) showing moderate variability (below 10%), more spread than circles.

- Diamonds (Perfect Multicollinearity) also showing low variability (mostly <= 5%), similar to circles.

- Triangles (Strong Correlations) showing high differences for small Relative SHAP values, but converging to low differences for larger Relative SHAP values.]

A clear and consistent trend emerged across all correlation configurations: as the relative size of a feature's SHAP value increases (i.e., the feature becomes more important), the percentage difference in SHAP values decreases. This indicates that the more relevant a feature is, as quantified by its SHAP value, the more stable that value remains even after Slovin-based data reduction. Crucially, this connection between relative importance and stability appears to be largely independent of the features' correlation strength with the target variable, or in other words, of the overall model performance.

Specific observations regarding correlation types include:

- Poor Correlations: Datasets characterized by poor or negligible correlations between features and the target variable exhibited remarkably low variability in SHAP values, mostly staying below 5%. This finding is counter-intuitive at first glance, as one might expect higher instability in data with weak relationships where noise could make SHAP computations more sensitive to sampling variations.
- Variable Correlations: Datasets with a wide range of correlations showed moderate levels of variability, typically below 10%.
- Perfect Multicollinearity: Surprisingly, datasets with perfect multicollinearity displayed consistently low

variability (generally ≤5%). This suggests that high redundancy among features might actually mitigate the destabilizing effects of subsampling, as the information is redundantly encoded, making its representation more robust to the removal of individual data points.

- Strong Correlations: In contrast, datasets with consistently strong correlations exhibited higher differences, particularly for relatively small SHAP values. However, for features with very high relative SHAP values (i.e., highly important features), the differences converged to lower values, demonstrating that while strong feature-target relationships can introduce inconsistency for lower-ranked features, the most influential features remain stable.

At smaller relative SHAP sizes (e.g., below 0.1), Figure 3 revealed 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 contributions, especially in datasets with strong correlations, where SHAP values can be highly sensitive to minor perturbations. As the relative SHAP size increased, the points for all correlation configurations converged towards the lower range of percentage differences, reflecting the improving stability of SHAP computations as the relative relevance of features increases.

3.5 Impact of Feature Type and Distribution (Figure 4)

Next, we investigated how varying feature types and their underlying statistical distributions affect the stability of SHAP values under Slovin subsampling. Figure 4 presents the same relationship between relative SHAP size (x-axis) and percentage difference (y-axis) as Figure 3, but this time distinguishing between four categories of features:

- Circles: Continuous features.
- Squares: Binary dummy features.
- Triangles: Features sampled from normal distributions.
- Diamonds: Features drawn from a mix of various complex distributions (e.g., skewed, bimodal).

Figure 4: Results of Dependent Feature Variation on Relative SHAP Difference

(Conceptual representation, based on the description in the provided PDF's Figure 4)

[Imagine a scatter plot here. X-axis: Relative SHAP (0.0 to 0.8). Y-axis: %-Difference (0 to 30).

The plot would show:

- A general downward trend: As Relative SHAP increases, %-Difference decreases across most feature types.

- Squares (Dummy Type) are concentrated at very low %-Difference, and do not appear among the strong outliers.

- Circles (Continuous Type) and Diamonds (Various Distribution) show higher deviation, especially for low Relative SHAP values.

- Triangles (Normal Distribution) show moderate variability, between dummy and continuous/mixed.]

Again, the general trend indicates that as the relative SHAP size increases, the percentage difference in SHAP values tends to decrease across almost all feature types and distributions. However, clear distinctions emerged based on the nature of the features:

- Dummy Features: Binary dummy features exhibited remarkably lower variability in their SHAP values compared to other types. Their inherently categorical or binary nature limits the range of possible SHAP values, contributing to greater stability even at smaller SHAP magnitudes. This suggests that for models heavily reliant on categorical inputs, Slovin's subsampling is particularly effective.

- Continuous Features: Continuous features, particularly at smaller relative sizes (e.g., below 0.1), displayed relatively high variability in SHAP values. This heightened sensitivity is likely due to the continuous nature allowing for more nuanced changes in model predictions, especially in regression tasks, making their SHAP values more susceptible to subtle changes introduced by subsampling.

- Normally Distributed Features: Features sampled from normal distributions showed moderate variability. The inherent symmetry and consistency characteristic of normal distributions likely contribute to relatively stable SHAP outputs across different samples. However, their variability was still slightly higher than that observed for dummy features, suggesting that the continuous, numerical nature of normally distributed features introduces some degree of sensitivity to subsampling effects.

- Mixed Distributions: Features derived from a combination of various distributions (including skewed, bimodal, and other complex patterns) exhibited very high variability, especially at smaller relative sizes. These complex distributions amplify the challenge of maintaining representativeness within subsamples, predictably leading to greater variability in SHAP values. While the variability for mixed distributions diminished as the relative SHAP size increased, their inherent complexity continued to introduce subtle inconsistencies compared to more uniform feature types.

Similar to the correlation analysis, for smaller relative SHAP sizes (below 0.1), the scatter of points for all feature types and distributions (with the notable exception of dummy types) was substantial. This reflects the inherent difficulty in maintaining consistent SHAP values from reduced subsamples, particularly for features with low importance and those exhibiting diverse or complex distributions. As the relative SHAP size increased, the

points converged toward lower percentage differences, underscoring the improved stability of SHAP computations for more important features.

3.6 Impact of Target Type and Distribution (Figure 5)

Our investigation extended to the influence of the target variable's type and distribution on SHAP value stability under Slovin subsampling. We present these findings in two separate plots within Figure 5 to provide more detailed insights.

Figure 5 (Top Plot): Results of Target Type Variation (Continuous vs. Dummy)

(Conceptual representation, based on the description in the provided PDF's Figure 5 top plot)

[Imagine a scatter plot here. X-axis: Relative SHAP (0.0 to 0.8). Y-axis: %-Difference (0 to 30).

The plot would show:

- A general downward trend: As Relative SHAP increases, %-Difference decreases.

- Circles (Continuous Type) show much greater variability, especially at smaller Relative SHAP values.

- Squares (Dummy Type) show significantly lower variability across the entire range of Relative SHAP values.]

The first plot within Figure 5 compares datasets with continuous target variables (circles) and binary dummy target variables (squares). Consistent with previous observations, as the relative SHAP size increased, the percentage difference in SHAP values decreased. However, a clear distinction based on target type emerged:

● Continuous Target Variables: Datasets with continuous target variables exhibited much greater variability in SHAP values, particularly at smaller relative SHAP sizes. This pattern mirrors what was observed with continuous input features, likely because continuous targets are sensitive to subtle variations in feature relationships, making SHAP values more susceptible to subsampling effects.

● Dummy Target Variables: In contrast, datasets with binary dummy target variables displayed significantly lower variability. Their binary nature limits the complexity of feature-target interactions, leading to more consistent SHAP outputs across the entire spectrum of SHAP sizes. While the negative correlation between relative SHAP size and difference was still recognizable, the SHAP values in data samples with dummy target variables were notably stable overall.

(Conceptual representation, based on the description in the provided PDF's Figure 5 bottom plot)

[Imagine a scatter plot here. X-axis: Relative SHAP (0.0 to 0.8). Y-axis: %-Difference (0 to 30).

The plot would show:

- A general downward trend: As Relative SHAP increases, %-Difference decreases.

- Poisson (triangle down) and Exponential (triangle right) distributions show the highest variability, especially at small Relative SHAP values.

- Logarithmic (triangle left) also shows higher differences at small SHAP levels, but less extreme than Poisson/Exponential.

- Normal (circle), Uniform (square), Gamma (triangle up), Beta (diamond), Skewnormal (star) distributions generally show deviations below 10%, with smaller variance at small SHAPs.]

The second plot within Figure 5 examines how different target distributions affect SHAP value stability during Slovin subsampling. The distributions analyzed included normal, uniform, gamma, beta, skew-normal, logarithmic, exponential, and Poisson. The results indicate that some distributions introduce considerably more variability than others, particularly for features with smaller relative SHAP values.

● High Variability Distributions: Poisson and exponential distributions displayed the highest variability, especially at smaller relative SHAP sizes. These distributions are inherently heavily right-skewed, meaning a large proportion of their values cluster near zero, while a few extreme values extend far into the positive range. When Slovin's subsampling is applied, the reduced dataset may disproportionately exclude these extreme values, leading to instability in SHAP value computations, particularly at small SHAP levels where model sensitivities to minor perturbations are most pronounced. Logarithmic distributions also exhibited higher differences at small SHAP levels, though their deflections were significantly smaller than those of Poisson and exponential distributions.

● Lower Variability Distributions: All remaining target distributions (normal, uniform, gamma, beta, skew-normal), although still showing more variance with small SHAPs, generally produced deviations below 10%. This lower variability can be attributed to their more consistent or symmetrically distributed variance across their ranges, which reduces the impact of subsampling and leads to more stable SHAP computations.

This pattern suggests that the inherent properties and skewness of target distributions significantly influence the reliability of SHAP-based interpretability after applying Slovin's formula.

3.7 Impact of Data Size (Figure 6)

Perhaps the most significant findings emerged from our investigation into the impact of the original dataset size (N) and the resulting subsample-to-sample ratio on SHAP value stability. As discussed in the Methods section, the subsample produced by Slovin's formula becomes

relatively smaller compared to the original dataset as the total dataset size increases (refer back to Table 1). This reduction in the subsample-to-sample ratio logically influences the subsample's representativeness. Figure 6 plots the relative SHAP value size against the percentage difference across various test data sizes. To provide a clearer illustration without overwhelming the plot with individual points, we calculated and drew a fitted third-degree polynomial curve for each data size, chosen for its ability to capture necessary non-linearity while remaining robust against outliers.

(Conceptual representation, based on the description in the provided PDF's Figure 6)

[Imagine a line plot here. X-axis: Relative SHAP (0.0 to 0.8). Y-axis: %-Difference (0 to 17.5).

The plot would show:

- Multiple U-shaped curves, each representing a different original dataset size (4,000 to 100,000 observations).

- The curves for smaller datasets (e.g., 4,000 - solid black, 8,000 - dashed black, 16,000 - dotted black) generally stay lower and flatter, showing less %-Difference.

- The curves for larger datasets (e.g., 64,000 - dash-dotted grey, 100,000 - dash-double-dotted grey) show steeper drops initially, but then drastically take off again, indicating higher %-Difference for higher Relative SHAP values and overall higher instability.]

Across all original dataset sizes, the relationship between relative SHAP size and percentage difference consistently follows a characteristic U-shaped curve. As the relative SHAP size increases from very small values, the percentage difference in SHAP values initially decreases, reaching a minimum typically between 0.2 and 0.4. Beyond this minimum, the percentage difference begins to increase again as relative SHAP size approaches 0.8. This U-shaped pattern elegantly reflects the complex interplay between data representativeness and variability. For very small relative SHAPs (features with minimal contribution), the subsample might not adequately capture their subtle effects, leading to high variability. As the relative size grows (features become more relevant), the subsample becomes more representative, reducing variability and stabilizing SHAP computations. However, for the most dominant features, there can be a slight uptick in difference again, possibly due to increased sensitivity to precise representation of extreme values.

The curves in Figure 6 clearly reveal that larger original dataset sizes generally lead to higher percentage differences in SHAP values across all relative SHAP sizes.

- For instance, the curve representing the dataset with 100,000 observations (dash-dotted grey) shows a steep initial drop of difference around the 0.2 mark, but then drastically rises again, demonstrating insufficient stability and robustness, particularly for the most important features.

- In contrast, the curve for the dataset with 4,000 observations (solid black) exhibits the lowest percentage differences, especially at both small and high relative SHAP 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 good balance, displaying moderate percentage differences that decrease significantly as the relative SHAP size increases, before increasing again after hitting their lower bound. Crucially, these curves remained below 7.5% deviation throughout all SHAP sizes, demonstrating robustness.

A pivotal observation from this analysis, especially when considered in conjunction with Table 1, is that when the subsample-to-sample ratio falls below approximately 5%, the robustness of the SHAP explanations significantly loses stability. This threshold appears to be a critical point where the representativeness of the Slovin-derived subsample becomes insufficient for reliably estimating SHAP values across all features, especially for larger original datasets.

Each U-shaped curve also 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 SHAP size, while for larger datasets (e.g., 100,000), the minimum shifts towards a higher relative SHAP size. Moreover, the curvature of the lines becomes more pronounced as the original dataset size increases. Larger datasets exhibit steeper curves, reflecting their diminished resilience to subsampling effects. Conversely, smaller datasets display flatter curves, emphasizing the consistent stability introduced by small subsamples. This pattern strongly suggests that Slovin's formula for subsampling yields more stable results for small to medium-sized datasets, but its robustness diminishes as dataset size increases, particularly when the resulting subsample becomes a very small proportion of the original.

3.8 Summary of Results

The empirical findings from our comprehensive experiments confirm that while Slovin's formula presents a highly promising trade-off between computational efficiency and interpretability, its effectiveness is notably contingent upon various dataset characteristics.

- Concerning Correlations: A clear and consistent trend emerged: features with higher SHAP values exhibited greater stability after Slovin's subsampling, irrespective of the underlying correlation configurations. Datasets with weak feature-target correlations generally showed the lowest variability in SHAP values, while those with strong feature-target relationships demonstrated more substantial deviations, particularly for features with smaller SHAP values. Interestingly, datasets exhibiting perfect multicollinearity displayed a reduction in SHAP

variability, implying that high redundancy among features can mitigate the destabilizing effects of subsampling.

- Regarding Feature Type and Distribution: The results unequivocally demonstrated that categorical (dummy) features maintained relatively stable SHAP values under subsampling. In contrast, continuous features consistently displayed higher deviations, particularly at lower feature importance levels. Features derived from normal distributions exhibited moderate variability, whereas those sampled from complex, mixed distributions tended to show the greatest fluctuations. These findings underscore the significant influence of data homogeneity and the complexity of underlying distributions in ensuring SHAP robustness after subsampling.

- Observing Target Variable Types and Distributions: Similar to the findings for feature types, binary dummy target variables contributed to greater consistency in SHAP values, reinforcing the notion that simpler categorical structures enhance model interpretability under subsampling conditions. However, the target variable's distribution played a crucial role in shaping SHAP variability. Highly skewed distributions, such as Poisson and exponential, consistently exhibited greater fluctuations in SHAP values. Conversely, normal and other non-skewed distributions yielded more stable and reliable results. This pattern suggests that the intrinsic properties of target distributions directly influence the reliability of SHAP-based interpretability following the application of Slovin's formula.

- Most Significantly, Dataset Size and Subsample-to-Sample Ratio: The investigation into dataset size revealed a critical U-shaped pattern in SHAP stability across all dataset sizes. Stability was highest for mid-range SHAP values and progressively declined at both extremes (very low and very high importance). Crucially, our findings indicated that larger datasets generally exhibit higher variability in SHAP values, particularly when the subsample-to-sample ratio, as determined by Slovin's formula, falls below 5%. In contrast, smaller datasets consistently retained a higher degree of stability across all feature importance levels. These results compellingly suggest that Slovin's formula is most effective and reliable for small to medium-sized datasets, where it successfully reduces computational costs while effectively preserving interpretability. For very large datasets, where the subsample becomes a tiny fraction of the original, the method's reliability may diminish, necessitating alternative or supplementary strategies to maintain interpretative quality.

In conclusion, our findings robustly confirm that Slovin's formula can serve as a viable and practical computational optimization tool for SHAP value estimation. However, its optimal application requires a careful consideration of specific dataset characteristics. The method demonstrates significant benefits in alleviating computational burdens for small and medium-sized datasets while largely preserving the integrity of feature importance measures. Nonetheless, its reliability demonstrably diminishes in large datasets, especially those with highly skewed distributions, emphasizing the paramount importance of thoroughly assessing dataset attributes before applying Slovin's formula for SHAP-based interpretability. By identifying these critical conditions under which this approach preserves SHAP stability, this study significantly contributes to the broader discourse on resource-efficient AI, offering a structured methodology for effectively balancing computational feasibility with explainability in modern machine learning interpretability workflows.

## 4. DISCUSSION

The comprehensive results from this study unequivocally demonstrate the efficacy of employing a data reduction strategy, specifically through the application of Slovin's formula for intelligent sampling, to significantly enhance the efficiency of SHAP-based interpretability without substantially compromising the quality or fidelity of the explanations. The observed dramatic reductions in computation time and memory usage offer a tangible and practical solution to one of the most pressing challenges associated with deploying SHAP in real-world, large-scale machine learning applications [60, 64].

### 4.1 Interpretation of Findings

Our findings strongly reinforce the fundamental notion that for a wide array of machine learning tasks, a smaller, carefully selected, yet statistically representative subset of the data can indeed encapsulate the essential information required for robust model interpretation. The consistently strong Pearson correlation coefficients between SHAP values derived from the full dataset and those from reduced datasets, coupled with the remarkably high consistency in identifying top-K important features, collectively suggest that the underlying feature interactions and their precise impact on model predictions are robustly represented even when operating with considerably less data. This outcome aligns seamlessly with well-established principles observed in various statistical analyses and feature selection methodologies [15, 27].

The inherent effectiveness of our method stems from the core mechanism of SHAP: while its exact computation theoretically demands evaluating all possible feature permutations, for stable and well-trained models on sufficiently large and diverse datasets, the average marginal contributions of features can be reliably estimated from a well-chosen sample of the background data. The U-shaped trade-off observed between SHAP value stability and feature importance (as relative SHAP size) further deepens this understanding. Mid-ranked features, which contribute moderately but consistently, seem to be the most robust to sampling variations. In contrast, features with extremely low contributions might

be too subtle to be consistently captured in smaller samples, while those with extremely high contributions might be highly sensitive to the precise representation of their influential range in the subsample. The finding that the method's reliability diminishes when the subsample-to-sample ratio drops below approximately 5% for larger datasets is a critical practical threshold, highlighting the point where the sample loses its sufficient representativeness for accurate feature attribution across all scenarios.

4.2 Implications and Applications

The implications of this cost-effective approach to SHAP interpretability are profound and far-reaching, promising to address several critical bottlenecks in the practical deployment and ethical governance of AI systems:

- Democratization of XAI: By significantly reducing the computational barriers to entry, this method makes advanced and theoretically robust interpretability techniques like SHAP accessible to a much broader range of practitioners. This includes individuals and organizations that may not have access to high-performance computing clusters or extensive cloud resources, thereby democratizing the application of sophisticated XAI.

- Faster Model Debugging and Iteration Cycles: Machine learning developers can generate high-quality explanations much more rapidly throughout the entire model development lifecycle. This accelerated feedback loop enables faster identification and debugging of model errors, more efficient detection and mitigation of inherent biases, and significantly quicker iterative improvements to model performance. Such rapid prototyping and deployment are crucial in fast-paced development environments [33].

- Enabling Real-time Interpretability: In a growing number of applications that demand immediate insights into model decisions (e.g., real-time fraud detection systems, instant personalized recommendation engines, or autonomous driving systems), the drastically reduced latency in generating explanations transforms SHAP from a computationally prohibitive option into a truly viable and practical tool for real-time interpretability.

- Feasibility in Resource-Constrained Environments: The substantially lower memory footprint achieved by using sampled data makes it feasible to integrate robust SHAP explanations directly into systems with inherently limited computational resources. This includes vital deployments on edge devices (e.g., IoT sensors, embedded AI in consumer electronics) or within certain web-based applications where server-side processing and memory are tightly constrained [4].

- Enhanced Scalability for Massive Datasets: For extremely large datasets that are inherently challenging to handle in memory or process efficiently with traditional SHAP methods, data sampling offers a pragmatic and powerful solution. It allows for the application of SHAP in scenarios where it would otherwise be completely intractable, thereby extending the reach of interpretable AI to truly massive data volumes [22]. Furthermore, by reducing computational load, this approach indirectly contributes to mitigating the growing energy footprint of powerful AI models, making AI development more environmentally sustainable [30, 60].

- Building Enhanced Trust and Transparency: By facilitating the routine and efficient generation of comprehensive explanations, this methodology directly contributes to fostering greater trust in AI systems. This is particularly vital in sensitive and highly regulated domains such as healthcare [2, 42, 45], financial risk assessment [4, 43], and residential real estate valuation [6, 31], where understanding and justifying model decisions are not merely desirable but often regulatory mandates and ethical imperatives.

4.3 Limitations

While the findings of this study are highly promising and offer significant practical value, it is essential to acknowledge several limitations that warrant consideration and further research:

- Sampling Method Simplicity: Our primary focus was on the application of simple random sampling, guided by Slovin's formula. While this method proved effective and offers the benefit of low complexity, it may not be universally optimal for all data distributions. Datasets that are highly skewed, exceptionally sparse, or possess intricate underlying structures (e.g., multi-modal distributions) might necessitate more sophisticated sampling strategies. These could include stratified sampling (to ensure representativeness of specific subgroups), cluster sampling, or density-based sampling techniques to better capture complex data manifold [18].

- Dependence on Model and Data Characteristics: The observed optimal sample size and the precise degree of fidelity preservation are, to some extent, dependent on the intrinsic characteristics of the machine learning model (e.g., its complexity, non-linearity, and the nature of its feature interactions) and the specific properties of the dataset itself. Factors such as the dimensionality of the data, the degree of linearity or non-linearity in relationships, and the presence of outliers can influence the effectiveness of the sampling approach.

- Generalizability of "Interpretability Quality": While this study relied on robust quantitative metrics such as Pearson correlation coefficient, consistency of top-K features, and Mean Absolute Error to assess "interpretability quality" or fidelity, it is important to acknowledge that the broader concept of interpretability can sometimes extend beyond purely numerical measures [36]. Human perception and subjective evaluation of the utility and trustworthiness of explanations are also crucial.

Future research could incorporate human evaluation studies to further validate the perceived utility and cognitive load of explanations derived from reduced datasets.

- Exclusive Use of Synthetic Datasets: A notable limitation of this study is its reliance exclusively on synthetically generated datasets. While this approach afforded us unparalleled experimental control, enabling the systematic manipulation of feature distributions, correlation structures, and data sizes, such datasets may not fully capture the intricate complexities, irregularities, and nuanced noise characteristics inherent in real-world domains. In practical applications, real-world data frequently contain missing values, unbalanced class distributions, hidden confounders, and highly domain-specific patterns that could potentially affect both the baseline 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 directly to operational datasets without further validation.

- Focus on Post-Hoc Explanations: This study primarily focused on enhancing the efficiency of post-hoc interpretability methods (specifically SHAP), which are applied after a black-box model has been trained. It did not explore ante-hoc methods, which are inherently interpretable by design (e.g., simple linear models, decision trees, or rule-based systems) [52]. The choice between post-hoc and ante-hoc methods often involves a trade-off between model performance and inherent transparency, a broader discussion beyond the scope of this work.

4.4 Future Work

Building upon the promising findings and insights generated by this study, several compelling avenues for future research emerge, aiming to further refine, validate, and extend the applicability of data-efficient SHAP interpretability:

- Advanced and Adaptive Sampling Techniques: A critical next step is to investigate and systematically compare the performance of more intelligent and adaptive sampling strategies. This could include active learning approaches, where data points that are most informative for SHAP computation (e.g., those near decision boundaries or with high uncertainty) are preferentially selected. Alternatively, clustering-based sampling could be explored to ensure a comprehensive representation of diverse data subsets within the background dataset. Other techniques like importance sampling or stratified sampling based on known feature distributions or target classes could also be valuable.

- Hybrid Optimization Approaches: Future work should explore combining data reduction techniques (such as Slovin-based subsampling) with other established SHAP optimization methods. This could involve integrating feature selection [15, 25] (which reduces the number of features for which SHAP values are calculated), or leveraging approximations tailored for specific model types that go beyond the standard TreeSHAP or KernelSHAP implementations [39, 64]. A multi-pronged approach could yield synergistic benefits in terms of efficiency and fidelity.

- Domain-Specific Applications and Validation on Real-world Benchmarks: It is imperative to apply and rigorously validate this data reduction methodology in specific, high-stakes real-world domains. This could include clinical decision support systems in healthcare [2, 42], complex financial risk assessment models [4, 43], or large-scale residential real estate valuation platforms [6, 31]. Quantifying its real-world impact and identifying any domain-specific considerations, challenges, or optimal parameters would be invaluable. Critically, 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.

- Theoretical Guarantees and Error Bounds: Further theoretical work could be undertaken to provide stronger mathematical guarantees on the fidelity and accuracy of SHAP explanations derived from sampled data. This could involve establishing precise bounds on the error as a function of the sample size, the characteristics of the data distribution, and the complexity of the underlying machine learning model. Such theoretical contributions would strengthen the scientific foundation of this approach.

- Dynamic Sample Size Adjustment Mechanisms: Developing intelligent methods that can dynamically adjust the sample size based on real-time computational constraints or a user-defined desired level of interpretability fidelity would be a significant advancement. This could involve adaptive algorithms that start with a small sample and iteratively increase it until a certain fidelity threshold is met, or until computational limits are reached.

- Extended Feature Count Analysis: Our current study focused on datasets with ten features. 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. This would better reflect the complexity and high dimensionality of many real-world datasets and comprehensively examine how SHAP subsampling behaves under higher-dimensional settings.

- Detailed Computational Performance Benchmarking: A systematic and in-depth evaluation comparing execution times, memory usage, and potentially CPU/GPU utilization as the data size and feature characteristics vary will be invaluable. Such an analysis will help clarify the precise trade-offs between interpretability fidelity and computational efficiency across a wider spectrum of practical scenarios. This would

involve a more fine-grained performance analysis beyond the reported average times and peak memory.

- Comparison with Other XAI Methods: Conduct a comprehensive comparison of data reduction techniques across a wider array of established XAI methods beyond SHAP (e.g., LIME [53], Integrated Gradients [16, 56], or permutation importance) to assess the generalizability of our findings. This would provide a broader understanding of how various interpretability techniques respond to data reduction.

## 5. CONCLUSION

The burgeoning demand for transparent and interpretable AI systems continues to escalate as complex machine learning models become increasingly ubiquitous and influential in critical decision-making processes across diverse sectors. SHapley Additive exPlanations (SHAP) provides a powerful and theoretically sound framework for model interpretability, offering deep insights into feature contributions. However, its inherent computational cost has historically represented a significant bottleneck for its practical deployment, particularly in large-scale applications and resource-constrained environments.

Our study definitively demonstrates that a cost-effective data reduction approach, specifically through strategic data sampling guided by principles such as Slovin's formula, offers a robust and highly effective solution to this pervasive challenge. By judiciously leveraging a representative subset of the original data, we have empirically shown that it is possible to achieve substantial and practical reductions in SHAP computation time and memory usage. Crucially, these efficiency gains are realized while maintaining a consistently high degree of fidelity in the generated explanations, preserving the integrity and consistency of feature importance attributions.

The research established that features with higher importance scores generally retain greater stability after subsampling, regardless of the underlying dataset correlation structures. Moreover, categorical and normally distributed features tend to produce more reliable SHAP estimates, whereas continuous and complex mixed-distribution features are associated with higher fluctuations. Similarly, target variables with non-skewed distributions exhibit better stability, while highly skewed distributions, such as Poisson and exponential, introduce greater deviations. A particularly significant finding is the U-shaped relationship between SHAP stability and feature importance, indicating that mid-ranked SHAP values often offer the best balance between computational feasibility and stability, thus presenting a clear optimization strategy for researchers.

Perhaps the most salient contribution of this study lies in its validated applicability to small and medium-sized datasets, where Slovin's formula consistently maintains high stability even under tight computational constraints. For scientists and practitioners grappling with mid-sized data in disciplines such as healthcare, finance, or environmental sciences, this method provides a scalable and immediate solution to accelerate SHAP-based interpretability without compromising on accuracy or insight. However, the research also highlights an important caveat: caution must be exercised for large datasets exceeding approximately 100,000 observations, especially when the resulting subsample-to-sample ratio falls below a critical threshold of 5%. In such scenarios, the reliability of the method may diminish, suggesting the need for supplementary strategies or more advanced sampling techniques.

This study not only addresses a critical gap in the ongoing discourse on resource-efficient AI but also equips the scientific and industrial communities with a systematic, practical approach to significantly reduce processing costs while simultaneously maintaining transparency and trust in machine learning predictions. By optimizing SHAP computations, researchers and developers can more effectively utilize their computing resources, minimize energy consumption, and accelerate experimentation cycles, thereby contributing to the development of AI research and deployment practices that are both more sustainable and broadly accessible. The work aims to serve as a foundational step for future advancements, promoting the widespread adoption of more sustainable and responsible AI practices for scientists and practitioners worldwide. It also encourages the exploration of hybrid techniques that combine Slovin’s subsampling with other optimization methods to further improve scalability and robustness in the complex landscape of explainable AI.

## REFERENCES

[1] A. M. Abdullahi. 2023. The challenges of advancing inclusive education: the case of somalia’s higher education.Journal of Law andSustainable 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 criticaldecision-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 interpretabilitySHAP values.Artificial Intelligence, 306, 103667.

[8] M. A. Batouei. 2024.A Feasibility Study On Artificial Neural Network-Based Prediction And Optimization Of Autoclave Curing Processutcomes 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 StatisticalSociety: Series B (Statistical Methodology), 59, 1, 3–54.

[13] T. Brown et al. 2020. Language models are few-shot learners. InAdvances 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 IntelligenceResearch, 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 modelinterpretation. InProceedings 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. InProceedings of the 6th Symposium onOperating 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. InProceedings ofthe 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 theRoyal 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-f or-AI-data-centers-Microsof t-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 residentialreal 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. InAdvances inNeural 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 artificialintelligence: a review, methods, and challenges. InPrimer 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 inindonesia.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 forqualitative primary, qualitative meta-analytic, and mixed methods research in psychology:

the apa publications and communicationsboard 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 emotionrecognition.Information Sciences, 619, 679–694.

[39] S. M. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. InAdvances in Neural Information ProcessingSystems. 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 StatisticalAssociation, 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 symbiosisin 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 ofwebstudent applications to improve service quality for budi luhur university students. InProceedings of the 4th International Conferenceon 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: asurvey 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: practicalapplications.Journal of Bioscience and Bioengineering, 115, 6, 579–589.

[48] J. R. Quinlan. 1993. Combining instance-based and model-based learning. InProceedings of the Tenth International Conference onMachine Learning, 236–243.

[49] A. Radford et al. 2021. Learning transferable visual models from natural language supervision. InProceedings of the 38th InternationalConference 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 highschool/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 ArtificialIntelligence 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-hocexplanations: 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, 386–404.

[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 ofmethods 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. InProceedings ofthe AAAI Conference on Artificial Intelligencenumber 9. Vol. 34, 13693–13696.

[61] M. Sundararajan and A. Najmi. 2020. The many shapley values for model explanation. InProceedings of the 37th International Conferenceon 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 IntelligenceResearch, 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. InAdvances 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.