Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sparse-Selective Similarity Regularisation for Enhanced Explanations

Ismail Alarab

ialarab@bournemouth.ac.uk

Bournemouth University

Mohammad Naiseh

Bournemouth University

Simant Prakoonwit

Bournemouth University

Research Article

Keywords: Explainable AI, Imbalanced Feature Importance, Regulariser, Machine Learning

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7316975/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sparse-Selective Similarity Regularisation for Enhanced Explanations

Ismail Alarab*[a], Mohammad Naiseh[a], Simant Prakoonwit[b]

[a]*Computing and Informatics, Bournemouth University, United Kingdom*
[b]*Department of Creative Technology, Bournemouth University, United Kingdom*

Abstract

We propose a regularisation method—Sparse-Selective Similarity (S_3)—designed to improve the local explainability of linear models by balancing sparsity and feature similarity in a context-aware manner. Unlike traditional regularisers such as L1 (which enforces sparsity) and L2 (which promotes smoothness), S_3 selectively applies a pairwise similarity penalty only when multiple features contribute significantly to a prediction. This avoids over-penalising models where one or few features dominate, while encouraging balanced importance distributions when several features are relevant. We evaluate the impact of S_3 on local explanation quality using three post-hoc explanation techniques—LIME, SHAP, and Feature Influence—on a real-world medical dataset. Explanation faithfulness is assessed via sufficiency, necessity, and prediction variability metrics across a range of top-K feature subsets. Our results show that models trained with S_3 achieve consistently higher necessity scores and more stable prediction behaviour, without sacrificing accuracy or sufficiency. This demonstrates that S_3 regularisation can significantly enhance the robustness and informativeness of local explanations in interpretable models.

Keywords: Explainable AI, Imbalanced Feature Importance, Regulariser, Machine Learning

Email addresses: ialarab@bournemouth.ac.uk (Ismail Alarab*), mnaiseh@bournemouth.ac.uk (Mohammad Naiseh), sprakoonwit@bournemouth.ac.uk (Simant Prakoonwit)

Preprint

1. Introduction

Explainable Artificial Intelligence (XAI) has become a foundation of trustworthy AI, enabling users to understand, justify, and trust the decisions made by machine learning models [1, 2, 3]. In fields such as healthcare, finance, and law, the ability to provide transparent and interpretable explanations is critical for building user confidence and supporting informed decision-making [4, 5, 6, 7, 8, 9].

Despite the current advances in XAI, several persistent challenges have been highlighted by the research community [10, 11, 12, 13]. These challenges can be broadly categorised into four main areas:

- **Contextual Complexity**: When the nature of the data or decision-making process makes explanations difficult to generate or understand (e.g., blockchain, financial data, rare diseases).
- **Audience Adaptability**: When explanations fail to account for the expertise, background, or cognitive load of different users, making them ineffective or inaccessible.
- **Fairness and Transparency**: When explanations inherit biases from data, models, or human annotators, leading to misleading or untrustworthy insights.
- **Usability and Engagement**: When explanations are not designed with user needs in mind, causing cognitive overload, lack of interactivity, or failure to support decision-making.

Focusing on the **Fairness** category, a well-known concern in AI is sampling bias, where some groups or feature patterns are underrepresented in the training data. This can lead to discriminatory outcomes, such as biased predictions based on gender or other societal factors [14, 15]. However, another important but less examined issue is what we refer to as the *imbalanced feature importance* problem. This occurs when certain features receive much larger weights than others, which can distort the way explanations are generated. In models such as logistic regression, a few features with large coefficients often dominate the explanation, while other relevant features are overlooked due to their smaller weights. This can result in misleading explanations.

Inspired by prior studies [16, 17, 18, 19], we note that such imbalance may arise from factors such as sampling bias, overreliance on dominant features, or suboptimal preprocessing. However, despite its implications for model interpretability, this issue remains underexplored in the literature. This is not merely a data quality concern, but also a modelling issue such that even well-curated datasets can lead to models with skewed coefficient distributions, compromising the quality of local explanations.

To address this, we introduce a novel regularisation technique named *Sparse-Selective Similarity (S_3)*. This regulariser improves local interpretability by jointly promoting sparsity and encouraging similarity among active coefficients in logistic regression models. Unlike traditional L_1 or L_2 penalties, the S_3 regulariser selectively reduces disparities between influential features, allowing smaller coefficients to meaningfully contribute to explanations.

Our contributions in this paper are as follows:

- We identify the *imbalanced feature importance problem* in local explainability, where dominant model coefficients hinder comprehensive feature attribution.
- We propose the S_3 regulariser as a principled solution to improve coefficient balance and enhance the interpretive value of logistic regression models.
- We empirically demonstrate the benefits of the S_3 regulariser using a real-world dataset, showing improved local faithfulness and model transparency.

Our work responds to recent calls for more principled XAI methods, such as those outlined by Freiesleben and König (2023) [20], who emphasise the role of model regularisation in achieving interpretable machine learning.

The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 presents our experiments and results; and Section 4 concludes with a discussion and directions for future research.

2. Related Work

XAI is known to be a hot topic with the increased complexity of AI models. Shahab et al. [9] reviewed how XAI can improve the interpretability of "black-box" models, showing its relevance for transparency. Petter

et al. [21] showed that XAI techniques helped improve explanation quality in credit scoring, making decisions clearer for users. Zhang et al. [22] discussed how explainable models can build trust in AI-based cyber-defence systems. Richmond et al. [23] looked at how XAI could help increase transparency and accountability in legal applications. Similarly, Deeks [24] argued that explainable models are useful for making legal decisions easier to understand. Despite these advancements, many studies have urged for more rigorous methodologies in applying XAI to go beyond simply justifying AI decisions and instead focusing on the interpretability and intuitiveness of explanations [25, 13, 7, 26]. This means that explanations become valuable with human-centric explanations.

2.1. The Evolving Landscape of XAI Challenges

Specifically, Debruijn et al. [10] argued that XAI should be viewed as a socio-technical challenge rather than merely a technical one. They emphasise that explanations of AI decisions must account for social and organisational contexts, as well as technical accuracy. Freiesleben et al. [20] criticised the XAI community for applying explanation techniques in a generic manner, without considering methodological rigour. They proposed steps to ensure that XAI research is grounded in solid methodologies and focused on producing actionable, understandable results. Swamy et al. [27] speculated that the future of human-centred AI will involve moving away from explaining black-box models. They suggested that the next wave of XAI research should prioritise models that are intrinsically interpretable, such as certain types of neural networks, which could mitigate the need for post-hoc explanation methods. Other recent work has shown that feature importance estimates from local explanation methods can vary depending on the data used [28]. Since explanations are influenced by both the data and the learning algorithm, it is normal to see that some features appear more important than others. We refer to this as imbalanced feature importance. Although common, this imbalance might affect how useful the explanations are, because feature importance plays a role in how we interpret model predictions, according to [16] and [18].

3. Explainability and Feature Imbalance

Before delving into imbalanced feature importance term, we introduce the different between global and local explainability in linear model.

3.1. Global and Local Explainability in Linear Model

To understand how feature imbalance impacts explainability, we clarify the distinction between feature importance and feature influence, supported by Framling (2023) [18]. Consider a binary logistic regression model expressed as:

$$\hat{y} = \beta_0 + \sum_{j=1}^{n} \beta_j x_j, \tag{1}$$

where $\mathcal{X}_{i,n} = \{x_1, \ldots, x_j, \ldots, x_n\}$ is the feature vector of the i^{th} instance in an n-dimensional dataset, and $\hat{y}$ is the model's predicted logit value. During training, the model learns the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ which define its decision boundary. The global feature importance of feature x_j is often interpreted as the magnitude of its coefficient:

$$\text{Global Importance}(x_j) = |\beta_j|. \tag{2}$$

This value reflects the overall contribution of feature x_j across the entire dataset and is commonly used in global explanations [18]. For a specific instance, the feature influence of x_j is determined by the product $\beta_j x_j$, which captures how that feature contributes to the individual (local) prediction:

$$\text{Feature Influence}(x_j) = \beta_j x_j. \tag{3}$$

This formulation supports local explanations, which aim to describe the model's behaviour for individual predictions rather than for the whole dataset.

3.2. Imbalanced Feature Importance

As discussed earlier, imbalanced feature importance refers to a situation where some features have much higher importance scores than others. This is common non-problematic in any AI model where the inputs do not contribute evenly. While some imbalance is expected, it becomes a concern when one or a few features dominate the model. For example, consider the linear model in Equation 1. Let β_1 be a large number in comparison to other coefficients, which can be expressed as follows:

$$|\beta_1| \gg |\beta_2|, |\beta_3|, \ldots, |\beta_n|.$$

In this case, it is more likely that the input feature x_1 dominates the explanations, due to the high feature imbalance.

3.3. Effect of Feature Imbalance on SHAP and LIME

Given the logistic regression model with disproportionate feature importance, both SHAP and LIME explanations can become influenced by the feature imbalance, leading to skewed interpretations of the model's decision. More formally, the prediction of a logistic regression model is given by:

$$f(\mathbf{x}) = \sigma\left(\beta_0 + \sum_{j=1}^{n} \beta_j x_j\right), \quad \text{where } \sigma(z) = \frac{1}{1+e^{-z}}.$$

SHAP values decompose the prediction as follows:

$$f(\mathbf{x}) \approx \sigma\left(\phi_0 + \sum_{j=1}^{n} \phi_j\right),$$

where ϕ_j is the SHAP value for feature x_j, and for linear models, these values are given by:

$$\phi_j = \beta_j \cdot (x_j - \mathbb{E}[x_j]).$$

When β_1 is significantly larger than the other coefficients and given non-zero deviation of xj from the feature average, the SHAP value for x_1 will get amplified, and consequently, increasing SHAP values. Similarly, LIME (Local Interpretable Model-agnostic Explanations) works by approximating the model locally with a surrogate model, typically a linear model, based on the instance to be explained. However, when there is significant feature imbalance in the logistic regression model, small perturbations in the most influential features highly affects the output. Since LIME relies on perturbing the input features and observing how the prediction changes, the dominant feature is expected to disproportionately affect the local surrogate model, skewing the explanation.

In both cases, when feature importance is imbalanced, explanations generated by SHAP and LIME are likely to be dominated by a few key features, leading to an incomplete and potentially misleading understanding of the model's behaviour. This reinforces the need to address feature imbalance for improved explainability, especially in case of underrepresented features. Although there may be various ways to address this issue, we focus on applying regularisation to a binary logistic regression classifier to mitigate the feature imbalanced importance effect.

3.4. Regularisation Techniques: L_1 Norm, L_2 Norm, ElasticNet

Regularisation techniques are essential in linear regression to prevent overfitting, especially in high-dimensional settings. Among the most widely used are Lasso ([29]) and Ridge regression ([30]). Lasso uses an L_1 penalty to promote sparsity, setting some coefficients exactly to zero. Ridge uses an L_2 penalty to shrink coefficients continuously toward zero without eliminating them. ElasticNet regularisation ([31]) combines the strengths of both to handle situations with highly correlated features, overcoming limitations of each method individually. To visualise the impact of regularization on logistic regression decision boundaries, we constructed a synthetic 2D dataset suitable for a logistic regression model. Two classes were generated from Gaussian distributions with means at [1.5, 1.5] and [-1.5, -1.5] in the 2D feature space, respectively. The standard deviations were chosen to maintain class separation either horizontally or vertically. Each class consisted of 500 samples, and the combined dataset was standardised to have zero mean and unit variance per feature. We then trained logistic regression classifiers us-

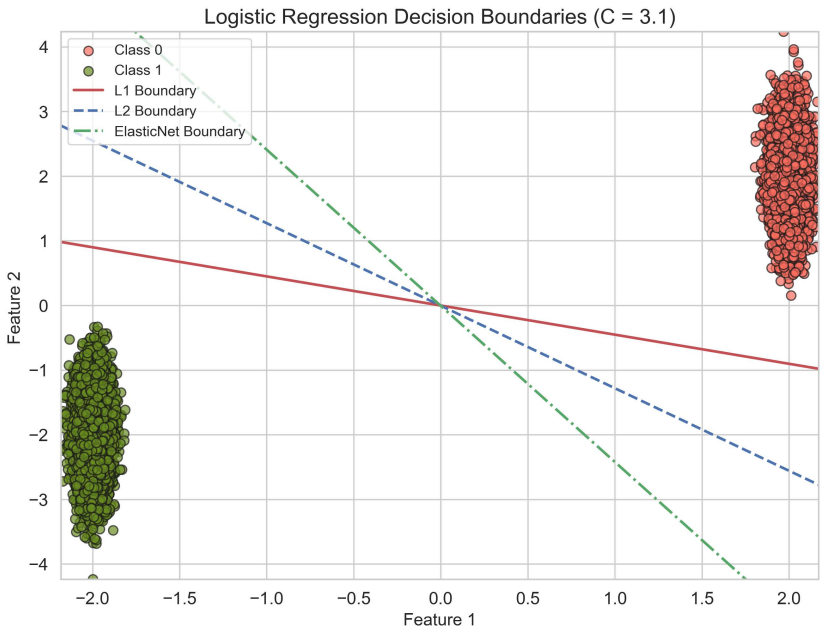


Figure 1: Logistic regression decision boundaries ($C = 2.5$) Comparison of L_1 (red), L_2 (blue), and ElasticNet (green) regularisation on a binary classification task. Class 0 and Class 1 are shown in red and green, respectively.

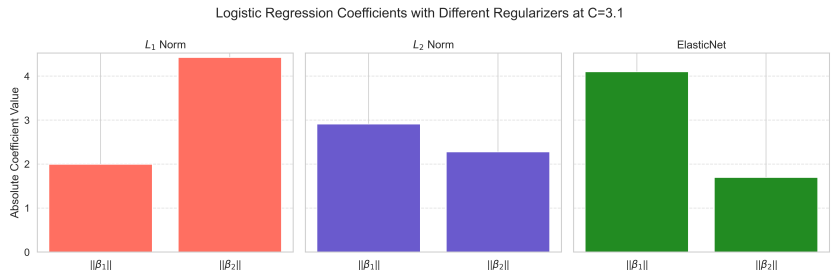


Figure 2: Logistic regression coefficients for different regularizers (L_1, L_2, ElasticNet) on the same dataset, with regularisation strength set to $C = 3.1$.

ing the three aformentioned regularisers: L_1, L_2, and ElasticNet, across a common inverse regularisation strength of $C = 2.5$. The resulting decision boundaries were visualised in Figure 1. All decision boundaries show the impact of each regulariser, which in turns affect the model coefficients (β_j) as depicted in Figure 2. As model coefficients are impacted by the choice of regularisation, their structure directly affects explainability. The patterns observed in Figure 2 motivate further investigation and suggest the need for a similarity-based regularisation approach that promotes pairwise coefficient similarity while preserving sparsity. This is explored in the following section.

4. Sparse-Selective Similarity (S_3) Regularisation

The proposed Sparse-Selective Similarity (S_3) regulariser extends the notion of coefficient uniformity by applying a selective pairwise similarity penalty only to features that contribute significantly to the model's output. Unlike the Fused Lasso [32], which enforces similarity between adjacent coefficients and is suited for ordered data, the S_3 regulariser generalises the similarity constraint globally across all important coefficients, while preserving sparsity by focusing only on impactful features.

The S_3 objective is defined as:

$$\min_{\beta,\beta_0} \mathcal{L}(\beta, \beta_0) + \lambda \cdot \mathcal{R}_{S_3}(\beta) \tag{4}$$

where $\mathcal{L}$ is the logistic loss, and $\mathcal{R}_{S_3}$ is the selective similarity regularisation term, defined only over a subset of active coefficients:

$$\mathcal{R}_{S_3}(\beta) = \sum_{i<j;\, i,j\in\mathcal{A}} |\beta_i - \beta_j| \tag{5}$$

Here, $\mathcal{A}$ denotes the set of active features, identified based on their relative contribution:

$$\mathcal{A} = \left\{ j \mid \frac{|\beta_j|}{\sum_k |\beta_k|} > \tau \right\} \tag{6}$$

with τ typically set to small ratio to maintain sparse features. The regulariser is conditionally applied only when $|\mathcal{A}| \geq 2$ to avoid degenerate or trivial penalties.

This formulation reflects the following assumptions:

- Feature interpretability benefits from a more uniform distribution of important coefficients.
- No single feature should dominate the model's prediction.
- Similar coefficients indicate functional similarity between features, improving explainability.

By applying the penalty selectively and conditionally, S_3 avoids enforcing unnecessary similarity when only one dominant feature exists—an important consideration for sparse models.

5. Experimental Setup and Results

We evaluate the proposed S_3 regulariser using the Breast Cancer dataset from `scikit-learn` [33], a widely used benchmark for binary classification. The task is to predict whether a tumor is malignant or benign based on 30 real-valued features. All features are standardised using z-score normalisation. The dataset is split into training (70%) and test (30%) sets.

As a baseline, we train a logistic regression model with L_1 regularisation to promote sparsity in feature coefficients. The learned coefficients are then used to initialise a second model trained with the proposed S_3 regulariser, referred to as the *regularised model*. This model is trained for up to 100 epochs with early stopping, using a regularisation weight $\lambda = 5$ and a feature importance threshold $\tau = 0.1$. The threshold ensures that only features contributing more than 10% are included in the S_3 penalty, as defined in Equation 6, thereby excluding sparse or low-importance features from the regularisation process.

Table 1 compares the classification performance of the baseline and regularised models. The regularised model preserves predictive performance

Table 1: Classification performance comparison of logistic regression model with and without S_3-regulariser on the Breast Cancer dataset.

Model	Accuracy	F_1-score (Weighted)
Base Model (without S_3)	98%	98%
Regularised Model (with S_3)	98%	98%

while modifying feature usage. Figure 3 visualises the learned coefficients before and after regularisation. The S_3 regulariser encourages similar magnitudes among dominant features, while leaving smaller coefficients largely unaffected. This behaviour demonstrates its effectiveness in mitigating feature imbalance without sacrificing accuracy. The next section investigates how this regularisation impacts post-hoc explainability.

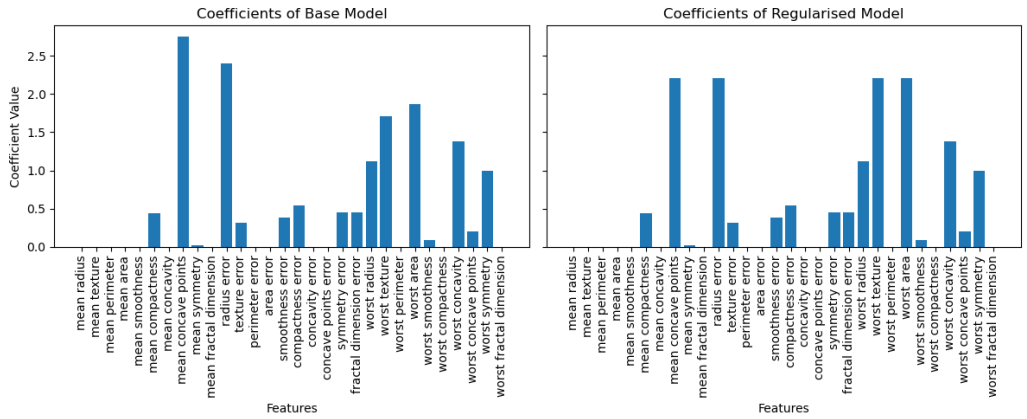


Figure 3: Coefficients of logistic regression models before and after S_3 regularisation.

6. Evaluating the Impact of S_3 Regulariser on Local Explanations

To evaluate the impact of S_3 regulariser on local explainability, we apply three post-hoc explanation techniques—LIME [34], SHAP [35], and the Feature Influence method—to generate local explanations for each test instance. Given a test point, we use these methods to compare the local explanations produced by the base model and the regularised model, thereby assessing the effect of regularisation on explainability. For each test sample, each explanation method produces a ranked list of features in descending order of importance. We compare the explanation quality by computing sufficiency,

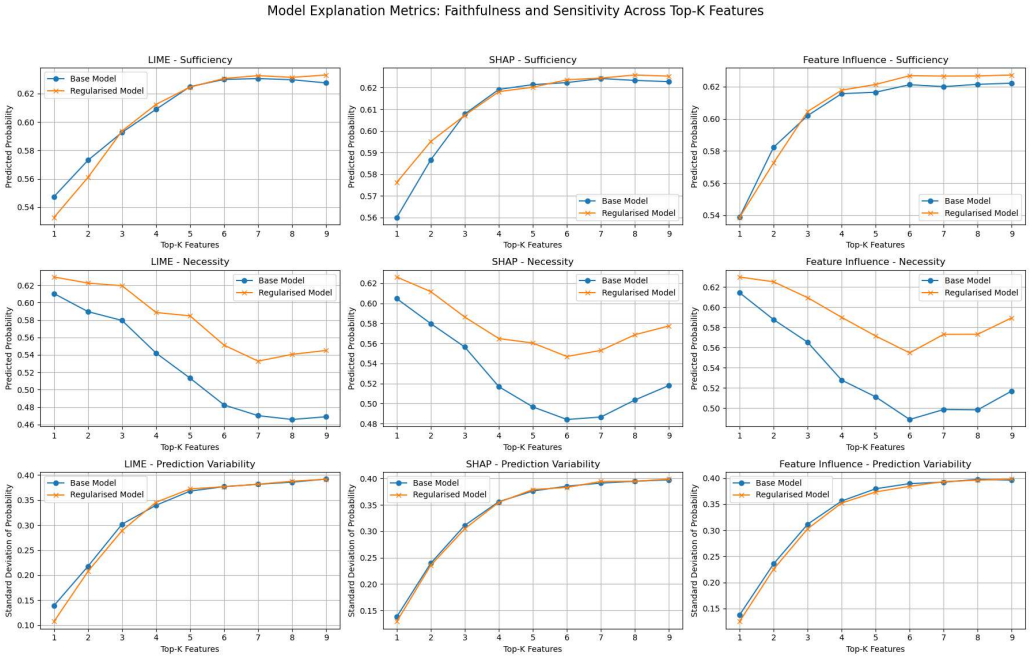


Figure 4: Comparison of explanation metrics across models and explanation methods. Rows correspond to sufficiency (top), necessity (middle), and prediction variability (bottom). Columns represent LIME, SHAP, and feature influence methods. Each plot shows results for the base model (blue) and the regularised model (orange) across different numbers of top-K features.

necessity and sensitivity metrics for both models, across multiple values of top-K features. This allows us to extract the top-K features for various values of K, enabling a detailed comparison of feature importance across models.

6.1. Sufficiency, Necessity, and Sensitivity Metrics

We used the following metrics to evaluate the reliability and informativeness of the local explanations:

- **Sufficiency**: Measures how well the top-K most important features alone can preserve the original model prediction.
- **Necessity**: Measures how much the model's confidence decreases when the top-K features are removed.

Metric (Method)	**P-value**(Base – Regularised)
Necessity (SHAP)	$\ll 0.01$
Necessity (Feature Influence)	$\ll 0.01$
Necessity (LIME)	$\ll 0.01$
Prediction Variability (FI)	0.016
Sufficiency (SHAP)	0.137
Sufficiency (Feature Influence)	0.178
Prediction Variability (SHAP)	0.240
Prediction Variability (LIME)	0.263
Sufficiency (LIME)	0.568

Table 2: Paired t-test results comparing the base and regularised models across three faithfulness metrics: sufficiency, necessity, and prediction variability (standard deviation). Significant improvements in necessity are observed with the regularised model ($p < 0.05$).

- **Prediction Variability**: Measures the sensitivity of the model's prediction to perturbations in the top-K features.

These metrics are computed for each test instance over a range of top-K values from $K = 1$ to $K = 10$ for each of base and regularised models. For every value of K, the metrics are averaged across all test points to summarised the overall explanations behaviour of each model. Figure 4 presents the aggregated results of the three metrics (rows) computed across the three explanation techniques (columns). Each subplot compares the metric values of the base model (blue curve) against the regularised model (orange curve) as K increases. Several key observations emerge:

- The regularised model consistently achieves **higher necessity scores**, indicating that the model relies on a richer set of informative features rather than overemphasising a few dominant ones.
- **Sufficiency values** are comparable between models, suggesting that the most important features in both models remain effective at reproducing predictions.
- **Prediction variability** remains stable or is slightly reduced under regularisation, indicating more robust and consistent explanations.

6.2. Statistical Significance

We conducted a paired t-test on metric values computed per test instance for each explanation technique. Table 2 summarises the results. Statistically

significant differences ($p < 0.05$) were observed in all *necessity metrics*, confirming that the S_3 regulariser encourages a more balanced use of features across instances. No significant change was observed in *sufficiency*, and only minor effects were seen on prediction variability.

6.3. Impact of S_3 on Explanation Quality

These findings demonstrate that the S_3 regulariser enhances the faithfulness and stability of local explanations by promoting more balanced feature reliance without sacrificing predictive performance or explanation sufficiency. In particular, the significant improvements in necessity metrics across all explanation methods suggest that S_3 encourages models to distribute importance more evenly across relevant features, thereby improving the robustness and trustworthiness of post-hoc interpretability techniques in logistic regression model on the given dataset. This positions S_3 as a practical tool for improving explanation quality in applications where interpretability is critical.

7. Conclusion

In this paper, we proposed the Sparse-Selective Similarity (S_3) regulariser to address the challenge of imbalanced feature importance in linear models. By combining sparsity with selective similarity among significant coefficients, S_3 aims to enhance the interpretability of local explanations without degrading predictive performance. Our experiments on the Breast Cancer dataset show that models trained with S_3 retain comparable accuracy to standard L_1-regularised models while improving the consistency and faithfulness of post-hoc explanations. In particular, S_3-regularised models achieved higher necessity scores and reduced prediction variability across multiple explanation methods. These results suggest that incorporating coefficient-level regularisation can offer a practical means to improve explanation quality in inherently interpretable models such as logistic regression. Future work may explore its applicability in neural network models by integrating S_3-style regularisation into the final dense layers or by encouraging similarity-aware sparsity in feature attribution layers, potentially improving explanation stability in complex architectures.

Funding

This research received no external funding.

References

[1] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy ai: From principles to practices, ACM Comput. Surv. 55 (9) (jan 2023). doi:10.1145/3555803.
URL https://doi.org/10.1145/3555803

[2] O. Williams, Towards human-centred explainable ai: A systematic literature review, Master's Thesis (2021).

[3] K. Ronanki, Towards an ai-centric requirements engineering framework for trustworthy ai, in: 2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2023, pp. 278–280. doi:10.1109/ICSE-Companion58688.2023.00075.

[4] A. Singh, S. Sengupta, V. Lakshminarayanan, Explainable deep learning models in medical image analysis, Journal of Imaging 6 (6) (2020) 52. doi:10.3390/jimaging6060052.
URL https://doi.org/10.3390/jimaging6060052

[5] T. Hulsen, Explainable artificial intelligence (xai): Concepts and challenges in healthcare, AI 4 (3) (2023) 652–666. doi:10.3390/ai4030034.
URL https://www.mdpi.com/2673-2688/4/3/34

[6] H. Durojaye, M. Naiseh, Explainable ai for intrusion detection systems: A model development and experts' evaluation, in: K. Arai (Ed.), Intelligent Systems and Applications, Springer Nature Switzerland, Cham, 2024, pp. 301–318.

[7] M. Naiseh, D. Al-Thani, N. Jiang, R. Ali, How the different explanation classes impact trust calibration: The case of clinical decision support systems, International Journal of Human-Computer Studies 169 (2023) 102941. doi:https://doi.org/10.1016/j.ijhcs.2022.102941.
URL https://www.sciencedirect.com/science/article/pii/S1071581922001616

[8] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, U. R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Computer Methods and Programs in Biomedicine 226 (2022) 107161.

doi:https://doi.org/10.1016/j.cmpb.2022.107161.
URL https://www.sciencedirect.com/science/article/pii/S0169260722005429

[9] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, H.-W. Liang, Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods, Informatics in Medicine Unlocked 40 (2023) 101286. doi:https://doi.org/10.1016/j.imu.2023.101286.
URL https://www.sciencedirect.com/science/article/pii/S2352914823001302

[10] H. de Bruijn, M. Warnier, M. Janssen, The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making, Government Information Quarterly 39 (2) (2022) 101666. doi:https://doi.org/10.1016/j.giq.2021.101666.
URL https://www.sciencedirect.com/science/article/pii/S0740624X21001027

[11] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin, B. Kang, Survey on explainable ai: From approaches, limitations and applications aspects, Human-Centric Intelligent Systems 3 (3) (2023) 161–188. doi:10.1007/s44230-023-00038-y.
URL https://doi.org/10.1007/s44230-023-00038-y

[12] M. Naiseh, Social explainable ai (social xai): Towards expanding the social benefits of xai, in: The Impact of Artificial Intelligence on Societies: Understanding Attitude Formation Towards AI, Springer, 2024, pp. 169–178.

[13] K. Alikhademi, B. Richardson, E. Drobina, J. E. Gilbert, Can explainable ai explain unfairness? a framework for evaluating explainable ai, arXiv preprint arXiv:2106.07483 (2021).

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (6) (2021) 1–35.

[15] U. Peters, M. Carman, Cultural bias in explainable ai research: A systematic analysis, Journal of Artificial Intelligence Research 79 (2024) 971–1000.

[16] I. Alarab, S. Prakoonwit, Effect of data resampling on feature importance in imbalanced blockchain data: comparison studies of resampling techniques, Data Science and Management 5 (2) (2022) 66–76. doi:https://doi.org/10.1016/j.dsm.2022.04.003.
URL https://www.sciencedirect.com/science/article/pii/S2666764922000145

[17] M. Saarela, S. Jauhiainen, Comparison of feature importance measures as explanations for classification models, SN Applied Sciences 3 (1) (2021) 272. doi:10.1007/s42452-021-04148-9.

[18] K. Främling, Feature importance versus feature influence and what it signifies for explainable ai, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 241–259.

[19] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, G. Valenza, From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks, Computer Methods and Programs in Biomedicine 236 (2023) 107550.

[20] T. Freiesleben, G. König, Dear xai community, we need to talk! fundamental misconceptions in current xai research, in: World Conference on Explainable Artificial Intelligence, Springer, 2023, pp. 48–65.

[21] P. E. de Lange, B. Melsom, C. B. Vennerød, S. Westgaard, Explainable ai for credit assessment in banks, Journal of Risk and Financial Management 15 (12) (2022). doi:10.3390/jrfm15120556.
URL https://www.mdpi.com/1911-8074/15/12/556

[22] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, F. Taher, Explainable artificial intelligence applications in cyber security: State-of-the-art in research, IEEE Access 10 (2022) 93104–93139.

[23] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, T. B. Moeslund, Explainable ai and law: an evidential survey, Digital Society 3 (1) (2024) 1.

[24] A. Deeks, The judicial demand for explainable artificial intelligence, Columbia Law Review 119 (7) (2019) 1829–1850.

[25] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (5) (2019) 206–215.

[26] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (11) (2021) e745–e750.

[27] V. Swamy, J. Frej, T. Käser, The future of human-centric explainable artificial intelligence (xai) is not post-hoc explanations, arXiv preprint arXiv:2307.00364 (2023).

[28] C. W. Ayad, T. Bonnier, B. Bosch, S. Parbhoo, J. Read, Feature importance depends on properties of the data: Towards choosing the correct explanations for your data and decision trees based models, arXiv preprint arXiv:2502.07153 (2025).

[29] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[30] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[31] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[32] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1) (2005) 91–108.

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

[34] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd

ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[35] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).