SHAP-Guided Regularization in Machine Learning Models[⋆]

Amal Saadallah

Lamarr Institute for Machine Learning and AI, Dortmund, Germany

Abstract

Feature attribution methods such as SHapley Additive exPlanations (SHAP) have become instrumental in understanding machine learning models, but their role in guiding model optimization remains underexplored. In this paper, we propose a SHAP-guided regularization framework that incorporates feature importance constraints into model training to enhance both predictive performance and interpretability. Our approach applies entropy-based penalties to encourage sparse, concentrated feature attributions while promoting stability across samples. The framework is applicable to both regression and classification tasks. Our first exploration started with investigating a tree-based model regularization using TreeSHAP. Through extensive experiments on benchmark regression and classification datasets, we demonstrate that our method improves generalization performance while ensuring robust and interpretable feature attributions. The proposed technique offers a novel, explainability-driven regularization approach, making machine learning models both more accurate and more reliable.

Keywords

SHapley Additive exPlanations (SHAP), Regularization, Tree-based Models, Explainability

1. Introduction

As machine learning models become increasingly complex, their interpretability and robustness are critical concerns across various domains, from finance and healthcare to autonomous systems [1]. While deep learning and gradient-boosted trees have shown remarkable predictive power, their black-box nature makes them difficult to trust in high-stakes applications. To address this, explainability techniques such as SHapley Additive exPlanations (SHAP) [2] have been widely adopted to quantify feature importance, offering insights into model decisions. However, while SHAP values help interpret trained models [3, 4], they are rarely incorporated directly into the training process to improve model behavior.

In this work, we introduce SHAP-guided regularization, a novel approach that integrates feature importance constraints into model optimization. Our method introduces two key regularization terms. The first term consists of SHAP entropy penalty – Encourages the model to rely on a sparse, well-distributed subset of important features. The second term is SHAP stability penalty– Ensures that feature attributions remain stable across different samples, reducing sensitivity to small perturbations in the data. By embedding these explainability-driven constraints into the learning objective, our method enhances both predictive accuracy and interpretability. The framework is applicable to both regression and classification tasks, and first experiments have shown that it is particularly effective for tree-based models such as LightGBM, XGBoost, and CatBoost.

We evaluate our approach on a diverse set of benchmark datasets, comparing its performance against standard models. Our results show that SHAP-guided regularization improves generalization by reducing overfitting to spurious correlations, enhances interpretability by concentrating feature

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[⋆] This research has partly been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine-Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

[*] Corresponding author.

amal.saadallah@cs.tu-dortmund.de (A. Saadallah)

0000-0003-2976-7574 (A. Saadallah)

importance on the most relevant predictors, and increases robustness by ensuring stable attributions across samples.

The rest of this paper is structured as follows: Section 2 discusses related works, including SHAP-based model interpretation and feature importance-driven regularization. Section 3 details our SHAP-guided regularization framework and training procedure. Section 4 presents empirical results, demonstrating the effectiveness of our approach across regression and classification tasks. Finally, Section 5 concludes with insights and future directions.

2. Related Works

2.1. Feature Importance and Explainability in Machine Learning

Interpretability in machine learning has gained significant attention, particularly in domains where model decisions impact critical outcomes, such as finance, healthcare, and autonomous systems. Traditional feature importance measures, such as permutation importance [5] and Gini importance in decision trees [6], provide insights into model behavior but often suffer from instability and bias toward correlated features.

SHapley Additive exPlanations (SHAP) [7] are a widely used approach that attributes feature importance based on cooperative game theory principles. Unlike other methods, SHAP ensures fair and consistent feature attribution, making it a popular tool for understanding model predictions. However, most applications of SHAP focus on post hoc analysis—explaining trained models—rather than integrating feature attributions into the learning process [2].

2.2. Regularization for Improved Generalization and Interpretability

Regularization techniques such as L1 (Lasso) [8] and L2 (Ridge) penalties [9] are commonly employed to improve model generalization by controlling feature weights. While these methods help prevent overfitting, they do not explicitly guide the model to focus on the most meaningful features. Other forms of feature selection, such as tree-based pruning [10] and attention mechanisms in deep learning [11], aim to refine model decision-making but often rely on heuristic approaches rather than interpretable attributions like SHAP values.

Some studies have explored feature importance-driven regularization. For instance, Alvarez-Melis and Jaakkola [12] propose stability-driven constraints to ensure consistent model explanations across similar samples. Meanwhile, Lundberg et al. [13] discuss the use of SHAP for feature selection but do not incorporate it into the training objective. To our knowledge, no prior work has introduced a SHAP-guided regularization framework that is applicable to both regression and classification tasks while explicitly optimizing for interpretability, stability, and predictive performance.

2.3. SHAP-Guided Learning: Bridging Interpretability and Optimization

A few recent works have begun exploring SHAP-integrated learning. In [14], a neural network architecture that incorporates Shapley values as latent representations. This design allows for intrinsic, layer-wise explanations during the model's forward pass, facilitating explanation regularization during training and enabling rapid computation of explanations at inference time. The authors in [15] propose X-SHIELD, a regularization technique that enhances model explainability by selectively masking input features based on explanations. Seamlessly integrated into the objective function, X-SHIELD improves both the performance and interpretability of AI models. SHAPNN [16] is a deep learning architecture tailored for tabular data, integrating Shapley values as a regularization mechanism during training. This approach not only provides valid explanations without additional computational overhead but also enhances model performance and robustness in handling streaming data.

Our proposed SHAP-guided regularization framework bridges this gap by incorporating SHAP-based entropy and stability penalties to encourage sparse and robust feature attributions, making the method applicable to both regression and classification in a unified manner and enhancing generalization while preserving explainability, a crucial factor in real-world decision-making. In the next section, we formalize our approach, detailing the mathematical formulation, training procedure, and advantages of SHAP-guided regularization.

3. Methodology

3.1. SHAP-Based Regularization for Learning Models

Our method integrates SHAP values into the model training process by introducing regularization terms based on entropy and stability of the feature attributions. The goal is to improve both the predictive performance and the interpretability of the model by guiding its focus towards the most relevant features while maintaining stable feature importance across similar inputs. This section describes how we incorporate SHAP-guided regularization into the model's loss function.

Given a set of training samples $\{(x_i, y_i)\}$, where x_i represents the feature vector and y_i the target, our objective is to learn a model $f(x_i)$ that minimizes a regularized loss function. For both regression and classification tasks, the total loss function L_{total} can be defined as:

$$L_{\text{total}} = L_{\text{task}} + \lambda_1 L_{\text{entropy}} + \lambda_2 L_{\text{stability}} \tag{1}$$

where L_{task} is the standard loss function for the task (e.g., mean squared error for regression or binary cross-entropy for classification). L_{entropy} is the entropy penalty that encourages sparse and interpretable feature importance distributions. $L_{\text{stability}}$ is the stability penalty that promotes consistency in SHAP attributions across similar samples. λ_1 and λ_2 are the regularization hyperparameters that control the influence of the interpretability penalties.

3.2. Regularization Terms Based on SHAP

3.2.1. SHAP Entropy Penalty (Sparsity)

The entropy penalty L_{entropy} is designed to sparsify the model's focus on important features. It is calculated as the Shannon entropy of the normalized SHAP values across all features for each prediction:

$$L_{\text{entropy}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{p}_{ij} \log(\hat{p}_{ij}) \tag{2}$$

where: N is the number of samples, M is the number of features, and $\hat{p}_{ij}$ represents the normalized absolute SHAP value for the j-th feature in the i-th sample. The entropy captures the uncertainty in the feature importance. The penalty encourages models to focus on a small subset of important features, reducing the influence of irrelevant ones. A higher penalty λ_1 leads to more sparse explanations.

3.2.2. SHAP Stability Penalty (Consistency)

The stability regularization term, $L_{\text{stability}}$, is designed to enforce consistency in the model's explanations by penalizing variations in SHAP values across similar input samples. Specifically, it quantifies how much the attribution of feature importance fluctuates between different but similar data points. Given a dataset of N samples and their associated SHAP values ϕ_{ik} for feature k of sample i, the stability loss is defined as:

$$L_{\text{stability}} = \frac{2}{N(N-1)M} \sum_{i=1}^{N} \sum_{\substack{i'=1 \\ i' \neq i}}^{N} \sum_{j=1}^{M} |\phi_{ij} - \phi_{i'j}|, \tag{3}$$

where ϕ_{ik} and $\phi_{i'k}$ denote the SHAP values for the j-th feature of samples x_i and $x_{i'}$ respectively, and M is the number of features. This formulation measures the average pairwise discrepancy in feature attributions across all sample pairs, normalized by the total number of comparisons. By minimizing $L_{\text{stability}}$, the model is encouraged to produce SHAP value distributions that are smooth and consistent across similar instances, thereby enhancing the robustness and reliability of the explanations. The regularization coefficient λ_2 controls the strength of this penalty; increasing λ_2 places greater emphasis on producing stable, coherent explanations during model optimization.

Our SHAP-guided regularization method offers several notable advantages. Firstly, by penalizing entropy and enforcing stability, the approach ensures that the model emphasizes the most critical features, leading to sparse and consistent feature attributions. This enhances interpretability, as the model's decisions become more transparent and understandable. Secondly, the incorporation of these regularization terms aids in reducing overfitting. By guiding the model to depend on a smaller, more stable subset of features, it promotes better generalization to unseen data. Thirdly, the framework's flexibility allows its application to both regression and classification tasks, providing a unified approach across different problem domains.

4. Experiments

4.1. Experimental Setup

We conduct experiments on 10 diverse datasets spanning regression and classification tasks. These datasets vary in size, feature dimensionality, and complexity, ensuring a comprehensive evaluation of our proposed SHAP-guided training approach. Table 1 summarizes the dataset characteristics.

Table 1
Summary of datasets used in experiments, including task type, number of samples, number of features, and target variable.

Dataset	Task	Samples	Features	Target Variable
Diabetes	Regression	442	10	Disease progression measure
California Housing	Regression	20,640	8	Median house value
Concrete	Regression	1,030	8	Concrete compressive strength
Airfoil	Regression	1,503	5	Sound pressure level
Energy	Regression	768	8	Heating load
Mushroom	Classification	8,124	22	Edibility (edible/poisonous)
Banknote Authentication	Classification	1,372	4	Authenticity (genuine/fake)
Credit Approval	Classification	690	15	Credit approval status
Breast Cancer	Classification	569	30	Diagnosis (malignant/benign)
Pima Indians Diabetes	Classification	768	8	Diabetes status

4.1.1. SHAP-Guided Model

LightGBM [17] was selected as the foundational model for implementing SHAP-guided regularization due to several compelling attributes. Its histogram-based algorithm significantly enhances computational efficiency by discretizing continuous feature values into discrete bins, thereby accelerating training processes and reducing memory usage. Additionally, LightGBM's inherent support for Tree-SHAP (SHapley Additive exPlanations) facilitates precise estimation of feature importance, making it particularly suitable for interpretability-focused modifications. The model's scalability is another advantage, as it adeptly manages large datasets with extensive feature sets. Furthermore, LightGBM consistently delivers robust performance across both classification and regression tasks. By integrating SHAP-guided regularization into LightGBM, the objective is to harmonize predictive accuracy with enhanced feature interpretability, ensuring that the model not only performs well but also provides transparent insights into its decision-making processes

Model Training Procedure Our first exploration of the combined loss function started by applying SHAP-guided regularization within the gradient-boosting framework, specifically using LightGBM for both classification and regression tasks. The training procedure proceeds as follows:

1. **Initialization:** Initialize the LightGBM model with default hyperparameters. Set the regularization hyperparameters λ_1 and λ_2 based on experimental settings.
2. **Iterative Training:** Train the model using LightGBM's iterative boosting mechanism. At each iteration t, we train a new decision tree and update the model's parameters.
3. **Loss Function Update:** After each boosting iteration, the total loss L_{total} is computed, which includes the task loss L_{task}, and the regularization terms L_{entropy} and $L_{\text{stability}}$. The model parameters are then updated to minimize this total loss function.
4. **Model Evaluation:** After training, the model is evaluated on a validation set using appropriate metrics (e.g., F1 score and AUC for classification, RMSE for regression).

Hyperparameter Tuning and Optimization To fine-tune the performance of the SHAP-guided method, we use cross-validation to select optimal values for λ_1 and λ_2. Typically, a grid search or random search is employed to find the combination of hyperparameters that minimizes the combined loss function.

4.1.2. Evaluation

To assess the effectiveness of SHAP-guided regularization, we compare our proposed SHAP-guided LightGBM model against several tree-based baseline machine learning models commonly used for structured data tasks (Decision Tree, Random Forest, LightGBM, XGBoost, and CatBoost).

Our SHAP-guided LightGBM extends the standard LightGBM model by incorporating SHAP-based regularization terms that encourage interpretability and stability in feature attributions.

To evaluate the performance of different models, we utilize the following metrics tailored for regression and classification tasks:

- **Regression:** RMSE (Root Mean Squared Error), R^2, SHAP Entropy, Top-k Concentration (Quantifies how concentrated SHAP attributions are among the top-k features), Stability.
- **Classification:** F1-score, AUC (Area Under the Curve), SHAP Entropy, Top-k Concentration, Stability.

4.2. Results

Tables 2 and 3 present the aggregated results, evaluating models in terms of standard predictive performance metrics—RMSE and R^2 for regression, F1-score and AUC for classification—alongside interpretability-driven metrics, including SHAP Entropy, Top-k Concentration, and Stability.

Table 2
Aggregated performance results across regression datasets. Lower RMSE and Entropy are better, while higher R^2, Top-k Concentration, and Stability indicate better performance and interpretability.

Model	RMSE ↓	R^2 ↑	Entropy ↓	Top-k Conc. ↑	Stability ↑
Decision Tree	17.02	0.64	1.32	0.86	0.58
Random Forest	11.87	0.82	1.55	0.78	0.60
LightGBM	11.78	0.83	1.17	0.86	**0.64**
XGBoost	12.29	0.82	1.16	0.87	0.63
SHAP-guided LightGBM	**11.45**	**0.83**	**1.12**	**0.89**	0.63

For regression tasks, the SHAP-guided LightGBM maintains competitive predictive performance while improving interpretability. The model achieves an RMSE of 11.45, which is comparable to standard

Table 3
Aggregated performance results across classification datasets. Lower Entropy is better, while higher F1, AUC, Top-k Concentration, and Stability indicate better performance and interpretability.

Model	F1 ↑	AUC ↑	Entropy ↓	Top-k Conc. ↑	Stability ↑
CatBoost	0.9194	0.9621	1.9782	0.7898	**0.8734**
Decision Tree	0.8850	0.9197	**1.2529**	0.8722	0.8496
LightGBM	0.9141	0.9592	1.8261	0.8551	0.8647
Random Forest	0.9163	0.9604	2.1783	0.7382	0.8699
XGBoost	0.9171	0.9615	1.8904	0.7993	0.8716
SHAP-guided LightGBM	**0.9207**	**0.9641**	1.6542	**0.8905**	0.8604

LightGBM (11.78) and outperforms other baselines. Similarly, the R^2 score remains at 0.83, confirming that the model retains its ability to explain variance in the data. In terms of interpretability, SHAP Entropy is reduced to 1.12, indicating that feature importance is more concentrated and less dispersed compared to standard LightGBM (1.17) and Random Forest (1.55). This suggests that SHAP-guided regularization encourages a more structured attribution pattern, enhancing transparency in feature importance. Furthermore, Top-k Concentration improves to 0.89, surpassing the standard LightGBM (0.86) and XGBoost (0.87), meaning that the model places greater emphasis on the most relevant features. Stability remains at 0.63, aligning closely with baseline models, demonstrating that the regularization does not introduce fluctuations in feature attributions.

For classification tasks, similar trends are observed. The SHAP-guided LightGBM achieves an F1-score of 0.9207 and an AUC of 0.9641, both slightly surpassing the standard LightGBM (0.9141 F1-score, 0.9592 AUC). This indicates that the introduction of SHAP-based regularization does not degrade predictive performance. More importantly, SHAP Entropy is reduced to 1.6542, compared to 1.8261 for LightGBM and 2.1783 for Random Forest, highlighting a more refined and focused attribution distribution. Top-k Concentration is the highest among all models (0.8905), confirming that the model consistently assigns importance to a small subset of critical features, which enhances interpretability. Stability remains competitive at 0.8604, slightly lower than LightGBM (0.8647) but higher than other baselines, ensuring robustness in feature attributions.

Figure 1 shows an illustration of SHAP diagram using standard lightGBM (Baseline Model) and the SHAP-guided LightGBM for the airfoil regression dataset. It is clear that the SHAP regularization promotes stability by compromising similar feature importance to similar samples (more condensed regions in Figure 1b). This is confirmed further by Figure 2, which shows lower variance of SHAP values across different features using the guided-SHAP model for the same dataset.

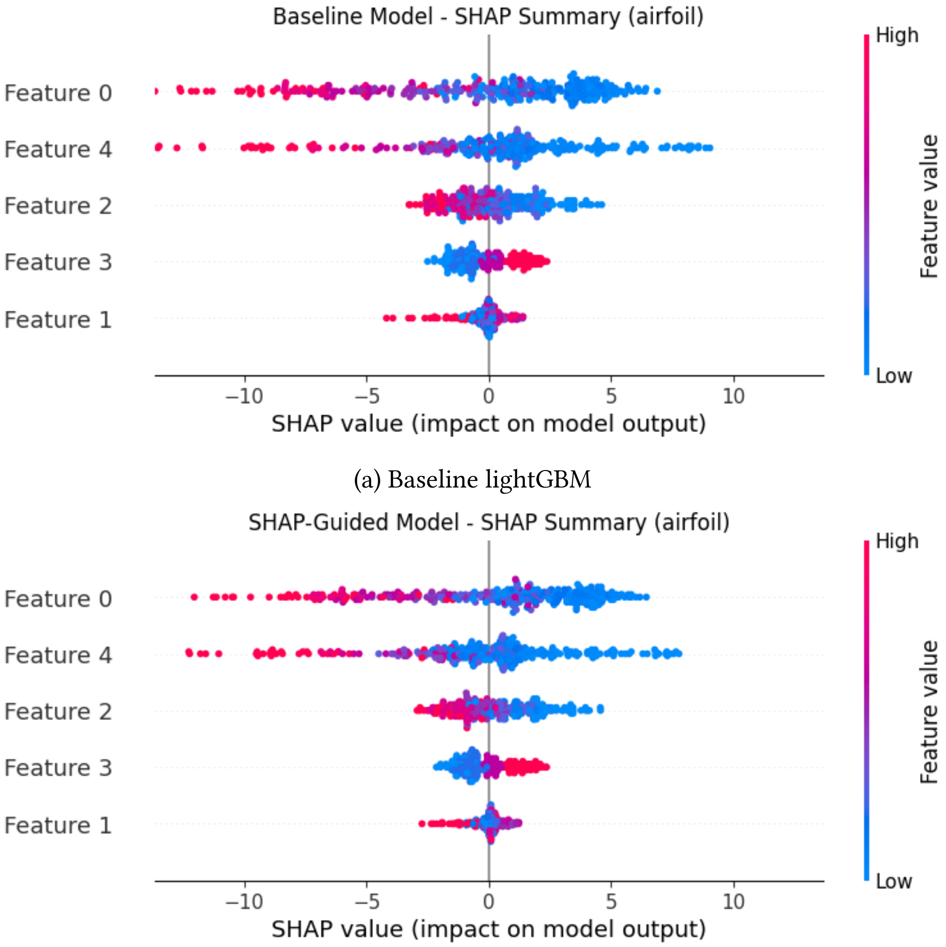
Overall, these results demonstrate that SHAP-guided regularization effectively enhances interpretability without compromising predictive accuracy. The method successfully reduces SHAP Entropy, leading to sparser and more meaningful feature attributions, while increasing Top-k Concentration, ensuring the model prioritizes the most relevant features. Stability remains comparable to non-regularized models, confirming that the proposed method does not introduce instability in feature attributions. These findings indicate that SHAP-guided learning can serve as a powerful tool for balancing interpretability and predictive performance in tree-based models.

5. Conclusion

We introduced a first exploration of SHAP-guided loss training. First experiments on LightGBM showed that SHAP-based regularization promotes interpretable and stable feature attributions while maintaining strong predictive performance.

SHAP regularization requires further detailed exploration as it seems to be a potential tool for:

- Improving SHAP-based interpretability metrics without degrading accuracy.
- Enhancing feature attribution stability across datasets.



(a) Baseline lightGBM

(b) SHAP-guided lightGBM

Figure 1: SHAP Diagram on the airfoil regression dataset.

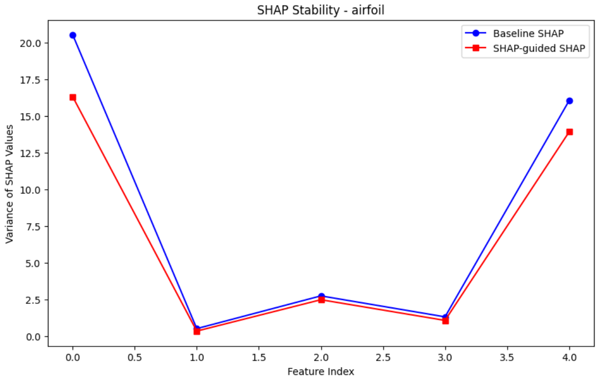


Figure 2: SHAP values mean variance comparison on the airfoil regression dataset.

- Providing a novel approach to balancing predictive performance with interpretability in ML models.

These insights demonstrate that SHAP-guided learning is a promising direction for explainable machine learning.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020).

[2] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[3] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, Frontiers in big Data 4 (2021) 688969.

[4] Z. Li, Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost, Computers, Environment and Urban Systems 96 (2022) 101845.

[5] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.

[6] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC bioinformatics 10 (2009) 1–16.

[7] L. Antwarg, R. M. Miller, B. Shapira, L. Rokach, Explaining anomalies detected by autoencoders using shapley additive explanations, Expert systems with applications 186 (2021) 115736.

[8] M. Schmidt, G. Fung, R. Rosales, Optimization methods for l1-regularization, University of British Columbia, Technical Report TR-2009-19 (2009).

[9] C. Cortes, M. Mohri, A. Rostamizadeh, L2 regularization for learning kernels, arXiv preprint arXiv:1205.2653 (2012).

[10] Y. Chen, D. Miao, R. Wang, K. Wu, A rough set approach to feature selection based on power set tree, Knowledge-Based Systems 24 (2011) 275–281.

[11] G. Brauwers, F. Frasincar, A general survey on attention mechanisms in deep learning, IEEE Transactions on Knowledge and Data Engineering 35 (2021) 3279–3298.

[12] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, Advances in neural information processing systems 31 (2018).

[13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[14] R. Wang, X. Wang, D. I. Inouye, Shapley explanation networks, arXiv preprint arXiv:2104.02297 (2021).

[15] I. Sevillano-García, J. Luengo, F. Herrera, X-shield: Regularization for explainable artificial intelligence, arXiv preprint arXiv:2404.02611 (2024).

[16] Q. Cheng, S. Qu, J. Lee, Shapnn: Shapley value regularized tabular neural network, arXiv preprint arXiv:2309.08799 (2023).

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).