Evaluating the Potential of SHAP-Based Feature Selection for Improving Classification Performance

Ashis Kumar Mandal
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur, Bangladesh
e-mail:ashis@hstu.ac.bd

Basabi Chakraborty
Madanapalle Institute of Technology and Science
Andhra Pradesh, India
Iwate Prefectural University
Iwate, Japan
email:basabi@iwate-pu.ac.jp

***Abstract*—Feature selection is an important preprocessing step in developing efficient and accurate classification models. Among various techniques, recently SHapley Additive exPlanations (SHAP)-based feature selection has gained attention for its interpretability and ability to quantify the contributions of individual features to model predictions. This study investigates the effectiveness of SHAP-based feature selection technique, specifically focusing on Linear SHAP, in improving classification performance. The research utilizes 10 diverse datasets to evaluate Linear SHAP's capability in identifying relevant features for classification tasks. The performance of Linear SHAP is assessed across varying percentages of selected features and compared to classification models without feature selection. Three popular filter-based feature selection approaches: Chi-square(Chi^2), Mutual Information, and Correlation-based methods are also used for feature selection with the same bench mark data sets. Comparative analysis, supported by statistical significance tests, demonstrates that Linear SHAP performs equally well to the traditional methods while offering the added benefit of interpretability. The findings suggest that Linear SHAP is a viable and promising alternative to established feature selection techniques in the realm of classification tasks.**

Keywords-Feature Selection; SHapley Additive exPlanations (SHAP); classification models; machine learning.

I. INTRODUCTION

Feature selection plays a pivotal role in machine learning models, particularly in classification tasks, by identifying and retaining the most relevant features from high-dimensional datasets in order to improve performance of the model. This process not only enhances model accuracy but also improves computational efficiency, reduces overfitting, and increases interpretability [1]. While traditional feature selection methods, including filter [2], wrapper [3], and embedded approaches [4], have been widely studied and applied to real world problems, they often struggle to find out efficient and optimal feature subset from high-dimensional datasets and sometime may not able to capture complex feature to feature interactions.

In recent years, eXplainable Artificial Intelligence (XAI) has gained significant attention in the development of trustworthy AI systems for recommendation and decision-making, particularly in high-risk application areas such as healthcare, finance, and control. As feature selection is an important preprocessing step, interpretable feature selection leads to the improvement of explanation ability of any pattern recognition or machine learning model. SHapley Additive exPlanations (SHAP) [5] has recently gained attention as an interpretable framework for understanding the contributions of individual features in machine learning models. While SHAP is widely recognized as a tool for model explanation, its utility as a feature selection mechanism is less explored. Linear SHAP [6], a variant designed for linear models, offers computational efficiency and interpretability, making it a promising candidate for feature selection. Its ability to quantify feature importance in an additive and consistent manner provides a unique advantage for understanding the relationship between features and predictions. However, a comprehensive evaluation of SHAP-based feature selection before classification tasks and its comparison with established methods is still lacking in the literature.

This paper seeks to fill the existing gap by thoroughly investigating the potential of SHAP-based feature selection to enhance classification performance. We outline a systematic approach to evaluate SHAP-based feature selection in comparison to traditional methods, focusing on the following key questions:

1) How does SHAP-based feature selection perform relative to approaches that do not utilize feature selection in terms of classification performance?
2) What is the impact of reducing the number of features on classification performance?
3) How does the performance of SHAP-based feature selection compare with popular filter-based feature selection methods?

To address the above questions, we have utilized a diverse array of benchmark datasets. Our methodology have involved a rigorous comparison of SHAP-based feature selection against well-established techniques, such as Chi-square, Mutual Information, and Correlation-based feature selection approaches.

The rest of this paper is organized as follows: Section II provides a theoretical background on feature selection in general and SHAP based approaches for feature selection. Section III describes our proposed methodology for evaluating SHAP-based feature selection in comparison to other existing state of the art approaches. Section IV presents the experimental results followed by a short section on discussion on the limitations of this study. Finally, Section VI concludes the paper and outlines directions for future research.

II. Theoretical background and related study

A. Feature selection

Feature selection aims to identify the most informative and discriminative features while eliminating irrelevant or redundant ones. Based on their interaction with the learning model, feature selection methods are categorized as filter, wrapper, and embedded approaches. Filter methods evaluate the relevance of features using statistical measures or intrinsic properties of the data. The most common and widely used techniques include Chi-square, Correlation-based analysis, Mutual Information and ANOVA [7] [8]. Wrapper methods involve the classifier model for evaluation of the feature subsets by training and validating the model on the data, with examples such as, Recursive Feature Elimination (RFE) and forward or backward selection [9]. Embedded methods integrate feature selection directly into the model training process, as seen with L1 regularization (Lasso) [10]. Feature selection can also be classified based on its approaches, such as feature ranking, where the top-K features are selected, or feature subset selection, which aims to identify an optimal or near-optimal subset of features [11]. These methods collectively improve model performance, reduce dimensionality, and enhance interpretability of the classification task as a whole.

B. SHAP

The idea of Shapley values originated from cooperative game theory. In game theory, the Shapley value of a player is the average marginal contribution of the player in a cooperative game. The Shapley value framework was developed by Llyod Shapley [12] which is based on some fairness axioms. This framework fairly allocates a contribution score to each player, reflecting their role in achieving the overall payoff. Lundberg applied the idea to machine learning, in which SHAP(SHapley Additive exPlanations) treats each feature as a player and calculates its contribution to the model's predictions [13]. SHAP approximates Shapley value by computing the contribution to a model's prediction of every subsets of features, given a dataset with m features. In this context, this approach offers a reliable and interpretable means of assessing the influence of individual feature or a subset of features on model outputs, facilitating both local and global insights into the model's behavior [14]. Computation of exact solution of Shapley values is quite infeasible for large number of inputs (players or features) due to the exponential nature of the problem. SHAP approximate the solutions through special weighted linear regression for any model or throughout different assumptions about feature dependence for ensemble tree models [15].

C. SHAP-Based Feature Selection Approaches

To date, several research efforts have utilized SHAP to improve model interpretability and examined its application in feature selection [16]. SHAP has been used effectively in medical diagnostics to improve the interpretability of the model. Huang et al. [17] developed a logistic regression model for the detection of heart failure, integrating SHAP for the global interpretation of the significance of features. This approach demonstrated superior precision compared to traditional methods by focusing on clinically relevant features. Luo et al. utilized SHAP to predict water quality indices, illustrating its capacity to highlight the most influential features in hydrological datasets [18]. Gehlot et al. [19] employed SHAP to enhance the explainability of machine learning models for surface electromyography-based hand gesture recognition. This study integrated SHAP scores to refine feature subsets, increasing the precision and interpretability of the model. SHAP has proven valuable in cancer detection, as demonstrated by a study that combined SHAP with machine learning for metabolomic analysis in breast cancer patients. This hybrid approach outperformed traditional selection methods, providing detailed insights into feature contributions [20]. In the domain of NLP, Ramanujam et al. [21] applied SHAP to select features for classifying spam SMS in Dravidian languages. Santos et al. [22] explored SHAP for efficient feature selection in the domain of industrial fault diagnosis. Here, an explainable artificial intelligence (XAI) technique is incorporated to meticulously select optimal features for the machine learning (ML) models. The chosen ML technique for the tasks of fault detection, classification, and severity estimation is the support vector machine (SVM). The interpretable analysis method based on Shapley values effectively enhances the performance of recognizing similar gestures and provides valuable insights into the decision-making process of recognition models in the research work of Wang et.al [23]. Overall, the motivation behind the use of shapley value or SHAP in selection of optimal features for a classification task is to build interpretable model capable of explaining the behavior of the model in the decision process.

III. Methodology

The primary objective of this research is to perform experiments using features identified by a SHAP-based approach to build an efficient classifier model. To achieve this, we collect 10 datasets for experimentation. The overall workflow of the task is illustrated in Figure 1. Initially, the datasets are preprocessed, which involves data cleaning, imputation of missing values, date encoding, normalization, and data balancing. After necessary processing, the datasets are partitioned into training and testing sets. The training datasets are then used for feature selection via the SHAP-based approach. A machine learning model is built using the selected features from the training datasets. The testing datasets are used to evaluate the performance of the models. Experiments are conducted with varying percentages of selected features for each dataset to analyze the effect of feature subset size on classifier performance. For comparison purposes, CHI2, Mutual Information, and Correlation-based feature selection methods are also employed for feature ranking.

A. Datasets

For our experimental analysis, we selected ten diverse datasets from the UCI Machine Learning Repository [24]. These datasets vary in their number of features, instances,

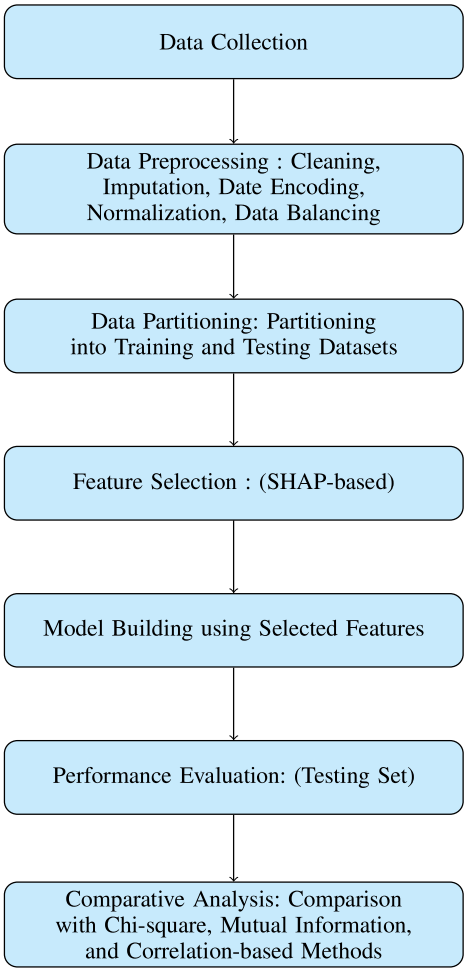

Figure 1. Workflow of the Research Methodology.

and classes, providing a comprehensive test bed for our experiments. Table I presents an overview of these datasets.

TABLE I. DATASETS

Datasets	No. of feature	No. of instances	No. of classes
Breast-w	9	683	2
Clean	166	476	2
Hepatitis	19	155	2
Parkinsons	22	195	2
Promoters	57	107	2
Qsar-biodeg	42	1055	2
Sonar	60	208	2
Spect	23	267	2
Spectf	45	349	2
Wisconsin	17	110	7

B. Data Preprocessing

We cleaned the datasets to remove inconsistent entries and address missing values. For continuous features, we imputed missing values using the mean, while for discrete features, we used the median. To process categorical data, we applied one-hot encoding, which transforms categorical variables into binary vectors by creating separate binary columns for each category. A value of 1 indicates the presence of a specific category, while 0 indicates its absence. Additionally, we used label encoding to convert each category into a unique integer, enabling machine learning models to process categorical data as numeric inputs.

To address class imbalances and reduce overfitting, we performed data balancing using the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for underrepresented classes.

Finally, we normalized the features to ensure they were represented on a uniform scale. We applied mini-max normalization to scale all feature values to a range between 0 and 1, preventing features with larger magnitudes from dominating the model training process.

C. Data Partitions and feature selection

After preprocessing, each dataset was partitioned into training and testing sets, with 80% of the data allocated for training and 20% for testing. The training dataset was primarily utilized for feature selection.

For feature selection, we employed a SHAP-based approach, specifically using the SHAP Linear Explainer to compute feature importance. Logistic Regression was selected as the underlying model for this process. The Linear Explainer was chosen due to its ability to calculate feature contributions with minimal computational overhead compared to kernel-based or tree-based SHAP methods. This approach provided a quantitative measure of feature importance, enabling us to rank the features and select the top n% for further analysis.

To compare SHAP-based feature selection with other methods, we employed three popular rank-based feature selection techniques: Chi-square (Chi^2), mutual information, and Correlation-based methods. Each of these methods was applied to the same dataset, and the top n% features were selected for each case.

For each feature selection approach, we evaluated the performance of machine learning models trained on the selected features. Logistic Regression was chosen as the classification model for this evaluation. Using the training dataset, we built models based on the features selected by each feature selection method. The classification accuracy of these models was then assessed using the testing dataset to evaluate the effectiveness of the respective feature selection approaches. To assess the impact of feature selection, we compared the models' performance with 25%, 50%, 75%, and 100% of the features, where 100% represents the dataset without any feature selection.

D. Experimental setup

For each feature selection approach, the experiment was conducted five times using different seed values to ensure robust results. The average classification accuracy across these runs was calculated to evaluate the performance of the models. The implementation was developed in Python. For SHAP-based feature selection, the SHAP library was utilized, while the scikit-learn library was used for building and evaluating machine learning models. The entire experiment was performed on a system with an Intel Core i5 processor, 8 GB of RAM, and running the Windows operating system.

IV. Experimental Results and Analysis

In this section, we present the results obtained using SHAP-based feature selection approaches and compare them with three other ranking-based filter type feature selection approaches.

Table II illustrates the average classification accuracy over 10 independent runs. It presents the results of the linear SHAP-based feature selection approach across different selection percentages. The analysis involves 10 datasets, considering feature selection percentages of 25%, 50%, 75%, and 100% (i.e., without feature selection).

When comparing feature selection with no feature selection, it is observed that reducing the features to 50% or 75% yields an average classification accuracy similar to that achieved without applying feature selection. Specifically, for 75% selected features using the SHAP-based approach, eight datasets exhibit the same average classification accuracy compared to the results obtained without feature selection. In one case, the SHAP-based approach outperforms, while in another case, it performs slightly worse compared to the scenario without feature selection.

Similarly, when 50% of the features are selected using the SHAP-based approach, four datasets demonstrate improved average classification accuracy compared to the results without feature selection. One case shows significantly better performance, while the remaining datasets show slightly lower performance compared to the case of without feature selection.

Table III. compares the performance of Linear SHAP with three filter-based feature selection methods: Chi2, Mutual Information, and Correlation. In this comparison, the average classification accuracy is considered with 50% of the features selected. It is found that Linear SHAP consistently demonstrates superior performance across various datasets, emphasizing its effectiveness in feature selection tasks. Specifically, it outperforms Chi2 in datasets such as Qsar-biodeg, Wisconsin, and Clean, while achieving equal performance in Breast-w, Parkinsons, Sonar and Spectf. Compared to Mutual Information, Linear SHAP shows notable advantages in Wisconsin and Qsar-biodeg, although Mutual Information performs better in Parkinsons and Promoters, with both methods yielding similar results in Breast-w. When evaluated against the Correlation method, Linear SHAP exhibits strengths in Wisconsin, Clean and Qsar-biodeg while Correlation is more effective in Parkinsons, with identical outcomes observed for Breast-w. Overall,

TABLE II. Performance Analysis of Linear SHAP-Based Feature Selection Based on Average Classification Accuracy Across Different Selection Percentages

Datasets	Percentage of selected feature			
	25%	50%	75%	100% (Without Feature Selection)
Breast-w	0.95	0.97	0.97	0.97
Clean	0.77	0.81	0.81	0.81
Hepatitis	0.71	0.76	0.78	0.78
Parkinson	0.74	0.76	0.76	0.76
Promoters	0.75	0.76	0.76	0.76
Qsar-biodeg	0.80	0.82	0.83	0.83
Sonar	0.76	0.75	0.78	0.78
Spect	0.71	0.69	0.69	0.68
Spectf	0.79	0.78	0.79	0.79
Wisconsin	0.95	0.96	0.96	0.97

Linear SHAP demonstrates better performance compared to the other methods in three out of 10 data sets, while maintains equal performance in six out of 10 data sets. Only in case of Parkinsons dataset, Mutual Information and Correlation based methods produce better results than Linear SHAP. From these findings, it can be inferred that Linear SHAP demonstrates robust and consistent performance, establishing itself as a reliable approach for feature selection across diverse datasets.

TABLE III. Accuracy Comparison of Linear SHAP Across Ranking-Based Feature Selection Approaches

Datasets	Linear Shap	Chi^2	Mutual Information	Correlation
Breast-w	0.97	0.97	0.97	0.97
Clean	0.81	0.78	0.77	0.78
Hepatitis	0.76	0.77	0.77	0.77
Parkinsons	0.76	0.76	0.78	0.80
Promoters	0.76	0.78	0.79	0.78
Qsar-biodeg	0.82	0.77	0.80	0.80
Sonar	0.75	0.75	0.76	0.74
Spect	0.69	0.70	0.70	0.70
Spectf	0.78	0.78	0.77	0.78
Wisconsin	0.96	0.93	0.93	0.93

TABLE IV. Pairwise Wilcoxon Signed-Rank Test p-values for Linear SHAP Compared to Other Methods

Method	**Chi2**	**Mutual Info**	**Correlation**
Linear SHAP	0.40	0.95	0.85

Table IV shows a statistical comparison of Linear SHAP with three different approaches using a pairwise Wilcoxon

Signed Rank test. All three comparisons show relatively high p-values (< 0.05), suggesting that Linear SHAP's performance is not statistically different from any of the other three approaches. This implies that Linear SHAP performs comparably similar to the alternative methods in this analysis.

V. Discussion and Limitation

The experimental results and their analysis has been presented in the previous section. In this study, we have used linear models of classification and Linear SHAP version of SHAP implementation is chosen for feature selection as it is computationally less expensive. The primary objective is to explore the viability of SHAP-based methods for rank based feature selection in comparison to traditional feature selection approaches: Chi2, Mutual Information, and Correlation-based methods. Linear SHAP is utilized to select features, with its performance evaluated using linear machine learning models across different percentages of selected features. The subset based feature selection methods are not examined in this work. Though we have not presented detail results of computational cost involved in feature selection methods, it seems that the computational cost of Linear SHAP is comparable with other popular optimal feature selection algorithms for datasets having moderate number of features.

Experimental results demonstrate that Linear SHAP is an effective tool for feature selection, consistently identifying relevant features and maintaining competitive performance across diverse datasets. The comparative analysis highlights its strengths and reliability when compared with Chi2, Mutual Information, and Correlation-based approaches. Statistical significance tests indicate that Linear SHAP performs equivalently to these traditional methods in several cases while offering the added advantage of interpretability, making it particularly valuable for applications where interpretability is a priority.

VI. Conclusion

In this work, a preliminary study has been done to assess the potential of SHAP- based feature selection method in comparison with other popular filter based methods with a limited number of bench mark data sets of moderate dimension. The results are encouraging. It shows that the performance of SHAP based method is comparable to other popular rank based feature selection algorithms. At present, the explanation capability of a decision model is very much valued in many practical applications in the area of medical or finance. In this context, SHAP based methods have a solid background and mathematical foundation to facilitate interpretability of the selected features. The development of effective SHAP based optimal feature selection algorithm can have a great impact in designing explainable decision systems.

As this study is limited to rank based feature selection algorithms involving linear SHAP, detail experiments with more sophisticated versions of SHAP implementations with more high dimensional data sets are needed for proper evaluation of SHAP based methods, especially regarding computational cost. Future study could also focus on developing hybrid methods that combine Linear SHAP with other feature selection techniques to leverage complementary strengths. Additionally, evaluating the scalability and generalization capability of these methods on more complex datasets and classification tasks would provide deeper insights into their broader applicability for interpretable feature selection.

Acknowledgment

This research project was supported by Japan Society of Promotion of Science (JSPS) KAKENHI Grant Number JP 24K15089 Type: Kaken C.

References

[1] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, 2022, ISSN: 1573-7497. DOI: 10.1007/s10489-021-02550-9.

[2] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Briefings in Bioinformatics*, vol. 23, no. 1, pp. 1–13, 2022, ISSN: 1477-4054. DOI: 10.1093/bib/bbab354. eprint: https://academic.oup.com/bib/article-pdf/23/1/bbab354/42229629/bbab354.pdf.

[3] J. Maldonado, M. C. Riff, and B. Neveu, "A review of recent approaches on wrapper feature selection for intrusion detection," *Expert Systems with Applications*, vol. 198, p. 116 822, 2022, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2022.116822.

[4] N. Mahendran and D. R. V. P M, "A deep learning framework with an embedded-based feature selection approach for the early detection of the alzheimer's disease," *Computers in Biology and Medicine*, vol. 141, p. 105 056, 2022, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.105056.

[5] S. M. Lundberg and S.-I. Lee, *Consistent feature attribution for tree ensembles*, 2018. arXiv: 1706.06060 [cs.AI].

[6] L. Schulte, B. Ledel, and S. Herbold, "Studying the explanations for the automated prediction of bug and non-bug issues using lime and shap," *Empirical Software Engineering*, vol. 29, no. 4, p. 93, 2024, ISSN: 1573-7616. DOI: 10.1007/s10664-024-10469-1.

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[8] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[9] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[12] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[14] C. Molnar, *Interpretable machine learning: A guide for making black box models explainable*. Lulu. com, 2020.

[15] S. Lundberg, G. Erion, and H. e. a. Chen, "From local explanations to global understanding with explainable ai for trees," *Nat Mach Intell*, pp. 56–67, 2020. DOI: https://doi.org/10.1038/s42256-019-0138-9.

[16] E. Wilson and D. M. Eler, "From explanations to feature selection: Assessing shap values as feature selection mechanism," *Proc.33rd IEEE SIBGRAPH Conference on Graphics, Patterns and Images*, pp. 340–346, 2020.

[17] H. Huang, J. Guan, and C. Feng, "Fluid volume status detection model for patients with heart failure based on machine learning methods," *Heliyon*, vol. 11, no. 1, e41127, 2025.

[18] H. Luo, C. Xiang, and L. Zeng, "Shap based predictive modeling for water quality index calculation in hydrological studies," *Scientific Reports*, 2024.

[19] N. Gehlot, A. Jena, and A. Vijayvargiya, "Surface electromyography based explainable ai fusion framework for hand gesture recognition," *Engineering Applications of Artificial Intelligence*, vol. 137, pp. 109–119, 2024, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2024.109119.

[20] F. Yagin and Y. Gormez, "Hybrid methodology integrating shap for metabolomic analysis in breast cancer patients," *Frontiers in Oncology*, 2024.

[21] E. Ramanujam, K. Thirumalai, and A. Abirami, "Fsshap: Global interpretable feature selection using xai for the classification of spam sms in dravidian languages," *IEEE MultiMedia*, pp. 1–11, 2024. DOI: 10.1109/MMUL.2024.3508765..

[22] M. L. Santos, A. Guedis, and I. S. Gendris, "Shapley additive explanations (shap) for efficient feature selection in rolling bearing fault diagnosis," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 316–341, 2024.

[23] F. Wang, X. Ao, M. Wu, S. Kawata, and J. She, "Explainable deep learning for semg-based similar gesture recognition: A shapley-value-based solution," *Information Sciences*, vol. 672, p. 120 667, 2024, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2024.120667.

[24] M. Kelly, R. Longjohn, and K. Nottingham, *The UCI machine learning repository, https://archive.ics.uci.edu*.