A Comparative Study on the Impact of Categorical Encoding on Black Box Model Interpretability

Hajar Hakkoum[1] and Ali Idri[1,2]
[1]*SPM, ENSIAS, Mohammed V University, Rabat, Morocco*
[2]*Faculty of Medical Sciences, Mohammed VI Polytechnic University, Ben Guerir, Morocco*

Keywords: Interpretability, Black Box, Machine Learning, Data Mining, Categorical Encoding, XAI.

Abstract: This study explores the challenge of opaque machine learning models in medicine, focusing on Support Vector Machines (SVMs) and comparing their performance and interpretability with Multilayer Perceptrons (MLPs). Using two medical datasets (breast cancer and lymphography) and three encoding methods (ordinal, one-hot, and dummy), we assessed model accuracy and interpretability through a decision tree surrogate and SHAP Kernel explainer. Our findings highlight a preference for ordinal encoding for accuracy, while one-hot encoding excels in interpretability. Surprisingly, dummy encoding effectively balanced the accuracy-interpretability trade-off.

1 INTRODUCTION

Machine learning (ML) models, particularly black box models like Support Vector Machines (SVMs), are increasingly utilized in medical fields to provide critical insights and serve as adjuncts to human decision-making (London, 2019) (Idri et al., 2018) (Benhar et al., 2020) (Hosni et al., 2019) (Zerouaoui et al., 2020). Despite their superior performance over more transparent models (Heinrichs and Eickhoff, 2020), such as Decision Trees (DTs), their adoption is hindered by a lack of interpretability, a crucial factor in healthcare where understanding model rationale is as important as the predictions themselves. This has, therefore, made research into interpretability appealing to the research community (Hakkoum et al., 2022).

Interpretability techniques are distinguished by two criteria (Molnar, 2022): 1) whether they explain the behaviour of the black box model globally or locally (single instance), and 2) whether they are agnostic or specific to one type of black box model. Our systematic literature review (SLR) (Hakkoum et al., 2022) on 179 papers investigating interpretability in medicine discovered that 95 (53%) and 72 (40%) papers focused solely on global or local interpretability, respectively, while 10 papers (6%) proposed and/or evaluated both global and local interpretability techniques. Moreover, most of the data types that the selected studies worked on were numeric (46%, 111 papers) and categorical (24%, 59 papers). The categorical features employed were often encoded using ordinal or label categorical encoding (CE), both of which map the numerical values onto an integer in order to represent each category. Label CE can disregard any order a feature might have, such as the degree of malignancy in a cancer prognosis dataset. This can have a negative impact on the relevance of the feature and, therefore, on the performance of the model. It is for this reason that ordinal CE is often used.

The gap in interpretability, coupled with the significant influence of data preprocessing methods like CE on model performance, underscores the need for more focused research in this area. Building on our previous work that investigated Multilayer Perceptrons (MLPs) (Hakkoum. et al., 2023), this study extends the exploration to SVMs, aiming to compare the interpretability and performance of these models in medical applications, particularly considering the impact of different CE techniques.

This study, therefore, compares two well-known interpretability techniques: global surrogates using DT and Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) when used with an SVM trained on two categorical medical datasets: breast cancer (BC) and lymphography (Lymph) (Dua and Graff, 2017). Moreover, it compares the results to our previous analysis on MLP (Hakkoum. et al., 2023). Following the application of three different CEs, namely ordinal, categorical and dummy, the ML

Hakkoum, H. and Idri, A.
A Comparative Study on the Impact of Categorical Encoding on Black Box Model Interpretability.
DOI: 10.5220/0012766300003756

In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 384-391
ISBN: 978-989-758-707-8; ISSN: 2184-285X

models were optimised using the particle swarm optimisation algorithm (PSO) to ensure maximum accuracy. The accuracy of the models was first compared using the Wilcoxon test and the Borda count voting system, after which the CEs were compared using the SkottKnott (SK) statistical test. Moreover, the SK test was repeated on the basis of the fidelity of the interpretability techniques in different settings in order to discover the best CE in each situation. Finally, the effects of the CEs on accuracy and fidelity were compared in order to examine the effect on the trade-off of accuracy versus interpretability. In this respect, the research questions (RQs) listed below will be addressed:

- RQ1: What is the impact of the CEs on accuracy?
- RQ2: Which CE is best for global interpretability?
- RQ3: What is the best CE considering local interpretability?
- RQ4: Which CE best reduces the trade-off between accuracy and interpretability?

This paper's key contributions are: 1) A quantitative evaluation of interpretability techniques. 2) The identification of the effect of CE, along with the best encoding scheme(s) in terms of accuracy and interpretability, and 3) The determination of whether a specific CE alleviates the accuracy-interpretability trade-off.

The remainder of this paper is organised as follows: Section 2 describes related work while Section 3 describes the datasets, along with the metrics and statistical tests used to identify the best performing models. The experimental design used in this empirical evaluation is presented in Section 4. Section 5 presents and discusses the findings, while Section 6 reports conclusions, limitation, and future directions.

2 RELATED WORK

While prior research has extensively explored the accuracy of ML models, the impact of CEs on interpretability remains less examined. Elshawi et al.'s analysis of various interpretability techniques (Elshawi et al., 2021), including global surrogates and SHAP, within a cardio-respiratory fitness dataset highlights the need for more nuanced handling of categorical data. The data, as they presented it, contains categorical features such as gender, race, and reason for test (pains in chest, shortness of breath...). Unfortunately, no further information regarding how categorical data were handled was provided. In general, we believe that ordinal CE is, in most cases, used to maintain the order that the attribute might have. In fact, in some of the datasets in the UCI repository (Dua and Graff, 2017), the categorical variable is already encoded with ordinal CE.

Similarly, Crone et al. and Potdar et al. have underscored the significant influence of CEs like ordinal, one-hot, and dummy on model accuracy, with disparities up to 14% observed across different encoding schemes (Crone et al., 2006; Potdar et al., 2017). These findings underscore the intricate balance between accuracy and interpretability in ML models, particularly in SVMs and MLPs, prompting a deeper investigation into the role of CEs within this context.

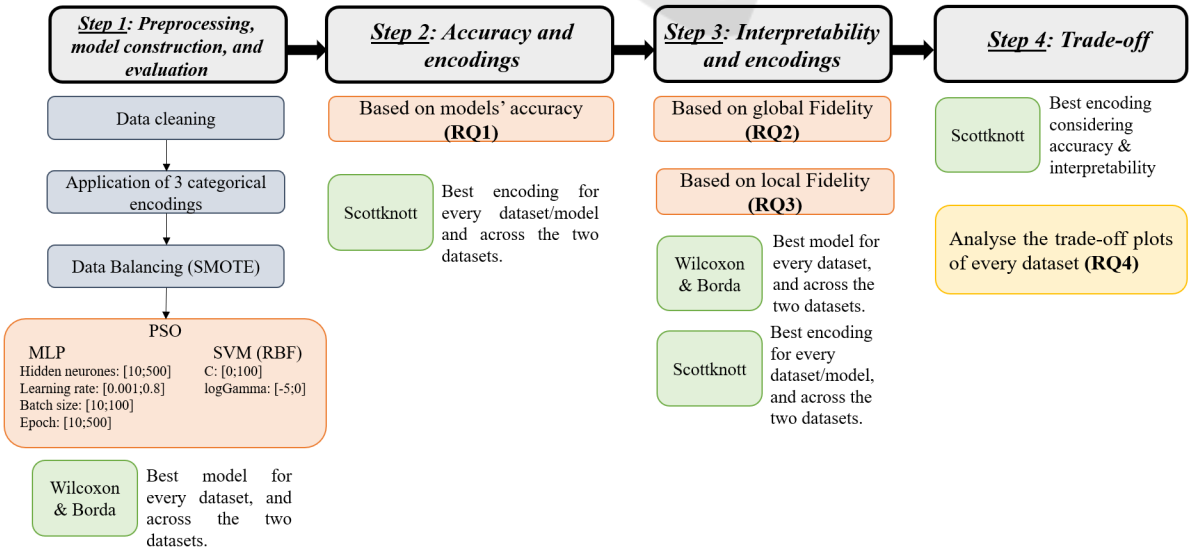

Figure 1: Experimental Design.

Table 1: Performance results.

Data	Encoding	MLP				SVM				Borda Winner
		Acc	F1	AUC	Spearman	Acc	F1	AUC	Spearman	
BC	Ordinal	0.6	0.476	0.398	0.167	0.6	0.153	0.455	-0.067	MLP
	One-hot	0.618	0.399	0.404	0.120	0.654	0.095	0.425	0.003	MLP and SVM
	Dummy	0.618	0.432	0.440	0.145	0.654	0.095	0.367	0.003	MLP and SVM
Lymph	Ordinal	0.758	0.695	0.808	0.497	0.896	0.857	0.555	0.778	SVM
	One-hot	0.827	0.782	0.580	0.641	0.827	0.782	0.656	0.641	SVM
	Dummy	0.793	0.625	0.699	0.583	0.827	0.782	0.575	0.641	SVM

3 DATASETS AND METRICS

This study utilizes two only online categorical datasets on the UCI repository: BC and Lymph (Dua and Graff, 2017). The datasets have different numbers of attributes (9 and 18 respectively) and a very low number of instances (286 and 148 respectively). The BC dataset describes a prognosis task of BC. Therefore the binary classification refers to the recurrence or no recurrence of BC. Meanwhile, four classes of the lymph dataset are available with different numbers of instances: normal find: 2, metastases: 81, malign lymph: 61, and fibrosis: 4. The first and last classes have very few instances: two in normal find and four in fibrosis. Therefore, the multi-classification in the lymph dataset was converted to a binary one by removing the classes normal find and fibrosis. Therefore, for both datasets, we keep our classification task binary.

Moreover, performance is assessed similarly to our previous study (Hakkoum. et al., 2023) using standard metrics such as accuracy, F1-score, Spearman, and Area under the ROC (Receiver operating characteristic) curve (AUC).

Interpretability is examined through the fidelity metric (which compares the black-box model's predictions to the interpretability technique's predictions using the usual classification metrics), the depth and number of leaves in global surrogate models as well as the Mean squared error (MSE) between SHAP and the black box certainties for the test sets.

Cross-validation is used to ensure the robustness of our findings, while the Borda count voting system (de Borda, 1781) help identify the most effective CE in optimizing both performance and interpretability. Moreover, the statistical tests Wilcoxon (for pair comparisons) and SK (for multiple comparisons) were conducted to determine the significance of the difference between methods based on the aforementioned metrics.

Wilcoxon was mainly used to check whether the pair of models compared (MLP and SVM) was statistically different for one or multiple datasets. The Wilcoxon test returns a p-value that can be used to interpret the results of the test. It can be considered as the probability of observing the performance of the two models given the base assumption/hypothesis that they were drawn from a population with the same distribution. In this study, the threshold is fixed at 5%. Therefore, the assumption of the methods belonging to the same distribution was rejected if the p-value was less than 5%.

Table 2: Appearances of SK ranks according to model performance.

Encoding	1st	2nd	3rd
Ordinal	8	4	4
One-hot	7	4	5
Dummy	4	10	2

4 EXPERIMENT

Our experiment, as depicted in Figure 1, comprises four primary stages: 1) Model construction and evaluation including our previous results on MLP (Hakkoum. et al., 2023); 2) We then delve into the effects of CEs on model accuracy, examining how different encoding strategies influence model performance; 3) The third stage focuses on interpretability, exploring how CEs impact both global and local interpretability techniques, evaluated via the fidelity metric; 4) Lastly, we address the trade-off between accuracy and interpretability, analyzing how CEs affect this balance in SVM and MLPs.

5 RESULTS

This section presents and discusses the findings of empirical evaluations carried out in this study in order to answer the RQs. The experiments were performed on a Lenovo Legion laptop with hexa-core Intel Core i7-9750H processor with 16GB of RAM.

Initial data preparation involved cleaning and ad-

Table 3: Results of DT global surrogate.

Dataset	Encoding	MLP global surrogate			SVM global surrogate			Borda Winner
		Spear. fidelity	Depth	Leaves	Spear. fidelity	Depth	Leaves	
BC	Ordinal	0.285	15	67	0.348	14	83	SVM
	One-hot	0.524	12	77	0.344	15	75	MLP
	Dummy	0.671	18	72	0.375	16	65	SVM
Lymph	Ordinal	0.368	6	16	0.454	8	25	MLP
	One-hot	0.716	6	17	0.715	6	17	MLP
	Dummy	0.629	7	19	0.607	7	17	MLP and SVM

dressing class imbalances in the Lymph dataset by removing exceptionally small classes and applying SMOTE to balance the training-validation sets, resulting in 159 no-recurrence and 63 recurrence cases for the BC dataset, and 63 metastases and 50 malignant cases for the Lymph dataset. The model hyperparameters were optimised using PSO on the basis of accuracy with a 10-fold cross validation using only the training-validation set.

5.1 RQ1: What Is the Impact of the CEs on Accuracy?

The concern of this empirical evaluation is the impact of the CE on interpretability. Nevertheless, it starts by investigating the impact of the CE on accuracy in order to study the similarities of the effect of different CEs on accuracy and interpretability, which can explain changes in the trade-off.

Table 1 shows the performances of SVM along with MLP from our previous study (Hakkoum. et al., 2023). The Wilcoxon test was performed on the accuracy of both models with the three CEs for: 1) each dataset, and 2) both datasets. In order to further assess the differences between the models (MLP vs. SVM) the Borda count voting system was carried out to discover whether one model was outperforming the other according to all the metrics.

For each dataset as well as regardless of the dataset, the Wilcoxon test provided a p value higher than 5% which validates the hypothesis of same distribution. Table 1 also shows the Borda winner where SVM always outperformed MLP on Lymph while MLP outperformed only once on BC with ordinal CE. Although the Wilcoxon test showed that the models are not significantly different, it is possible to consider that, according to the Borda count voting system, SVM slightly outperforms MLP.

Investigating the influence of CEs on model accuracy, the SK test was used to evaluate and rank the performance of ordinal, one-hot, and dummy CEs first for each dataset, then across both datasets. The evaluation considered four metrics: Accuracy, F1-score, AUC, and Spearman correlation. The number of appearances of each CE in a SK rank (cluster) was computed by considering the ranks of the CEs for each metric in Table 1, and these are presented in Table 2.

The appearances of each CE in a SK rank was computed by considering each performance metric at a time in different settings (each dataset/model, both datasets/models). The ordinal CE generally outperformed the others, since it came first 8 times, second 4 times and last 4 times, followed by one-hot then dummy. Despite these differences, an aggregated SK analysis across both models and both datasets deemed all three CEs similarly effective.

5.2 RQ2: Which CE Is Best for Global Interpretability?

Exploring global interpretability, we assessed the performance of DT surrogates constructed with various CEs. Despite ordinal encoding's superior accuracy in RQ1, this phase evaluated global surrogate efficacy via Spearman fidelity, tree depth, and leaf count as shown in Table 3.

Based on Table 3, it appears that there is a higher performance for the Lymph dataset when compared to the BC dataset, which was also noticed in the case of model performance (RQ1). When comparing SVM to the MLP, it is also noted that, according to the Borda count voting system, MLP slightly outperformed SVM by 3 wins to 2 with 1 draw. Nevertheless, according to the Wilcoxon test, this difference was not significant, since it led to very high p-values (28%, 100%, and 46% on BC, Lymph, and both datasets, respectively).

The SK clustering statistical test was used to help rank the three CEs on the basis of global surrogate metrics. The number of appearances of each CE in each SK rank was computed according to Spearman fidelity, depth, and number of leaves, and is summarised in Table 4. The one hot CE generally outperformed the others, since it came first 6 times, second 4 times and last twice, followed by dummy and ordinal.

Despite these differences, an aggregated SK analysis across both models and both datasets deemed all three CEs similarly effective.

Table 4: Appearances of SK ranks according to global surrogates.

Encoding	1st	2nd	3rd
Ordinal	4	3	5
One-hot	6	4	2
Dummy	4	5	3

5.3 RQ3: What Is the Best CE Considering Local Interpretability?

This phase investigated local interpretability using SHAP Spearman fidelity to discern the most effective CE in enhancing model interpretability as shown in Table 5. The results indicated that MLP classifiers, particularly with the Lymph dataset, showcased commendable performances. However, the Wilcoxon test's high p-values, both dataset-specific and overall, suggested negligible differences in SHAP fidelity between SVM and MLP models.

Further examination through SK rankings and the frequency of appearances in these rankings as per SHAP fidelity underscored one-hot encoding's superiority, consistently securing top positions as shown in Table 6. Despite these differences, an aggregated SK analysis across both models and both datasets deemed all three CEs similarly effective considering local fidelity.

The influence of CEs on both interpretability and model accuracy was also analysed, exploring if these aspects are affected similarly by CEs. According to the SK test on Spearman fidelity, CEs belong to same cluster, with a nuanced preference for one hot encoding.

The SK clustering of the CEs for global and local fidelities is illustrated in Figure 2. All CEs of global scope appeared to rank better than the local ones. The one hot and dummy CEs of the global surrogates came first, followed by ordinal. Meanwhile, the one hot CE from the local scope was placed in the third cluster, followed by ordinal and dummy in the last cluster. Nonetheless, regardless of the interpretability scope (global or local), the CEs were assigned to the same cluster, with an insignificant preference for the one hot CE.

5.4 RQ4: Which CE Best Reduces the Trade-Off Between Accuracy and Interpretability?

For further comparison, the errors of model and global surrogate were computed along with the MSE of SHAP probabilities against model probabilities. The errors are presented in Table 7 for both models with both datasets, along with the Borda count voting system results, while Figure **XXX** illustrates the differences between the CEs with the average number of errors for both datasets. The analysis revealed that while all CEs produced acceptable interpretability errors (below 0.21), local interpretability errors were notably lower than global ones, with ordinal CE leading in local interpretability and dummy CE excelling globally, particularly for SVMs.

The Borda count voting system's results further refined these insights, showing dummy CE's superiority in balancing the trade-off, as evidenced by its consistent high rankings. Conversely, despite one-hot CE's high interpretability fidelity, it did not secure top ranks in balancing the trade-off according to Borda count, suggesting that dummy CE might offer the best compromise between maintaining model accuracy and ensuring interpretability.

6 CONCLUSIONS AND PERSPECTIVES

Two interpretability techniques (global surrogate and SHAP) were empirically evaluated in this study. The primary goal was to identify the influence of CE on interpretability techniques using two opaque models (SVM and MLP) trained on two medical categorical datasets.

In most cases, the quantitative evaluations revealed that SVM outperformed MLP and that ordinal CE is preferred for better model performance. The one hot CE appeared to provide better interpretability results when using the global surrogate and SHAP. However, according to SK clustering, global surrogate performed slightly better than SHAP when using the Spearman fidelities of both datasets. Finally, the three CEs helped the black box models gain a degree of transparency by reducing the trade-off with the surprising outperformance of the dummy CE when comparing accuracy errors.

Ongoing work concerns studying the effect of continuous attribute encoding on the accuracy and interpretability of ML black box models. In fact, it might be interesting to analyse the effect of many DP

Table 5: SHAP fidelity and Wilcoxon tests.

Dataset	Encoding	MLP Spear. fidelity	SVM Spear. fidelity	Wilcoxon (p value)
BC	Ordinal	-0.330	-0.341	
	One-hot	-0.146	0.103	0.785
	Dummy	0.049	-0.200	
				0.141
Lymph	Ordinal	0.697	-0.134	
	One-hot	0.384	0.298	0.108
	Dummy	0.067	-0.196	

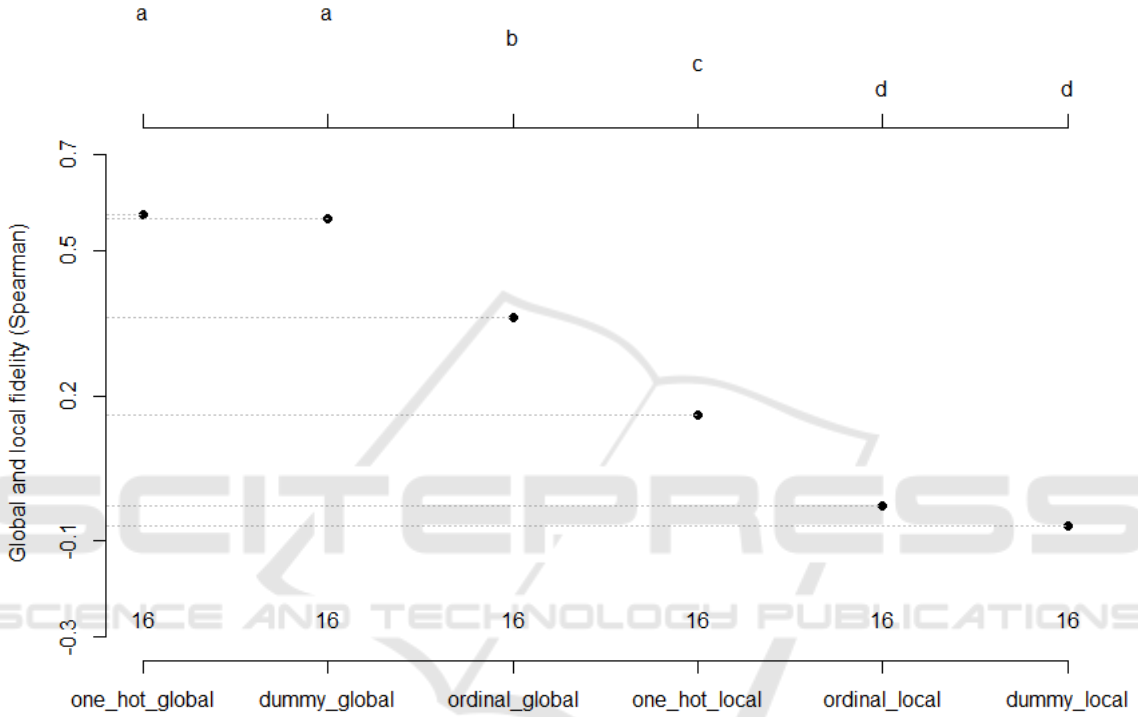

Figure 2: SK clustering for CEs considering global/local scopes.

Table 6: Appearances of SK ranks according to SHAP fidelity.

Encoding	1st	2nd	3rd
Ordinal	1	1	2
One-hot	2	2	0
Dummy	1	1	2

methods.

This study's validity is influenced by the use of only two datasets, potentially limiting the generalizability of the findings. The performance of the models on these datasets may not fully capture the challenges of applying CEs and interpretability techniques to more complex or diverse medical data. Reproducibility efforts must consider the variability inherent in ML models, especially with different hyperparameter configurations.

ACKNOWLEDGEMENTS

This work was conducted under the research project "Machine Learning based Breast Cancer Diagnosis and Treatment", 2020-2023. The authors would like to thank the Moroccan Ministry of Higher Education and Scientific Research, Digital Development Agency (ADD), CNRST, and UM6P for their support.

REFERENCES

Benhar, H., Idri, A., and Fernández-Alemán, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195:105635.

Crone, S. F., Lessmann, S., and Stahlbock, R. (2006). The impact of preprocessing on data mining: An evalua-

Table 7: Model and global/local errors.

Dataset	Encoding	MLP			Borda Ranks	SVM			Borda Ranks
		Accuracy error	Global error	SHAP MSE		Accuracy error	Global error	SHAP MSE	
BC	Ordinal	0.4	0.128	0.041	1	0.4	0.237	0.030	2
	One-hot	0.382	0.164	0.090	3	0.346	0.237	0.063	3
	Dummy	0.382	0.164	0.076	2	0.346	0.219	0.060	1
Lymph	Ordinal	0.242	0.173	0.029	3	0.104	0.173	0.039	1
	One-hot	0.173	0.138	0.107	2	0.173	0.138	0.065	3
	Dummy	0.207	0.035	0.046	1	0.173	0.069	0.061	2

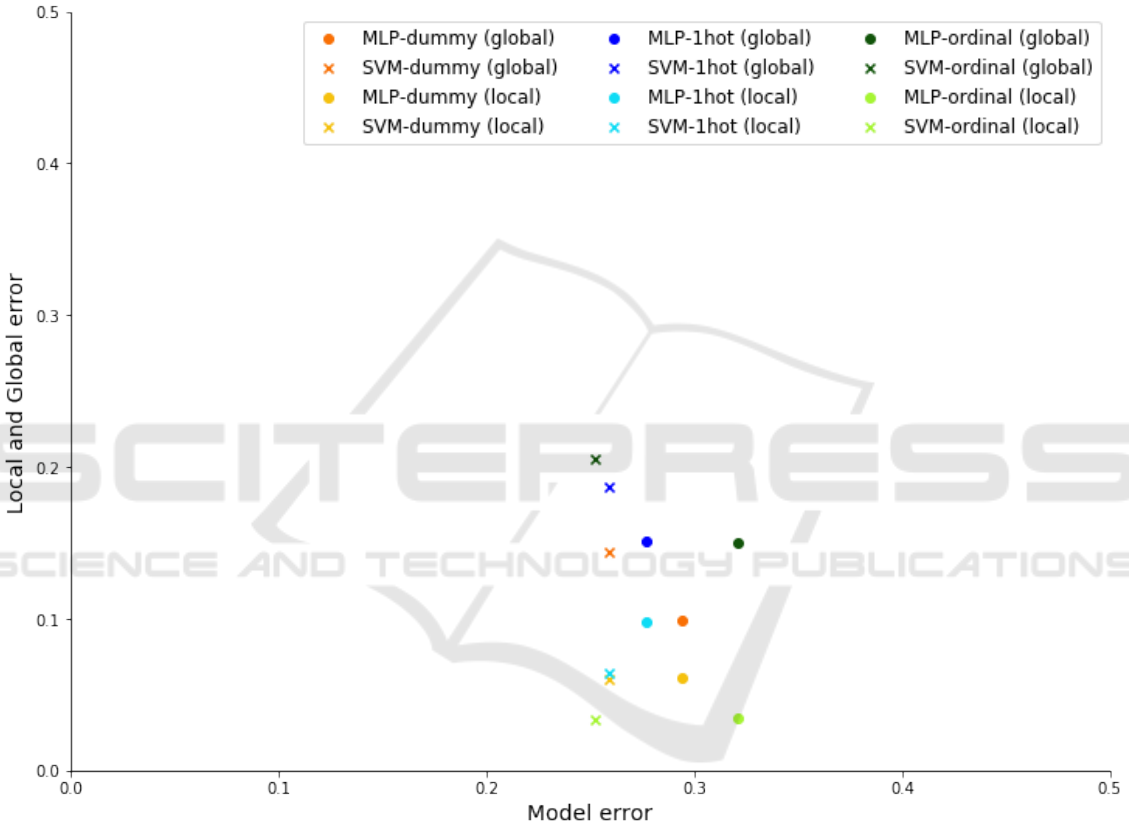

Figure 3: Model vs. global and local interpretability techniques trade-off based on error.

tion of classifier sensitivity in direct marketing. *European Journal of Operational Research*, 173(3):781–800.

de Borda, J.-C. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Elshawi, R., Al-Mallah, M., and Sakr, S. (2021). On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak*, 19(146).

Hakkoum, H., Abnane, I., and Idri, A. (2022). Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, 117:108391.

Hakkoum., H., Idri., A., Abnane., I., and Luis Fernades-Aleman., J. (2023). Does categorical encoding affect the interpretability of a multilayer perceptron for breast cancer classification? In *Proceedings of the 12th International Conference on Data Science, Technology and Applications - DATA*, pages 351–358. INSTICC, SciTePress.

Heinrichs, B. and Eickhoff, S. B. (2020). Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, 41(6):1435–1444.

Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., and Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. *Computer*

Methods and Programs in Biomedicine, 177:89–112.

Idri, A., Chlioui, I., and Ouassif, B. E. (2018). A systematic map of data analytics in breast cancer. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '18, New York, NY, USA. Association for Computing Machinery.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1):15–21.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Potdar, K., Pardawala, T., and Pai, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175:7–9.

Zerouaoui, H., Idri, A., and El Asnaoui, K. (2020). Machine learning and image processing for breast cancer: A systematic map. In Rocha, Á., Adeli, H., Reis, L. P., Costanzo, S., Orovic, I., and Moreira, F., editors, *Trends and Innovations in Information Systems and Technologies*, pages 44–53, Cham. Springer International Publishing.