ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094701

INTERPRETABILITY IN THE CONTEXT OF SEQUENTIAL COST-SENSITIVE FEATURE ACQUISITION

Yasitha Warahena Liyanage[1]*, Daphney–Stavroula Zois*[2]

[1]Microsoft Corporation, Redmond, WA

[2]Department of Electrical and Computer Engineering, University at Albany, State University of New York, Albany, NY

yliyanage@microsoft.com, dzois@albany.edu

ABSTRACT

Despite the popularity of complex machine learning models, domain experts often struggle to understand and are reluctant to trust them due to lack of intuition and explanation of their predictions. Moreover, these cannot be used in many real–world applications, where features are not readily available but acquired at a cost. To address the latter challenge, dynamic instance–wise joint feature selection and classification selects both the order and the number of features to individually classify each data instance when features are sequentially acquired one at a time. Herein, its model–based and post hoc interpretability is demonstrated validating its utility in high–stakes applications. As a case study, predicting the credit risk of an individual based on financial and other data is considered. Experimental results show that the proposed method is indeed interpretable without sacrificing prediction accuracy.

Index Terms— model–based interpretability, instance–level sparsity, glass–box models, explainability, datum–wise decisions.

1. INTRODUCTION

Recent advances in machine learning set paths to complex function approximators that can achieve high performance in many domains [1, 2]. However, humans are often reluctant to deploy such complex models in practice, particularly in health care, criminal justice, and financial markets, since they do not have formal justifications about what the model is doing and why it outputs specific classification decisions [3, 4].

Inherently interpretable machine learning models can be used to discover relevant knowledge about domain relationships in data, debug or justify the model and its outputs, and control and improve the model [5–7]. Examples include the generalized additive models (GAMs) [8] and the decision tree. GAMs combine single–feature models through a linear function, identifying the contribution of individual features to the model output. Common GAMs include the logistic regression and the explainable boosting machines, which use linear and boosted decision tree shape functions [9].

This material is based upon work supported by the National Science Foundation under Grants ECCS–1737443 & CNS–1942330.

Using a sparse set of features to classify data instances is essential for model interpretability [6, 10], since we can explicitly observe which features contribute to each model output. However, in standard settings, sparsity is achieved globally by incorporating a regularizer to the model parameters (e.g., GAMs with L1–norm regularizer), where the *same* subset of features is used to classify all test instances. In contrast, the decision tree achieves *instance–level sparsity* by evaluating features along different decision paths, using different features to classify different test instances. Nonetheless, it uses a greedy approach to build the tree structure, where locally optimal splits are obtained at every tree node [11], hence, using more features than necessary.

In our prior work [12], we proposed an algorithm for Instance–wise Feature selection and Classification with optimum feature Ordering (IFCO), which dynamically selects both the *order* and the *number of features* to classify each data instance *individually* when features sequentially arrive one at a time during testing. Herein, the model–based and post hoc interpretability of IFCO is justified. Experimental results on a credit risk classification task show that IFCO can be used in high–stakes applications, where model interpretations are required without sacrificing test accuracy. Other sequential methods can be potentially analyzed in a similar way to verify their interpretability.

2. BACKGROUND

In [12], we introduced the following optimization problem:

$$\underset{\sigma,\sigma(R),D_{\sigma(R)}}{\text{minimize}}\ \mathbb{E}\left\{\sum_{k=1}^{R} e\big(F_{\sigma(k)}\big)+\mathcal{L}(D_{\sigma(R)})\right\}, \tag{1}$$

where σ denotes a permutation of the features, $\sigma(R)$ denotes the number of features acquired before the framework terminates assuming feature ordering σ, and $D_{\sigma(R)}$ denotes the classification rule used. The term $e(F_k) > 0$, $k \in \{1, \ldots, K\}$, denotes the cost of acquiring feature F_k, and $\mathcal{L}(D_{\sigma(R)}) = \sum_{j=1}^{L}\sum_{i=1}^{L} Q_{ij} P\big(D_{\sigma(R)} = j, \mathcal{C} = c_i\big)$ is the cost associated with the classification rule $D_{\sigma(R)}$, where c_i denotes an assignment to class variable $\mathcal{C}$, and $Q_{ij} \geqslant 0$

represents the cost of selecting class c_j when the true class is c_i, $i, j \in \{1, \ldots, L\}$. The optimum classification strategy $D^*_{\sigma(R)}$ for any number $\sigma(R)$ and ordering σ was shown to be:

$$D^*_{\sigma(R)} = \arg \min_{1 \leqslant j \leqslant L} \left[Q_j^T \pi_{\sigma(R)} \right], \tag{2}$$

where $Q_j \triangleq [Q_{1,j}, Q_{2,j}, \ldots, Q_{L,j}]^T$, $\pi_{\sigma(k)} \triangleq [\pi^1_{\sigma(k)}, \pi^2_{\sigma(k)}, \ldots, \pi^L_{\sigma(k)}]^T$, and $\pi^i_{\sigma(k)} \triangleq P\big(\mathcal{C} = c_i | F_{\sigma(1)}, \ldots, F_{\sigma(k)}\big)$. The posterior probability vector $\pi_{\sigma(k)} \in [0,1]^L$ is updated recursively via Bayes' rule as follows:

$$\pi_{\sigma(k)} = \frac{\text{diag}\left(\Delta\big(F_{\sigma(k)} | F_{\sigma(1)}, \ldots, F_{\sigma(k-1)}, \mathcal{C}\big)\right) \pi_{\sigma(k-1)}}{\Delta^T(F_{\sigma(k)} | F_{\sigma(1)}, \ldots, F_{\sigma(k-1)}, \mathcal{C}) \pi_{\sigma(k-1)}}, \tag{3}$$

where $\Delta\big(F_{\sigma(k)} | F_{\sigma(1)}, \ldots, F_{\sigma(k-1)}, \mathcal{C}\big) \triangleq [P(F_{\sigma(k)} | F_{\sigma(1)}, \ldots, F_{\sigma(k-1)}, c_1), \ldots, P(F_{\sigma(k)} | F_{\sigma(1)}, \ldots, F_{\sigma(k-1)}, c_L)]^T$, $\text{diag}(A)$ represents a diagonal matrix with elements of vector A, $\pi_{\sigma(0)} \triangleq [p_1, p_2, \ldots, p_L]^T$, and $p_i = P(\mathcal{C} = c_i)$. The optimum ordering σ^* and the optimum number $\sigma^*(R^*)$ of features were then derived using dynamic programming:

$$\begin{aligned} &\bar{J}_k(\pi_{\sigma^*(k)}) = \min \left[g(\pi_{\sigma^*(k)}), \bar{\mathcal{A}}_k(\pi_{\sigma^*(k)}) \right], k = 0, \ldots, K-1, \\ &\bar{J}_K(\pi_{\sigma^*(K)}) = g(\pi_{\sigma^*(K)}), \end{aligned} \tag{4}$$

where $\bar{\mathcal{A}}_k(\pi_{\sigma^*(k)}) \triangleq \min_{F_{k+1} \in Z_k} \Big[e(F_{k+1}) + \sum_{F_{k+1}} \Delta^T(F_{k+1} | F_{\sigma^*(1)}, \ldots, F_{\sigma^*(k)}, \mathcal{C}) \pi_{\sigma^*(k)} \bar{J}_{k+1}\big(\pi_{\sigma^*(k+1)}\big) \Big]$, $g(\pi_{\sigma^*(k)}) \triangleq \min_j [Q_j^T \pi_{\sigma^*(k)}]$, and Z_k is the set of remaining features at stage k. The optimum number $\sigma^*(R^*)$ of features is equal to the first $k < K$ features for which $g(\pi_{\sigma^*(k)}) \leqslant \bar{\mathcal{A}}_k(\pi_{\sigma^*(k)})$, or $\sigma^*(R^* = K)$ if there are no more features to be acquired.

Finally, we proposed IFCO using the fact that $g(\pi_{\sigma^*(k)})$ and $\bar{\mathcal{A}}_k(\pi_{\sigma^*(k)})$ are continuous, concave, and piecewise linear functions. To classify a test instance, IFCO starts by setting $k = 0$ and assigning the prior distribution of the class variable $\mathcal{C}$ to the posterior probability $\pi_{\sigma^*(0)}$. If $\bar{J}_0(\pi_{\sigma^*(0)})$ belongs to a hyperplane of $g(\pi_{\sigma^*(0)})$, IFCO stops and classifies the instance using Eq. (2). Otherwise, it acquires the feature associated with the hyperplane of $\bar{\mathcal{A}}_0(\pi_{\sigma^*(0)})$, increments k by 1, and updates the posterior probability $\pi_{\sigma^*(0)}$ using Eq. (3). IFCO repeats these steps until it stops or there are no more features available (see [12] for more details).

3. INTERPRETABILITY

The glass–box nature of IFCO (*model–based interpretability*) is demonstrated herein, such that a human can understand its behavior and which factors influence its decision–making process. The relationships that IFCO has learned from a given dataset (*post hoc interpretability*) [5] are also analyzed.

We consider logistic regression (LR) with L1–norm regularizer, and decision tree (DT) as baselines, since they are well–studied machine learning models that are interpretable on a modular level [6]. For demonstration purposes, we use the German credit risk dataset [13]. The goal is to classify individuals as *high* or *low credit risk* based on a set of 20 features (see Table 1). Similar observations can be made in other domains (e.g., IMDB movie review classification [14]), but are not included herein due to space limitations.

Table 1. Credit risk dataset features.

Feat.	Description	Feat.	Description
F_1	Checking account status	F_{11}	Present residence
F_2	Duration in months	F_{12}	Property
F_3	Credit history	F_{13}	Age in years
F_4	Purpose of the credit	F_{14}	Other installment plans
F_5	Credit amount	F_{15}	Housing
F_6	Savings account status	F_{16}	Existing credits
F_7	Present employment (years)	F_{17}	Job
F_8	Installment rate	F_{18}	Number of dependents
F_9	Personal status	F_{19}	Telephone
F_{10}	Other debtors	F_{20}	Foreign worker

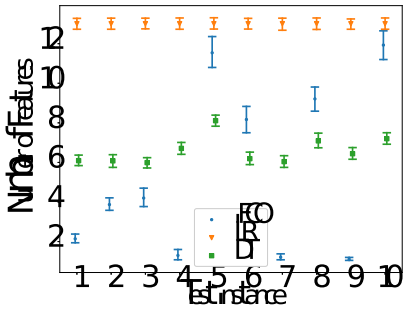

Fig. 1. Number of features used to classify ten random test instances in the credit risk dataset. Range corresponds to the 95% confidence interval.

3.1. Model–Based Interpretability

Model–based interpretability considers three criteria: *sparsity*, *simulatability*, and *modularity* [5]. Sparsity refers to using a sparse set of features to classify each data instance. Simulatability represents the ability to simulate and reason about the entire decision–making process. Modularity denotes the ability to interpret the meaningful portions of the decision–making process independently.

3.1.1. Sparsity

IFCO imposes *instance–level sparsity* by utilizing the feature acquisition cost, i.e., $\sum_{k=1}^{R} e\big(F_{\sigma(k)}\big)$, in the optimization function in Eq. (1). Specifically, acquiring features in different orderings σ and terminating at different number $\sigma(R)$ of features results in different accumulated costs. By penalizing these accumulated feature acquisition costs, IFCO optimizes the number of features used to classify individual data instances. It also uses a varying number of features to classify different data instances.

Similar to IFCO, DT uses a varying number of features to classify different data instances by evaluating features along different decision paths in the tree. In contrast, LR imposes global sparsity by utilizing the L1–norm penalty on the model parameters in the optimization function. Global

sparsity degrades the model's interpretability since it uses the same subset of features to classify all test instances. However, before interpreting a sparse solution, the stability of the sparsity should be validated [5]. In particular, if the sparsity varies drastically due to a small perturbation in the training dataset, the resulting interpretations are meaningless. Herein, we train 100 instances of each machine learning model using 100 bootstraps of the training data and observe the variation of the number of features used to classify a fixed set of test instances (see Fig. 1). Observe that the variation in the number of features used to classify the same test instance is less than one feature with 95% confidence. Therefore, the instance–level sparsity of IFCO and DT and the global sparsity of LR are stable to small perturbations in the training dataset.

3.1.2. Simulatability

A human can simulate and reason about IFCO's entire decision–making process. The functions $\bar{J}_k(\pi_{\sigma^*(k)}), k = 0, \ldots, K$, in Eq. (4) can be decomposed into linear hyperplanes because $g(\pi_{\sigma^*(k)})$ and $\bar{\mathcal{A}}_k(\pi_{\sigma^*(k)})$ are continuous, concave, and piece–wise linear [12]. Hence, there are unique regions in the domain of $\bar{J}_k(\pi_{\sigma^*(k)})$, i.e., the posterior probability space generated by $\pi_{\sigma^*(k)}$, that determine whether to acquire a specific feature or stop the acquisition process. Fig. 2 illustrates this decision–making process for the credit risk dataset. At stage $k = 0$, IFCO starts by assigning the prior probability $P(\textit{credit risk} = \text{high})$ to the posterior probability $\pi_{\sigma^*(0)}$. If this probability falls in the blue region, IFCO acquires F_1. If it falls in the orange region, IFCO acquires F_6. Otherwise, IFCO stops and reaches a decision, i.e., the person is classified as *low* or *high credit risk* if this probability falls in the green or red region, respectively. If IFCO decides to acquire a feature (i.e., F_1 or F_6), it updates the posterior probability $\pi_{\sigma^*(0)}$ using Eq. (3) and continues the decision–making process similarly in the next stage, as shown in Fig. 2.

DT is also a simulatable model due to its hierarchical decision–making process [5]. Each node compares the feature value with a fixed threshold and decides whether to follow the left or right branch. Final decisions are at the leaves. However, DT uses a greedy approach to learn the tree structure, where locally optimal splits are obtained at every tree node [11]. In contrast, IFCO optimizes the *instance–wise feature ordering*, hence uses fewer features on average to classify data instances compared to DT (see Table 2). In LR, a human only needs to compute the dot product between the feature vector and the corresponding weight vector to obtain a classification decision. Each weight is proportionate to the effect of the corresponding feature on the class variable when the rest features are kept unchanged.

3.1.3. Modularity

IFCO enforces modularity by employing a sequential decision–making process based on a sufficient statistic, i.e., the posterior probability vector $\pi_{\sigma^*(k)}$, that is recursively updated as seen in Eq. (3). At each stage k, the only information required for this update is the observation vector $\Delta\big(F_{\sigma^*(k)}|F_{\sigma^*(1)}, \ldots, F_{\sigma^*(k-1)}, \mathcal{C}\big) = [P(F_{\sigma^*(k)}|F_{\sigma^*(1)}, \ldots, F_{\sigma^*(k-1)}, c_1), \ldots, P(F_{\sigma^*(k)}|F_{\sigma^*(1)}, \ldots, F_{\sigma^*(k-1)}, c_L)]^T$. In other words, the posterior probability decomposes into the probability of each feature given the already acquired features and the class variable. Note that probabilistic models can enforce modularity by specifying a conditional independence structure, making it easier to reason about different parts of a model independently [5]. IFCO adopts such an assumption, which simplifies the observation vector to $\Delta\big(F_{\sigma^*(k)}|\mathcal{C}\big) = [P(F_{\sigma^*(k)}|c_1), \ldots, P(F_{\sigma^*(k)}|c_L)]^T$. This assumption helps to decompose the posterior probability into simple and meaningful portions in terms of the probability of each feature given the class variable. It also speeds up computations and enables prediction–level interpretations.

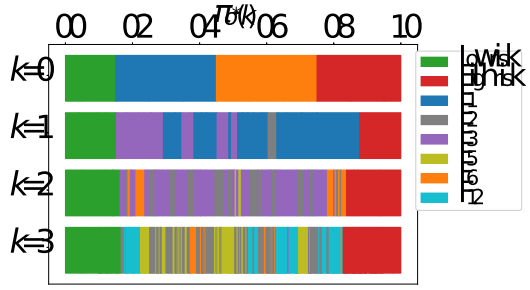

Fig. 2. First four stages of the IFCO's decision making process for the credit risk dataset.

LR inherits modularity by having a decision function based on an affine transformation of the input feature space [8], while each node in DT can be viewed as a modular block that contributes to the final classification decision.

3.2. Post Hoc Interpretability

This section analyzes the dataset– and prediction–level information learned by IFCO about the credit risk dataset.

3.2.1. Dataset–level Interpretations

Partial Dependence: Partial dependence captures the marginal effects of an individual feature on the output of a machine learning model [15]. Specifically, for feature F_i, the partial dependence function is approximated by $PD(F_i) \approx \frac{1}{N}\sum_{n=1}^{N} \hat{f}(F_i, \bar{F}_i^{(n)})$, where $\hat{f}$ is the model output, $\bar{F}_i^{(n)}$ is the nth training instance without feature F_i, and N is the total number of training instances. Fig. 3 shows the partial dependence functions of two features on the probability of credit risk being high. Observe that the probability of credit risk being high decreases as the feature *savings account status* increases. Feature *savings account status* $= 4$ represents "no known savings accounts" [13], which increases credit risk. In contrast, feature *age* does not seem to affect credit risk.

Feature Importance: IFCO computes feature importance by averaging the number of times each feature contributes to a

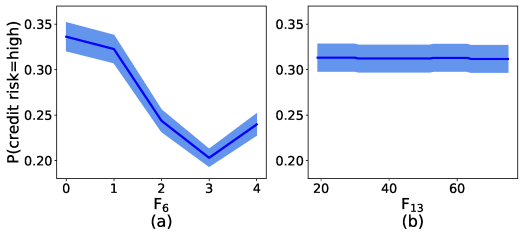

Fig. 3. Partial dependence plots of features (a) *savings account status*, $PD(F_6)$, and (b) *age*, $PD(F_{13})$, using IFCO. *savings account status* $\in \{0, 1, \ldots, 4\}$ and *age* is in years. The dark blue line represents the mean, while the light blue region represents the 95% confidence interval.

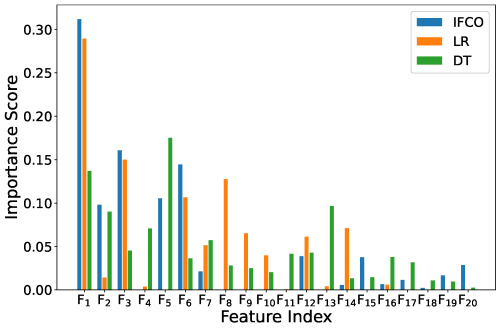

Fig. 4. Normalized feature importance for credit risk dataset.

particular classification decision. Fig. 4 shows the feature importance for each feature in the credit risk dataset. Note that both IFCO and LR choose *checking account status* and *credit history* as the top two important features. On the other hand, DT selects *credit amount* as the most important feature, which ranks fourth according to IFCO.

Accuracy Stability: Test accuracy should be stable for any perturbations in the training data [6]. Herein, we compute out–of–sample accuracy using 10–fold cross–validation (see Table 2). In addition to interpretable machine learning models, we also consider a *black–box* model, i.e., gradient boosted trees (XGB) [16], which provides an upper bound on the achievable accuracy. The deviation in the out–of–sample accuracy achieved by IFCO is ± 0.04 over the 10–folds. Hence, the accuracy is stable to any perturbations in the data. Note that XGB achieves the highest accuracy (i.e., only 0.1% better), but requires 3.4 times more features than IFCO.

3.2.2. *Prediction–level Interpretations*

This section analyzes the ability of IFCO to provide prediction–level interpretations using 4 test instances of the credit risk dataset. Recall that IFCO assigns the data instance to the class with the highest posterior probability [12]. Fig. 5(a) represents a correctly predicted low credit risk instance, where IFCO acquires features F_1, F_3, and F_6, representing that the person has a bad checking account status, a good credit history, and a good savings account status, respectively. Note that having a bad checking account status is discounted by having a good credit history and a good savings account status. Fig. 5(c) represents a correctly predicted high credit risk instance, where IFCO acquires features F_1, F_3, F_2, and F_{12}, representing that the person has a bad checking account status, a bad credit history, a credit history of 36 months, and no known property, respectively. Fig. 5(b) represents a high credit risk instance incorrectly predicted as low, where IFCO acquires features F_1, F_3, F_6, F_2, F_5, F_{20}, and F_{12}, representing that the person has a bad checking account status, a good credit history, a bad savings account status, a credit history of 12 months, an existing credit amount of 1056, is a foreign worker, and owns real state property, respectively. Having a bad checking and savings account status is discounted by having a good credit history with low existing credit amount and owning real state property, hence incorrectly predicting the person to be low credit risk. Fig. 5(d) represents a low credit risk instance predicted as high, where IFCO acquires features F_1, F_3, F_2, and F_5, representing that the person has a bad checking account status, a moderate credit history, a credit history of 39 months, and an existing credit amount of $11,760$, respectively. The high credit amount within a short credit length and the bad checking account status justify IFCO's high–risk prediction.

Table 2. Accuracy ($\pm$ standard deviation) and average number of acquired features (Feat.).

Method	Accuracy	Feat.	Method	Accuracy	Feat.
IFCO	$\mathbf{0.754} \pm 0.040$	**5.85**	LR	0.740 ± 0.034	14.0
DT	0.70 ± 0.044	6.78	XGB	$\mathbf{0.755} \pm 0.037$	19.9

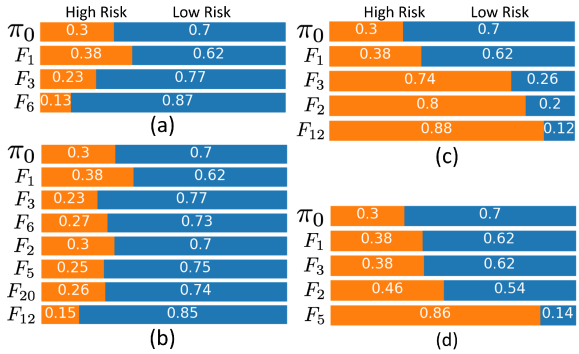

Fig. 5. Variation of $P(\textit{credit risk} = \text{high}|F_1, \ldots, F_k)$ (in orange) and $P(\textit{credit risk} = \text{low}|F_1, \ldots, F_k)$ (in blue) for 4 test instances in the credit risk dataset.

4. CONCLUSION

This paper demonstrates the glass–box nature of IFCO, an algorithm that dynamically selects both the order and the number of features to individually classify each data instance when features sequentially are sequentially acquired one at a time during model testing. Model–based interpretability is shown using instance–level sparsity, simulatability, and modularity. Dataset– and prediction–level interpretations are analyzed during the post hoc stage. Experimental results validate that IFCO outperforms state–of–the–art interpretable models while applicable in high–stakes applications, where model interpretations are necessary. In the future, we plan to extend this model to capture causal effects in the data.

5. REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[3] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[4] Cynthia Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[6] Christoph Molnar, *Interpretable machine learning*, Lulu. com, 2020.

[7] Roberto San Millán-Castillo, Luca Martino, Eduardo Morgado, and Fernando Llorente, "An exhaustive variable selection study for linear models of soundscape emotions: rankings and Gibbs analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2460–2474, 2022.

[8] Trevor J Hastie and Robert J Tibshirani, *Generalized additive models*, Routledge, 2017.

[9] Yin Lou, Rich Caruana, and Johannes Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 150–158.

[10] Georg Heinze, Christine Wallisch, and Daniela Dunkler, "Variable selection–a review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[11] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al., *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2001.

[12] Yasitha Warahena Liyanage and Daphney-Stavroula Zois, "Optimum feature ordering for dynamic instance–wise joint feature selection and classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3370–3374.

[13] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[14] Yasitha Warahena Liyanage, Daphney-Stavroula Zois, and Charalampos Chelmis, "Dynamic Instance-Wise Classification in Correlated Feature Spaces," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 6, pp. 537–548, 2021.

[15] Jerome H Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[16] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.