Fall 1-1-2022

# Towards a Performance-explainability-fairness Framework for Benchmarking ML Models

Shuvro Chakrobartty

Omar El-Gayar

# Towards a Performance-explainability-fairness Framework for Benchmarking ML Models

Shuvro Chakrobartty
*Dakota State University*, shuvro.chakrobartty@trojans.dsu.edu

Omar El-Gayar
*Dakota State University*, omar.el-gayar@dsu.edu

# Towards a Performance-explainability-fairness Framework for Benchmarking ML Models

*Completed Research Paper*

**Shuvro Chakrobartty**
Dakota State University
shuvro.chakrobartty@trojans.dsu.edu

**Omar El-Gayar**
Dakota State University
omar.el-gayar@dsu.edu

## Abstract

Artificial Intelligence (AI) holds great promise in beneficial, accurate, and effective predictive and real-time decision-making in a wide range of use cases. However, there are concerns regarding potential risks, harm, trust, and fairness issues arising from some AI algorithms' opacity and potential unfairness because of their un-explainability and concern with objectivity. This study proposes a framework for evaluating a machine learning model that incorporates explainability for AI fairness as currently, no such framework exists. We evaluate its applicability with a classification problem using multiple classifiers. The experimental case study demonstrates the successful application of the performance-explainability-fairness framework to the classification problem. The framework can guide means for improving fairness in machine learning models.

**Keywords**

Artificial intelligence, Explainability, Fairness.

## Introduction

While the very first AI systems were easily interpretable, recent years have witnessed the rise of opaque (black-box) decision systems such as Deep Neural Networks (DNNs) (Barredo Arrieta et al. 2020). Black-box approaches do not foster trust and acceptance of ML among end-users (Holzinger et al. 2017). The opposite of black-box-ness is transparency, i.e., a direct understanding of the mechanism by which a model works as it makes a decision (Barredo Arrieta et al. 2020).

Explainability is an essential aspect of trust since trust would depend on the visibility that a human has into the working of the AI system. Therefore, DNN and other algorithms should provide human-understandable justifications for its output, leading to insights into the AI system's inner workings. Interpretable models can explain why a certain prediction was made for a specific patient by showing characteristics that led to the prediction. Therefore, lack of interpretability limits otherwise powerful deep and ensemble learning models in critical domains such as medical decision support (Lundberg et al. 2018).

There is no apparent consensus within the literature as to the definition of fairness, and the fairness metrics for any given ML model should be given in each situation (Mehrabi et al. 2021; Verma and Rubin 2018). However, it is understood that fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making (Makhlouf et al. 2021). For example, criminal-sentencing and hiring-decision algorithms have been shown to discriminate against people of certain ethnic or gender backgrounds because of representation bias (Du et al. 2020; Ho 2019). Multiple definitions of fairness and mathematical formulas have been proposed, such as equal odds, positive predictive parity, counterfactual fairness, etc. Verma and Rubin (2018) collected the most prominent definitions of fairness for the algorithmic classification problem explaining the rationale behind these definitions, and demonstrated them on a single unifying case study. Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice (Mehrabi et al. 2021).

There is an obvious tradeoff between the accuracy of prediction and the transparency of algorithms. The European Union's General Data Protection Regulation (GDPR) provides a right to explanation for the individual on automated decision-making for them that has "legal effects" on the individual. Therefore, in addition to their prediction performance, machine learning methods have to be assessed on how they can explain their decisions (Fauvel et al. 2020). While Fauvel et al. (2020) proposed a performance-explainability analytical framework to benchmark ML algorithms, the framework does not account for the fairness of the models. Accordingly, this paper aims to extend the framework by incorporating the AI fairness aspect to build a comprehensive framework for evaluating machine learning model and demonstrate its applicability with a classification problem using multiple classifiers. The outcome of this research would help ML researchers and developers benchmark and identify the best ML model that can be deployed, satisfying the explainability and fairness criteria. Additionally, this helps the business deploy models whose outcome would be more accepted by the users and build trust. Moreover, the end-users would be benefited because they would be able to trust the decisions of the ML models. This research project also contributes to the literature related to ML models' performance, explainability, and fairness topics.

The rest of the paper is organized as follows: first we elaborate on the notion of explainability, performance, fairness in the context of prior work, then we propose a performance-explainability-fairness framework. Later, we demonstrate the application of the framework to a classification problem. Following that we analyze the outcomes and present its results. Finally, we discuss some future research directions and conclude the paper.

## Background and Related Works

Explainable AI is not a new field since, in expert systems of the 1980s, there were reasoning architectures to support an explanation function for complex AI systems (Holzinger 2018). In expert system-based AI, human knowledge is first codified, then an inference engine is used to provide an expert decision to a non-expert user through an interface (London 2019). This is an explainable system by design since the inference engine follows specific rules to make the decision. Explainable AI is an important topic in the context of ML models use within the medical domain (Chakrobartty and El-Gayar 2021). Guidotti et al. (2018) provides a classification of the problems addressed in the literature with respect to the notion of explanation and the type of black box system.

Fairness is a highly desirable human value in day-to-day decisions that affect human life. In recent years many successful applications of AI systems have been developed, and increasingly, AI methods are becoming part of many new applications for decision-making tasks that were previously carried out by human beings. Questions have been raised 1) can the decision be trusted? 2) is it fair? Overall, are AI-based systems making fair decisions, or are they increasing the unfairness in society? This is because defining fairness is not easy, as stakeholders are unlikely to agree on "fair" in different spheres of life. Moreover, something may be deemed fair in one context but may seem unfair in another context. Castelnovo et al. (2022) discussed some important aspects about the relationships between fairness metrics highlighting the clash of individual vs. group as well as observational vs. causality-based fairness.

The performance of a machine learning method can be assessed by the extent to which it correctly predicts unseen instances. Various metrics such as accuracy, F-score, Area Under the Receiving Operating Characteristic (ROC) Curve score commonly used to measure the performance of a classification model. Gunning and Aha (2019) identified the inherent tension between ML performance, such as predictive accuracy and explainability. Often the highest-performing methods such as Deep Learning (DL) are the least explainable, and the most explainable method such as decision trees, are the least accurate (Gunning and Aha 2019). With a critical decision-making process in the medical domain, one cannot take precedence over another. We see researchers try to achieve a balance by developing a blended system to optimize for both of these characteristics in identifying patients at a high risk of death (Kanda et al. 2020).

In that regard, Fauvel et al. (2020) proposed a performance-explainability analytical framework by using a set of characteristics that systematize the performance-explainability assessment to evaluate and benchmark ML algorithms. Moreover, Naylor et al. (2021) introduced a framework through which practitioners and researchers can assess the frontier between a model's predictive performance and the quality of its available explanations. However, there is a gap within the literature regarding its applicability to different classifiers and introducing fairness in addition to performance and explainability when benchmarking ML algorithms for suitability to a particular use case. Hence, our research incorporates AI fairness characteristics to extend the framework so that fairness evaluation can be part of the framework making it comprehensive for benchmarking ML algorithms for use cases where the algorithm's fairness is important. As for the methodological framework of this research first we identify a classification problem and its corresponding dataset. Through literature search we identify the best performing ML models for the dataset. We extend Fauvel's framework and evaluate the framework with the identified best performing model class for the same dataset, then we analyze and interpret the result.

## Performance-explainability-fairness Framework

The framework aims to respond to the different questions an end-user may ask to take an informed decision based on the predictions made by a machine learning model. The evaluation characteristics and the questions along with the assessment values are described in Table 1. For explainability, Fauvel et al.'s position their framework as a further development of the fourth step of the method described in Hall et al. (2019) by detailing a set of explanations characteristics that systematize the assessment of existing methods without including application-specific implementation constraints like time, memory usage and privacy (Fauvel et al. 2020). Table 1 summarizes the framework.

| **Evaluation Characteristics** | **Question** | **Assessment Answer Values** |
|---|---|---|
| Performance | What is the level of performance of the model? | Best, Similar, Below |
| Comprehensibility | Is the model comprehensible? | Black-box, White-box |
| Granularity | Is it possible to get an explanation for a particular instance? | Global, Local, Global & Local |
| Information type | Which kind of information does the explanation provide? | Importance, Patterns, Causal |
| Faithfulness | Can we trust the explanations? | Imperfect, Perfect |
| User category | What is the target user category of the explanations? | Machine Learning Expert, Domain Expert, Broad Audience |

**Table 1. Fauvel et al.'s performance-explainability analytical framework**

We extend the framework (Table 2) by adding two characteristics to Fauvel et al.'s framework. The component of the extended framework characterizes the fairness of a machine learning model. The fairness context evaluation characteristics answer the question, *it is fair for whom?* It is important to identify the fairness context because while statistical parity-based group fairness equalizes outcomes across protected and non-protected groups, the outcome could still be very unfair from the point of view of an individual (Dwork et al. 2012). Additionally, according to Kearns et al. (2017), for group fairness, a classifier may appear to be fair on each individual group, but still can seriously violate the fairness constraint on one or more structured subgroups defined over the protected attributes, i.e., certain combinations of protected attribute values. We consider individual, group (Speicher et al. 2018), sub-group fairness (Kearns et al. 2017, 2019) and a combination of those for assessing the fairness context of the model for which it's fairness is measured.

| **Evaluation Characteristics** | **Question** | **Assessment Answer Values** |
|---|---|---|
| Fairness Context | It is fair for whom? Is it fair for individual or group/sub-group or both? | Individual, Group, Subgroup, Both (individual, group), All |
| Fairness | What is the level of fairness of the model? | Best, Similar, Below |

**Table 2. Extended characteristics for performance-explainability-fairness analytical framework**

The *fairness* evaluation characteristics answers *what is the level of fairness of the model?* Various fairness notions exist to evaluate the fairness of an ML model. Heidari et al. (2019) provide an interpretation of existing notions of algorithmic fairness for binary classification as special instances of Equality of Opportunity (EOP). Speicher et al. (2018) provide different fairness notions and their corresponding fairness conditions to evaluate the fairness of an ML model. However, there is no consensus on an evaluation procedure to assess the fairness of a machine learning model.

The choice of a metric to evaluate the fairness of a machine learning model depends on the application. According to the application, a metric aligned with the goal of the experiments is selected, which prevents the fairness comparison of machine learning models across applications. Therefore, the fairness component in the framework is defined as the first step towards a standard procedure to assess the fairness of machine learning models. It corresponds to the relative fairness of a model on a particular application. More specifically, it indicates the relative fairness of the models compared to the state-of-the-art model on a particular application and an evaluation setting. This definition allows the categorization of the models' fairness on an application and an evaluation setting. In the case of different applications with a similar machine learning task, the fairness component can give the list of models which outperformed current state-of-the-art models on their respective applications. Thus, it points to certain models that could be interesting to evaluate on a new application, without guaranteeing that these models would perform the same on this new application. Following Fauvel et al. (2020), we propose an assessment of the performance in three categories:

- **Best**: best fairness. It corresponds to the fairness of the first ranked model on the application following an evaluation setting (models, evaluation method, datasets)
- **Similar**: fairness similar to that of the state-of-the-art models. Based on the same evaluation setting, it corresponds to all the models which do not show a statistically significant fairness difference with the second ranked model.
- **Below**: fairness below that of the state-of-the-art models. It corresponds to the fairness of the remaining models with the same evaluation setting.

## Case Study Demonstration

For our purpose, we select the default payments prediction binary classification problem. We adopted an open dataset from the University of California Irvine (UCI) Machine Learning Repository named "default of credit card clients Data Set" (Yeh 2016) that records customers' credit card payment history. This data was used for research aimed at the case of customers' default payments in Taiwan to predict customers who are likely to default on their payments. This dataset consists of a total of 30,000 samples. There are 6,636 samples for the minority class with a binary label "Yes" that indicates the customers' defaults in the next month. Furthermore, there are 23,364 samples for the majority class for class with a binary label "No" that indicates the customers will not default in the next month. The twenty-three explanatory variables feature is separated into five static eighteen dynamic features, and the ID column is described in Table 3. Thirteen research papers (Table 4) have used the same default credit data set and published prediction accuracy results.

| **Attribute Name** | **Description** |
|---|---|
| ID | User ID |
| X1 (LIMIT_BAL) | Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. |
| X2 (SEX) | Gender (1 = male; 2 = female) |
| X3 (EDUCATION) | Education (1 = graduate school; 2 = university; 3 = high school; 4 = others) |
| X4 (MARRIAGE) | Marital status (1 = married; 2 = single; 3 = others) |
| X5 (AGE) | Age (year) |
| X6 – X11 (PAY_0 – PAY_6) | X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above |
| X12 – X17 (BILL_AMT1 – BILL_AMT6) | Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005 |
| X18 – X23 (PAY_AMT1 – PAY_AMT6) | Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005. |
| Y (default payment next month) | The binary variable Y is a response column, representing the default payment (1: Yes, 0: No). |

**Table 3. Default of credit card clients Data Set**

As described in the framework here we identify the overall fairness metrics for application specific to the “default of credit card client” dataset. These metrics help us determine fairness of the models while answering the question of what is the level of fairness of the model? Although the dataset has multiple demographic features, we would use only SEX as the demographic group among which we want to build models and measure metrics to assess fairness of models between the groups.

| Article | Algorithms | Result |
|---|---|---|
| Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach (Ebiaredoh-Mienye et al. 2021) | SSAE+LDA | A=90, P=91, S=90, F1=90 |
| Credit Card Default Prediction using Machine Learning Techniques (Sayjadah et al. 2018) | Random Forest (RF) | A=81, AUC=77 |
| Credit Default Mining Using Combined Machine Learning and Heuristic Approach (Islam et al. 2018) | Extremely Random Trees (ET) | A=96, P=96, R=86 90 |
| Credit Scoring : A Comparison between Random Forest Classifier and K- Nearest Neighbours for Credit Defaulters Prediction (Dewani et al. 2020) | Random Forest | A=95, P=94, R=79, F1=84 |
| Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer (Chishti and Awan 2019) | DNN | A=82, P=84, Sp=67 Se=96, R=96, F1=89 |
| Default Payment Analysis of Credit Card Clients (Sharma and Mehra n.d.) | Logistic Regression (LR) | A=79, P=55, R=77 F1=76, AUC=72 |
| Design and Comparison of Data Mining Techniques for Predicting Probability of Default on a Loan (Akcura and Chhibber n.d.) | Support Vector Machine (SVM) | A=81, P=75, R=77, F1=76 |
| Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction (Hsu et al. 2019) | Recurrent Neural Network with Random Forest (RNN-RF) | Lift index= 66, AUC=79, AAC=80 |
| Estimation of Credit Card Customers Payment Status by Using kNN and MLP (Koklu 2016) | Multilayer Perceptron (MLP) | A=81, MAE=32, RMSE=39 |
| Prediction of default payment of credit card clients using Data Mining Techniques (Subasi and Cankurt 2019) | Random Forest | A=89, F1=89, AUC=95 |
| Predictive Analysis of Credit Score for Credit Card Defaulters (Torvekar and Game 2019) | Random Forest | A=82 |
| Using Neural Network Techniques to Predict Possibility of Default Payment on Credit Card (Lin n.d.) | MLP + Genetic Algorithm (GA) | A=83 |
| Web service based credit card fraud detection by applying machine learning techniques (Prusti and Rath 2019) | Ensemble of SVM, K-Nearest neighbor (KNN), Decision Tree (DT) | A=83, P=97, Se=84, Sp=73, F1=90 |

**Table 4. List of papers with benchmarking results for algorithmic performance for default of credit card clients Data Set**

The following metrics are provided by the Microsoft’s Fairlearn toolkit (Bird et al. 2020) which is a toolkit for assessing and improving fairness in AI. The Fairlearn metrics package documentation (2021) provides more details on these metrics calculations.

- Demographic parity difference: Demographic parity is a fairness metric that is satisfied if a model's classification results are not dependent on a given sensitive attribute. In our case, it’s the feature SEX. Demographic parity is achieved if the percentage of males being defaulted is the same as that of females, irrespective of other characteristics between the two groups. The lower it is, the better demographic parity between groups.
- Demographic parity ratio: The demographic parity ratio is defined as the ratio between the smallest and the largest group-level selection rate across all values of the sensitive feature(s). The demographic parity ratio of 1 means that all groups have the same selection rate.
- Equalized odds difference: A fairness metric that checks if, for any attribute, a classifier predicts that label equally well for all values of that attribute. Equalized odds are satisfied provided that no matter whether an individual is a male or a female if they are qualified for defaults, they are equally as likely to get a default decision. If they are not qualified, they are equally as likely to get a non-default decision. The lower it is, the better equalized odds parity between groups. The equalized odds difference of 0 means that all groups have the same true positive, true negative, false positive, and false negative rates.

Along with the fairness metrics, we identify the performance metrics that are used for this case study. These metrics are also available through the Fairlearn toolkit python library:

- Overall balanced error rate: The Balanced Error Rate (BER) is the average of the errors on each class.
- Balanced error rate difference: This is the difference of BER between the groups. The lower it is the better, close to zero.
- Overall AUC: AUC (Area Under the Curve) where the curve is the ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking a classification model's performance. The closer the score toward 1, is the better.
- AUC difference: This is the difference of AUC score between the groups. The lower it is, the better.

Inspired by the results from the literature review (Table 4), where tree-based classifiers showed better model performance, we conduct an experiment using four tree bases classifiers namely, 1) Decision Tree (DT), 2) Gradient Boosting Machines (GBM), 3) Extremely random Trees (ET) and 4) Random Forest (RF). The following Table 5 shows the result from the experiment where we present the fairness and performance metrics for each of the models. For the case study, we used the Microsoft Fairlearn (Bird et al. 2020) tool and its Python library to measure the fairness of the mentioned machine learning classifiers. First, we removed the SEX feature from the dataset, then we split the dataset into 70% train and 30% for test. Then we used a GridSearch to mitigate disparities. The predictors produced by GridSearch do not access the sensitive feature at test time. Also, rather than training a single model, we train multiple models corresponding to different trade-off points between the performance metric (balanced accuracy) and fairness metric (equalized odds difference). We do this for all four classifiers mentioned above and present the result in Table 5.

| | **Metrics** | **Classifiers** | | | |
|---|---|---|---|---|---|
| | | **DT** | **GBM** | **ET** | **RF** |
| **Fairness** | Demographic parity difference | 0.013346 | 0.027488 | 0.042925 | 0.040284 |
| | Demographic parity ratio | 0.969635 | 0.909465 | 0.879934 | 0.881325 |
| | Equalized odds difference | 0.006312 | 0.01592 | 0.03179 | 0.029033 |
| **Performance** | Overall balanced error rate | 0.372879 | 0.285551 | 0.297963 | 0.297079 |
| | Balanced error rate difference | 0.004229 | 0.007767 | 0.006734 | 0.006627 |
| | Overall AUC | 0.627185 | 0.778476 | 0.756681 | 0.770468 |
| | AUC difference | 0.004331 | 0.007176 | 0.012103 | 0.017096 |

**Table 5. Four different classifiers fairness and performance metrics**

The experimental result shows that the GBM classifier has the highest overall AUC value measured for performance. It also has a negligible AUC difference between the different SEX, so we consider it the best performing model. For fairness, DT has the highest demographic parity ratio and the lowest equalized odds difference making it the fairest model. The GBM model is very close to these metrics to the DT model with a far higher AUC value. However, the GBM model has a negligible difference in the fairness metrics values than the DT model.

Now we apply the results in the extended performance-explainability-fairness framework. Table 6 summarizes the extended framework results of the four default credit classifiers.

| **Evaluation Characteristics** | **DT** | **GBM (with SHAP)** | **ET (with SHAP)** | **RF (with SHAP)** |
|---|---|---|---|---|
| Performance | Below | Best | Below | Similar |
| Comprehensibility | White-box | Black-box | Black-box | Black-box |
| Granularity | Local | Global & Local | Global & Local | Global & Local |
| Information type | Feature importance | Feature importance | Feature importance | Feature importance |
| Faithfulness | Perfect | Imperfect | Imperfect | Imperfect |
| User category | Domain Expert | Domain Expert | Domain Expert | Domain Expert |
| Fairness Context | Group | Group | Group | Group |
| Fairness | Best | Similar | Below | Below |

**Table 6. Summary of extended framework results of the four default credit classifiers**

The GBM has the best performance based on overall AUC. RF has an AUC score close to GBM, so we consider it similar performance, whereas the DT and ET have below performance. For comprehensibility, only the DT model is white-box, the other three being an ensemble model are all block-box models. For granularity, while DT can provide local level, all other classifiers with SHapley Additive exPlanations SHAP (Lundberg and Lee 2017) can provide both global and local level granularity of explanation. All the classifier's explanation provides feature importance information. Since an explanation can be extracted directly from the DT original model, it is a perfect faithful model. In contrast, all other models requiring a post-hoc SHAP explanation method using surrogate models would be considered as having imperfect faithfulness. The fairness context for all the classifiers is group fairness and DT is the fairest model. However, GBM comes close, and its fairness can be considered as close to DT. The other two classifiers provide a level of fairness that is considered below than DT and GBM.

An illustration of the result is also presented in Figure 1. Overall, we determine that the GBM model would be the best choice when considering all different characteristics of the performance-explainability-fairness framework as applicable to the default credit score dataset.

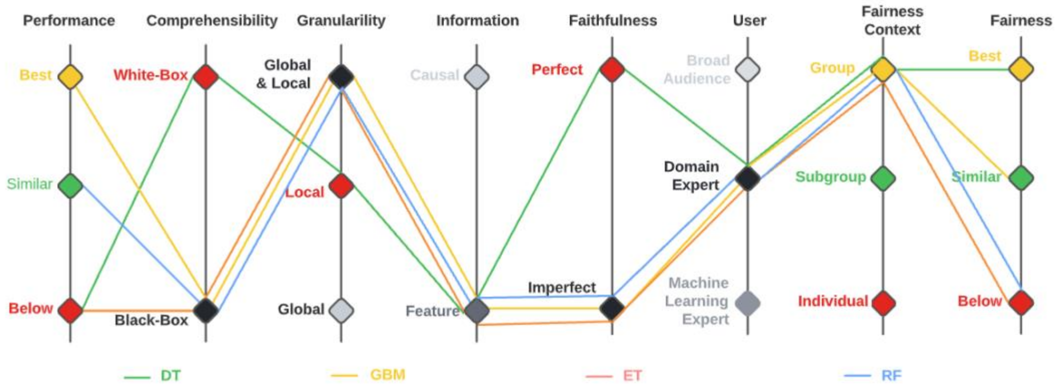

**Figure 1. Parallel coordinates plot of the default credit classifiers (extended framework)**

## Discussion

In our benchmarking classifiers, the first DT classifier is a pure decision tree classifier, and while it provided the lower performance, it was the best in terms of fairness. DT classifier also having white-box comprehensibility has perfect faithfulness. However, since its performance on accuracy is lower, it might not be the best choice for decision-making purpose if higher performance levels are required. Given the nature of the decision problem, we would like to maximize both the performance and the fairness while being more flexible on the explainability if we have the choice. However, if perfect faithfulness is required and performance expectations are relaxed, DT would be the only choice here. Similarly, if performance and fairness receive similar priorities, GBM gives us the best choice. This tells us that when multiple model metrics are available, we must look at the use case on hand and find a model that closely fits all metrics requirements in a priority-adjusted manner.

It is evident through the result presented in Table 6 that the proposed performance-explainability-fairness framework can be used to assess and benchmark the default of the credit card classifiers. Therefore, we claim that the framework can benchmark multiple classifiers for a binary decision classification problem. Further, we should also note that the framework is flexible enough that new fairness metrics can be implemented and benchmarked using several classifiers for a specific decision problem.

## Conclusion

We have presented a new performance-explainability-fairness analytical framework to assess and benchmark ML models. The proposed framework details a set of characteristics that systematize the performance-explainability-fairness assessment of ML models. The experimental case study demonstrates the successful application of the performance-explainability-fairness framework to a classification problem. The proposed framework can be used to identify ways to improve current machine learning models and design new ones. Future research may explore additional fairness metrics pertinent to various applications, enhance the process for applying the proposed framework, and investigate its applicability to other machine learning models.

## REFERENCES

Akcura, K., and Chhibber, A. (n.d.). *Design and Comparison of Data Mining Techniques for Predicting Probability of Default on a Loan*, p. 8.

Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115. (https://doi.org/10.1016/j.inffus.2019.12.012).

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. 2020. "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI," No. MSR-TR-2020-32, Microsoft, May. (https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/).

Castelnovo, A., Crupi, R., Greco, G., Regoli, I., and Cosentini, A. 2022. *A Clarification of the Nuances in the Fairness Metrics Landscape | Scientific Reports*, (4209). (https://doi.org/10.1038/s41598-022-07939-1).

Chakrobartty, S., and El-Gayar, O. 2021. "Explainable Artificial Intelligence in the Medical Domain: A Systematic Review," *AMCIS 2021 Proceedings. 1*, p. 11.

Chishti, W. A., and Awan, S. M. 2019. "Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer," in *2019 International Conference on Innovative Computing (ICIC)*, Lahore, Pakistan: IEEE, November, pp. 1–8. (https://doi.org/10.1109/ICIC48496.2019.8966723).

Dewani, P., Sippy, M., Punjabi, G., and Hatekar, A. 2020. *Credit Scoring : A Comparison between Random Forest Classifier and K- Nearest Neighbours for Credit Defaulters Prediction*, (07:10), p. 7.

Du, M., Yang, F., Zou, N., and Hu, X. 2020. "Fairness in Deep Learning: A Computational Perspective," *IEEE Intelligent Systems*, pp. 1–1. (https://doi.org/10.1109/MIS.2020.3000681).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. "Fairness through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, Cambridge, Massachusetts: ACM Press, pp. 214–226. (https://doi.org/10.1145/2090236.2090255).

Ebiaredoh-Mienye, S. A., Esenogho, E., and Swart, T. G. 2021. "Artificial Neural Network Technique for Improving Prediction of Credit Card Default: A Stacked Sparse Autoencoder Approach," *International Journal of Electrical and Computer Engineering (IJECE)* (11:5), p. 4392. (https://doi.org/10.11591/ijece.v11i5.pp4392-4402).

"Fairlearn.Metrics Package — Fairlearn 0.7.0 Documentation." 2021. (https://fairlearn.org/v0.7.0/api_reference/fairlearn.metrics.html, accessed February 28, 2022).

Fauvel, K., Masson, V., and Fromont, É. 2020. "A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers," *ArXiv:2005.14501 [Cs, Stat]*. (http://arxiv.org/abs/2005.14501).

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.* (51:5), New York, NY, USA: Association for Computing Machinery. (https://doi.org/10.1145/3236009).

Gunning, D., and Aha, D. 2019. "DARPA's Explainable Artificial Intelligence Program," *AI Magazine* (40:2), La Canada, pp. 44–58. (https://doi.org/10.1609/aimag.v40i2.2850).

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., and Preece, A. 2019. *A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems*, p. 2.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. 2019. "A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, January 29, pp. 181–190. (https://doi.org/10.1145/3287560.3287584).

Ho, A. 2019. "Deep Ethical Learning: Taking the Interplay of Human and Artificial Intelligence Seriously," *Hastings Center Report* (49:1), pp. 36–39. (https://doi.org/10.1002/hast.977).

Holzinger, A. 2018. "From Machine Learning to Explainable AI," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, , August, pp. 55–66. (https://doi.org/10.1109/DISA.2018.8490530).

Holzinger, A., Plass, M., Holzinger, K., Crisan, G. C., Pintea, C.-M., and Palade, V. 2017. "A Glass-Box Interactive Machine Learning Approach for Solving NP-Hard Problems with the Human-in-the-Loop," *CoRR* (abs/1708.01104). (http://arxiv.org/abs/1708.01104).

Hsu, T.-C., Liou, S.-T., Wang, Y.-P., Huang, Y.-S., and Che-Lin. 2019. "Enhanced Recurrent Neural Network for Combining Static and Dynamic Features for Credit Card Default Prediction," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, May, pp. 1572–1576. (https://doi.org/10.1109/ICASSP.2019.8682212).

Islam, S. R., Eberle, W., and Ghafoor, S. K. 2018. "Credit Default Mining Using Combined Machine Learning and Heuristic Approach," *ICDATA*, p. 7.

Kanda, E., Epureanu, B. I., Adachi, T., Tsuruta, Y., Kikuchi, K., Kashihara, N., Abe, M., Masakane, I., and Nitta, K. 2020. "Application of Explainable Ensemble Artificial Intelligence Model to Categorization of Hemodialysis-Patient and Treatment Using Nationwide-Real-World Data in Japan.," *PloS One* (15:5), United States, p. e0233491.

Kearns, M., Neel, S., Roth, A., and Wu, Z. 2017. *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. 2019. "An Empirical Study of Rich Subgroup Fairness for Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, January 29, pp. 100–109. (https://doi.org/10.1145/3287560.3287592).

Koklu, M. 2016. "Estimation of Credit Card Customers Payment Status by Using KNN and MLP," *International Journal of Intelligent Systems and Applications in Engineering* (4:Special Issue-1), pp. 249–251. (https://doi.org/10.18201/ijisae.2016SpecialIssue-146983).

Lin, Y. (n.d.). *Using Neural Network Techniques to Predict Possibility of Default Payment on Credit Card*, p. 6.

London, A. J. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *Hastings Center Report* (49:1), pp. 15–21. (https://doi.org/10.1002/hast.973).

Lundberg, S., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," *CoRR* (abs/1705.07874). (http://arxiv.org/abs/1705.07874).

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., and Lee, S.-I. 2018. "Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery.," *Nature Biomedical Engineering* (2:10), pp. 749–760. (https://doi.org/10.1038/s41551-018-0304-0).

Makhlouf, K., Zhioua, S., and Palamidessi, C. 2021. "On the Applicability of Machine Learning Fairness Notions," *ACM SIGKDD Explorations Newsletter* (23:1), pp. 14–23. (https://doi.org/10.1145/3468507.3468511).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys* (54:6), pp. 1–35. (https://doi.org/10.1145/3457607).

Naylor, M., French, C., Terker, S., and Kamath, U. 2021. "Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff," *ArXiv:2107.05693 [Cs]*. (http://arxiv.org/abs/2107.05693).

Prusti, D., and Rath, S. K. 2019. "Web Service Based Credit Card Fraud Detection by Applying Machine Learning Techniques," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, Kochi, India: IEEE, October, pp. 492–497. (https://doi.org/10.1109/TENCON.2019.8929372).

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., and Kasmiran, K. A. 2018. "Credit Card Default Prediction Using Machine Learning Techniques," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Subang Jaya, Malaysia: IEEE, October, pp. 1–4. (https://doi.org/10.1109/ICACCAF.2018.8776802).

Sharma, S., and Mehra, V. (n.d.). *Default Payment Analysis of Credit Card Clients*, p. 7.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. 2018. "A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom: ACM, July 19, pp. 2239–2248. (https://doi.org/10.1145/3219819.3220046).

Subasi, A., and Cankurt, S. 2019. "Prediction of Default Payment of Credit Card Clients Using Data Mining Techniques," in *2019 International Engineering Conference (IEC)*, Erbil, Iraq: IEEE, June, pp. 115–120. (https://doi.org/10.1109/IEC47844.2019.8950597).

Torvekar, N., and Game, P. S. 2019. *Predictive Analysis of Credit Score for Credit Card Defaulters*, (7:5), p. 5.

Verma, S., and Rubin, J. 2018. "Fairness Definitions Explained," in *Proceedings of the International Workshop on Software Fairness*, Gothenburg Sweden: ACM, May 29, pp. 1–7. (https://doi.org/10.1145/3194770.3194776).

Yeh, I.-C. 2016. "UCI Machine Learning Repository: Default of Credit Card Clients Data Set," , January 26. (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients, accessed January 23, 2022).