Explainable AI for Non-Experts: Energy Tariff Forecasting

Hongnan Ma, Kevin McAreavey, Ryan McConville, Weiru Liu
School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths
University of Bristol, Bristol, UK
{ex20249, kevin.mcareavey, ryan.mcconville, weiru.liu}@bristol.ac.uk

***Abstract*—Non-expert users are increasingly affected by the decisions of systems that rely on machine learning (ML), yet it is often difficult for these users to understand the predictions of ML models. In this paper, we propose a web-based platform to evaluate explainable AI (XAI) for non-experts in the context of time series forecasting, focusing on energy price predictions as an exemplary use case. The XAI methods we consider include local feature importance and counterfactual explanations. The platform relies on gamification to encourage user engagement. Our research objective is to evaluate the effectiveness of these different approaches from the perspective of non-expert understanding of machine learning models.**

***Index Terms*—XAI, global, local, counterfactual explanation**

I. Introduction

Machine learning (ML) and Artificial Intelligence (AI) has demonstrated effectiveness across a range of domains such as finance [1], medicine [2], social media [3] and autonomous driving [4], [5]. However, for those who interact with AI, the lack of transparency and ability to understand their actions may affect their trust in such systems. This is especially noticeable when the system is making high-stake decisions [6] where human understanding of those decisions is vital. In the medical industry [2], for example, diagnosing a condition early can be critical. In the banking sector, AI-enabled banking systems may use credit scores to decide if a loan request is granted, and suitable explanations are expected if a loan request is denied. Explanations are required not just by researchers, but may also be required by regulatory obligations that are becoming more prevalent, such as the *General Data Protection Regulation* (GDPR) [7] which requires that the underlying principles for algorithmic predictions be transparent, and that the rationale behind those predictions be explainable, if necessary.

One common categorisation of machine learning models, within this context, is white box and black box models, with the former models decision-making considered transparent, and the latter less so. Further, XAI approaches can be characterized as producing *global* or *local explanations*, or as being *post-hoc* or *ante-hoc* [8]. That is, global explanations provide a higher level insight into a model while local explanations provide individual instance-specific explanations. Post-hoc methods provide explanations for an already trained model whilst ante-hoc methods produce an additional component (to the trained model) to aid explanations.

This work is partially funded by the EPSRC CHAI project (EP/T026820/1).

In [9], the authors divided users of an AI system into three classes based on their expertise: *AI beginners, domain experts*, and *AI experts*. This is based on the belief that distinct groups have different requirements. While AI experts may find utility in global explanations of how machine learning models function, beginners and (non-AI) domain experts may benefit more from local explanations about the relationship between specific inputs and outputs. Although explanation has been found to increase understanding of ML systems for a wide range of audiences [10] [11], non-experts are chosen as an important but under-represented class of stakeholder within XAI research [12].

Gamification is the incorporation of game elements into systems or activities for the purpose of motivating and engaging users [13]. Gameplay data from users can be logged and analyzed, with the aim of to improve the underlying system. For example, [27] utilised a type of gamification called games with a purpose (GWAP) [14] to evaluate XAI at scale. The authors concentrated on explaining deep learning models that have been built for image recognition, and claim it is the first time that GWAP has been used within XAI research. In [15], the authors replaced the game design with user objectives in XAI planning, and added money as an incentive. They connect the user's objective (or goal) to payment via a bonus that grows in proportion to the goal value attained, which gives an incentive to develop a successful strategy.

In this paper, we design and develop an interactive XAI system for non-experts. As an application domain we use an ML model that performs energy price prediction. We design two interactive user interfaces. One interface is intended for users in a *control group*, whilst another interface is intended for users in a *treatment group*. Our proposed gamification uses a simple game to allow users to experience how the value of different features influences predictions of the underlying ML model. A simple *score* mechanism is used to inform a user how their own predictions differ from the system's predictions so that, with repeated play, a user can gain a better understanding of the underlying ML model.

The rest of the paper is as follows: Section 2 introduces the dataset and gamification. Section 3 discusses global, local and counterfactual explanations. Section 4 describes the system design. Finally, we provide a brief summary and our plans for the future in the conclusion.

II. Preliminaries

A. Dataset Description

The dataset used in this study is Octopus Energy import prices[1] for the London region during a two-year period beginning midnight on 1 January 2018 and ending at midnight on 31 December 2019. The original dataset is a time series dataset containing 34,993 instances at 30 minute intervals, with the current electricity price with and without tax[2]. In order to use standard regression algorithms we pre-process the dataset to extract the following set of time-based features: *year, month, day, hour, day of week, is weekend*. In addition to those, *weather* and *carbon intensity* features are also added to the dataset to improve the predictive performance of the model. Carbon intensity data comes from carbon intensity forecast in National Grid ESO[3], which is described in terms of a 96+ hour forecast of CO2 emissions per kWh of consumed electricity. Meteomatics[4] provides temperatures in degrees Celsius for the London area and is updated every half-hour.

To select a regression model we first evaluate several common regression models to identify a performant model for our XAI experiment. We emphasise that we are not interested in identifying a state-of-the-art model for this task, but that any reasonably well-performing model will suffice. We trained several regression models, using the final month for evaluation. Mean squared error (MSE) is the evaluation metric used which measures the mean squared difference between the predictions and the ground truth. Table I provides the performance of the top three regression models we tested, in terms of MSE, with the **Random Forest** model achieving the best performance. Furthermore, the performance from these models improved when carbon intensity and temperature were added as extra features, with the MSE for the best performing model reducing from 6.33 to 3.31.

TABLE I
Top three performing regression models.

Regression Model	MSE
Random Forest	3.31
Extra Trees	3.56
Histogram-based Gradient Boosting Regression Tree	3.75

B. Gamification

The inclusion of a gamification component in our XAI system provides a metric for us to evaluate user performance. In our study, we seek to measure the understanding that a non-expert has of an underlying ML model. To aid this, we design a game where the user receives a score based on their understanding. This ***Score*** will be received during gameplay. Despite the fact that there are presently a variety of explanatory approaches for black box models, such as LIME and SHAP, quantitative assessment methodologies for XAI are presently lacking. There is currently no agreement on how best to assess explanations [16]. In one example, [16] evaluated explanations in terms of application-ground, human-ground and function-ground. However, no precise objective evaluation techniques are provided. Several researchers have validated the efficacy of explanations by conducting user studies. In [17], over 200 participants were asked to evaluate an XAI interface. The study deployed an assessment measure called the Explanation Satisfaction Scale to measure satisfaction after providing explanation. However, this assessment approach is static and not based on incentives, while the subjective nature of user-reported satisfaction makes it difficult to judge the real usefulness of explanations. Our gamification and scoring mechanism aims to overcome this limitation.

Our game interface asks the user to estimate the relationship between a model's predicted price X and a hypothetical estimated price Y, given a specific input instance, with a limited number of qualitative options for the user to choose from. Those options are *much less than, less than, similar to, greater than, much greater than*. The reason we ask for a qualitative estimation is to place less cognitive load on the (non-expert) user. Each option is associated with a qualitative numeric value from $\{-2, -1, \ldots, 2\}$ as shown in Table II. The absolute difference between the qualitative numeric value of the model's prediction and the user's choice, normalised by the maximum absolute difference (i.e. $|2 - (-2)| = 4$), is then taken as the user's score. Thus, scores are taken from $\{0, 1, \ldots, 4\}$ with 4 being the best score and 0 being the worst. For example, if the user estimates that X is much greater than Y, and X is actually less than Y, then the user will be awarded the score of $4 - |2 - (-1)| = 1$. Scores thus provide a way to measure user understanding of the underlying ML model, and follow the common-sense understanding in games that a higher score is better.

TABLE II
An overview of the scoring system.

Option	Semantics	Numeric value
Much less than	X <Y-20	-2
Less than	Y-20 <= X <Y-5	-1
Similar to	Y-5 <= X <= Y+5	0
Greater than	Y+5 <X <= Y+20	1
Much greater than	Y+20 <X	2

III. Explanations for black box models

Broadly speaking, machine learning models can be divided into two categories related to their interpretability, namely, black box models and white box models. Black box models models are difficult to understand on their own [18], and as a result, their is a greater need for explanation. On the other hand, white box models [18] are regarded as interpretable by design, and hence are easier to explain.

According to the results of our experiments, the Random Forest model, which may be considered a black box model,

[1]https://api.octopus.energy/v1/products/AGILE-18-02-21/electricity-tariffs/E-1R-AGILE-18-02-21-C/standard-unit-rates/

[2]The original column is:value_exc_vat, value_inc_vat, valid_from, valid_to

[3]https://carbonintensity.org.uk/

[4]https://www.meteomatics.com/en/

performs the best at our task. A random forest model is a collection of decision trees, each trained on a different subset of the data with random subsets of the features. As a result, viewing each tree (of the typically large number) is not a viable explanation approach for non-experts.

Some explanation methods that do not consider the specifics of the machine learning model model, but only the inputs and outputs, are collectively referred to as ***model-agnostic explanations*** [19]. Model-agnostic explanation can be separated into global and local approaches, depending on the aim of interpretation.

A. Global Explanation

To interpret the model's global output, the model must be trained to understand the algorithm and the data. This level of interpretability refers to the model's decision-making process in relation to the full feature space and model structure. The most popular global approach is Permutation Feature Importance (PFI) [20] The PFI is used to determine which feature has the most influence on the prediction. The PFI selection approach evaluates a model's performance after eliminating each unique feature and replacing it with random noise. Individual feature importance may thus be directly compared, and a quantitative threshold can be utilised to determine feature inclusion. In Fig. 1, the rows towards the top are the most important features, and those towards the bottom matter least. Thus, *hour* is regarded as the most important feature, whereas *day* is seen as the least important.

Weight	Feature
1.6310 ± 0.0432	hour
0.0679 ± 0.0035	CO2
0.0036 ± 0.0014	temperature
0.0005 ± 0.0005	dayofweek_num
0.0003 ± 0.0003	IsWeekend
0 ± 0.0000	month
0 ± 0.0000	year
-0.0011 ± 0.0010	day

Fig. 1. Global explanation for the energy price dataset

B. Local Explanation

[21] points out that local explanation offers a tailored explanation that is focused on the particulars of each instance. It provides a thorough explanation of how a machine learning model may provide precise predictions about the features effecting a specific prediction. The authors studied the global and local interpretability of machine learning in Type 2 diabetes screening, and found that characteristics such as age, gender and body mass index (BMI) contributed significantly to global explanation. However, for a specific patient they found that depression, smoking status or physical health had a significant impact on the development of Type 2 diabetes in that patient.

In our work, we use SHapley Additive exPlanations (SHAP) [22] to generate the local explanation. SHAP can find the feature importance which can interpret the predictions of any machine learning classifier or regressor. SHAP includes two sub-methods: *KernelSHAP* and *TreeSHAP*. We utilise TreeSHAP to explain our Random Tree regressor.

To illustrate, we use the SHAP method to provide a local explanation for one instance of our dataset. Table III depicts the feature values of an instance that we want to investigate.

TABLE III
AN INSTANCE FROM OUR DATASET.

Temperature	**Carbon intensity**	**Year**
2.4°C	229 gCO2/kWh	2019
Month	**Day**	**Hour**
December	1	0 am
Day of week	**Isweekend**	**-**
Saturday	Yes	-

According to Fig. 2, we can not only see the influence trend of each feature, but we can also see how the features contribute to a single prediction. Shap values deconstruct a prediction to demonstrate the impact of each feature. The resulting price is £7.52, while the base value is £12.24. The feature values that cause greater predicted values are highlighted in pink, and their visual size indicates the magnitude of the feature's effect. Blue represents feature values that decrease the predicted value.

Fig. 2. Local Explanation

In this scenario, the local explanation concludes that features "temperature" and "carbon intensity" have a positive influence on the prediction, while "hour", "year", and "day" have a negative influence. Based on this plot, we could generate the following textual local explanation in a pop-up window: *"temperature" and "carbon intensity" are the two most inferential features for energy price in this case".*

C. Counterfactual Explanation

[23] discussed that the end-user may not be particularly interested in why a certain prediction was obtained and which features of the input led to the prediction. Instead, they may be more interested in understanding the changes that can be made to obtain other predictions. Counterfactual explanations primarily address the issue of how the prediction will change if a certain change in the features of the input occurs. It compares the user's expectations with the actual predicted outcomes, and provides suggestions about how to change feature values in order to alter prediction results. In some cases, counterfactual explanation is more intuitive and useful within the local explanation category. Furthermore, some psychologists [24] [25] have demonstrated that counterfactuals elicit causal reasoning in humans.

Counterfactual explanations are commonly selected based on some measure of proximity to the original input. By restricting which features to focus on (and what range of

values they can take), specific counterfactual instances can be found to better meet a user's expectation.

We can use the same instance in Table III as an example to generate explanations for using DiCE [26], which generates counterfactual explanations of machine learning models. In Fig. 3, the top row represents the feature values of the explainee datapoint and the bottom rows represent two counterfactual datapoints. In these two rows, features with "-" mean that these features have the same values as the explainee datapoint. Accordingly, the features with values present are those which are different from the explainee datapoint's features.

	degree	CO2	year	month	day	Hour	dayofweek_num	IsWeekend	price
0	2.4	229.0	2019.0	12.0	1.0	0.0	6.0	1.0	7.524851

Diverse Counterfactual set (new outcome: [10, 11])

	degree	CO2	year	month	day	Hour	dayofweek_num	IsWeekend	price
0	-3.4	48.0	-	1.0	-	-	5.0	-	10.012791374439013
0	-5.3	48.0	-	1.0	-	-	4.0	-	10.012791374439013

Fig. 3. Counterfactual Explanation

We implemented and integrated DiCE as part of the interface and Fig. 3 is part of the interface, where a pop-up arrow will point to the feature values that have influenced the ML to produce the alternative outcomes.

IV. System Design

A. System Architecture

Fig. 4 illustrates how the architecture of our proposed system. The system involves three main components. They are *front-end*, *back-end* and the *dataset*. The *front-end* is end-user facing, which receives user input to be transmitted to the *back-end* via an API, which then performs tasks such as prediction and explanation generation. The dataset component is used purely to train the underlying ML model which the user is seeking to understand.

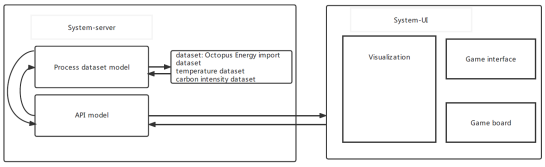
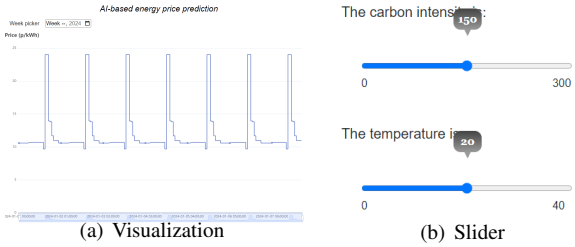


Fig. 4. System architecture

B. Interface Design

The interface has three panels. They are the *visualisation, game interface* and *game board panel*. Fig. 5(a) depicts the *visualisation panel*. The visualisation is presented to the user with a stepped line chart. In our system it displays the energy tariff price in the future (e.g., one year from now), with a point at every half-hour interval. Additionally, for exploration, the top of this panel displays a week calendar picker to allow users to change calendars week by week. Fig. 5(b) illustrates two sliders for the *carbon intensity* and *temperature* feature values. The default carbon intensity is set to 150 gCO2/kWh, and the default temperature is set to 20 °C. The scale of the carbon intensity is from 0 gCO2/kWh to 300 gCO2/kWh and the scale of temperature is from 0°C to 40°C.

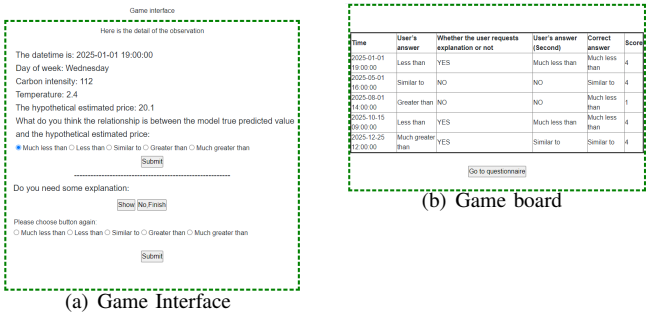


(a) Visualization (b) Slider

Fig. 5. Visualisation and feature slider controls.

The *game interface panel* is depicted in Fig. 6(a). The purpose of this panel is for users to explore the system interactively. Fig. 6(b) shows the *game board* displaying the results of each round of the game. There are five rounds for each player to play. The game board is divided into six columns for treatment group participants. They are as follows:

- *day/time*: it shows the date and time of each round.
- *User's answer*: it show the result inferred from the visualisation screen.
- *Whether the user requests explanation or not*: it shows whether an explanation is required (i.e, yes or no)
- *User's answer (second)*: it shows the result inferred after viewing the textual explanation.
- *Correct answer*: it shows the true answer.
- *Score*: it shows the score received for each round.



(a) Game Interface (b) Game board

Fig. 6. Game interface and game board panel

C. Interaction Design and Gamification for Local Explanation

The treatment group user interface uses gamification for local explanation. That is, each round of games is about one specific datapoint where data feature values are given, and through gamification, a user is tasked with the predicting the prediction of the underlying ML model for that datapoint. Although gamification used in this paper does not provide counterfactual examples, through the process of playing games, a user can obtain comparisons between different instances. The user can analyse the correlation between feature input values and their predictions and could learn from each round of play.

We provide a case study of the gamification interface for the treatment group. Fig. 7 demonstrates the whole interface for treatment group participants.

Step 1: The game interface is populated by the feature values of the datapoint whose price is to be predicted.

- *datetime*: 07:00–07:30 on Wednesday 1 January 2025 (this should occur in the future)
- *carbon intensity*: 112 gCO2/kWh.
- *temperature*: 2.4°C.

Step 2: The task is for the user to utilise explanations in order to determine if the a hypothetical price is close to what the underlying model would predict given the features. For this reason, each round of the game includes a hypothetical price. To utilise the visualisation, the user can modify the carbon and temperature features via the sliders (e.g., move temperate slider to 2.4°C, and the carbon intensity slider to 112 gCO2/kWh). They can then inspect the chart for time-related trends by changing e.g. the month or year. For example, users may inspect:

- Prices at 07:00am-07:30am every day.
- Prices every Wednesday.
- Prices on the first day of every month.
- Prices in May every year.

After completing all operations and analysis, the user triggers the button to refresh the visualization with the matching parameters

Step 3: The system asks the following question: *What do you think the relationship is between the model true predicted value and the hypothetical estimated price?* A user needs to select a *Radio button* (corresponding to the qualitative options introduced previously) as an answer based on their analysis of the visualization. The system will then display a message to tell the user whether their choice is correct.

Step 4: Users are asked further questions *Do you require some explanation?*. User can select either *Show* or *No, Finish*. If the user selects *Show*, a local explanation will be displayed. If the user selects *No, Finish*, the game will move on to the next round. After selecting *Show*, the user is prompted to select again by the sentence *"Please choose again:"*. Once done, the user can *Submit*. The game repeated in this way 5 times.

Step 5: All recorded information is shown in the *game board* with *Score* being calculated by the system.

In the game board panel, two critical pieces of information will be recorded by the system. Firstly, it is important for users in the treatment group to determine whether the user requests the textual explanation or not. This information may indicate whether the user felt confident in their understanding of the visualisation. Secondly, the users score is important as a metric of their understanding and will aid in further analysis of user performance.

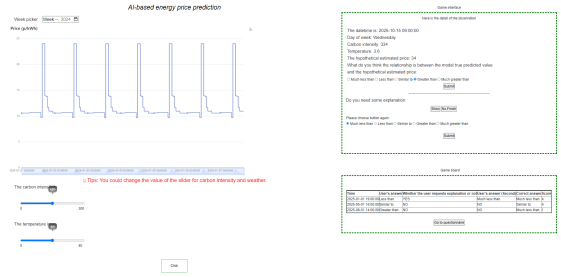


Fig. 7. The complete XAI user interface.

V. Conclusion

In this paper, we proposed a web-based XAI system consisting of visual and textual components using gamification to measure non-expert user understanding of a ML prediction model. For illustrative purposes, we used home energy price prediction as the application. The proposed system is flexible in that it supports not only various explanation modalities, but also different types of explanations, i.e., local and counterfactual explanations. In future work, we will conduct real-world user experiments using the proposed system.

References

[1] Z Shi, JP Cartlidge. "State Dependent Parallel Neural Hawkes Process for Limit Order Book Event Stream Prediction and Simulation." 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery (ACM), 2022.

[2] A. Holzinger. "Explainable AI and multi-modal causability in medicine," i-com, 2020, 19(3): 171-179.

[3] H Bo, R McConville, J Hong,W Liu. "Social Influence Prediction with Train and Test Time Augmentation for Graph Neural Networks." 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021.

[4] J.M Schraagen, P. Elsasser, H. Fricke, et al. "Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models," 2020, 64(1): 339-343.

[5] Y Li, H Wang, L.M Dang, et al. "A deep learning-based hybrid framework for object detection and recognition in autonomous driving," IEEE Access, 2020, 8: 194228-194239.

[6] W Samek, et al. "Explaining deep neural networks and beyond: A review of methods and applications." Proceedings of the IEEE 109.3 (2021): 247-278.

[7] P Voigt, A Von dem Bussche. "The eu general data protection regulation (gdpr),". A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017, 10(3152676): 10-5555.

[8] A Holzinger, R Goebel, R Fong, et al. "xxAI-Beyond Explainable Artificial Intelligence." International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. Springer, Cham, 2022.

[9] S Mohseni, N Zarei, E.D Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," ACM Transactions on Interactive Intelligent Systems (TiiS), 2021, 11(3-4): 1-45.

[10] J Krause, A Perer, K Ng. "Interacting with predictions: Visual inspection of black-box machine learning models." Procs. of the 2016 CHI conference on human factors in computing systems. 2016.

[11] Q.V Liao, D Gruen, S Miller. "Questioning the AI: informing design practices for explainable AI user experiences," 2020: 1-15.

[12] K McAreavey, K Bauters, W Liu. "A Smart Home Testbed for Evaluating XAI with Non-experts." ICAART (3). 2022.

[13] K.M. Kapp. "The gamification of learning and instruction: game-based methods and strategies for training and education." John Wiley and Sons, 2012.

[14] L Von Ahn, L Dabbish. "Designing games with a purpose." Communications of the ACM 51.8 (2008): 58-67.

[15] R Eifler, M Brandao, A.J Coles, et al. “Plan-Property Dependencies are Useful: A User Study.” ICAPS 2021 Workshop on Explainable AI Planning. 2021.
[16] F Doshi-Velez, B Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint arXiv:1702.08608 (2017).
[17] Y Zhang , K McAreavey, W Liu. “Developing and Experimenting on Approaches to Explainability in AI Systems.” ICAART (2). 2022.
[18] S Nidhra, J Dondeti. “Black box and white box testing techniques-a literature review.” International Journal of Embedded Systems and Applications (IJESA) 2.2 (2012): 29-50.
[19] M.T. Ribeiro, S Singh, C Guestrin. “Model-agnostic interpretability of machine learning.” arXiv preprint arXiv:1606.05386 (2016).
[20] L Breiman.“Random forests.” Machine learning 45.1 (2001): 5-32.
[21] L Kopitar, L Cilar, P Kocbek, et al. “Local vs. global interpretability of machine learning models in type 2 diabetes mellitus screening.” Artificial intelligence in medicine: Knowledge representation and transparent and explainable systems. Springer, Cham, 2019. 108-119.
[22] S.M.Lundberg, and S.I Lee. “A unified approach to interpreting model predictions.” Advances in neural information processing systems 30 (2017).
[23] S Wachter, B Mittelstadt, C Russell. “Counterfactual explanations without opening the black box: Automated decisions and the GDPR.” Harv. JL & Tech. 31 (2017): 841.
[24] J Woodward. “Psychological studies of causal and counterfactual reasoning.” Understanding counterfactuals, understanding causation. Issues in philosophy and psychology (2011): 16-53.
[25] D.R. Mandel, D.J. Hilton, P.E. Catellani. “The psychology of counterfactual thinking”. Routledge, 2005.
[26] R.K. Mothilal, A Sharma, C Tan. “Explaining machine learning classifiers through diverse counterfactual explanations.” Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.
[27] L.B. Fulton, J.Y. Lee, Q. Wang, Z. Yuan, J. Hammer, and A. Perer. “Getting playful with explainable AI: games with a purpose to improve human understanding of AI.” In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–8. 2020.