# CHILLI: A data context-aware perturbation method for XAI

**Saif Anwar** [1] **Nathan Griffiths** [1] **Abhir Bhalerao** [1] **Thomas Popham** [1] **Mark Bell** [2] **Shaun Hellman** [2]

## Abstract

The trustworthiness of Machine Learning (ML) models can be difficult to assess, but is critical in high-risk or ethically sensitive applications. Many models are treated as a 'black-box' where the reasoning or criteria for a final decision is opaque to the user. To address this, some existing Explainable AI (XAI) approaches approximate model behaviour using perturbed data. However, such methods have been criticised for ignoring feature dependencies, with explanations being based on potentially unrealistic data. We propose a novel framework, CHILLI, for incorporating data context into XAI by generating contextually aware perturbations, which are faithful to the training data of the base model being explained. This is shown to improve both the soundness and accuracy of the explanations.

## 1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) are increasingly being used to tackle problems in a variety of domains because of their prodigious performance in automated decision-making. Some of these domains have high associated risks, such as financial systems (Ala'raj & Abbod, 2016; Byanjankar et al., 2015), healthcare (Lodhi et al., 2017; Mikalsen et al., 2018) and criminal justice (Rigano, 2019). Incorrect decisions in these scenarios can have significant repercussions, and making decisions with intentional or inadvertent biases can lead to discrimination and other social consequences (Reuters, 2018; Sweeney, 2013). Therefore, it is essential that the decisions made by an ML model are trusted before being acted upon. The foundation of such trust is dependent on both developers and end users understanding the reasoning behind a model's decisions.

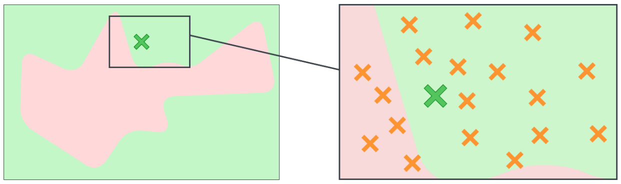
Due to the complexity of many ML techniques, they are often treated as a 'black-box' where the reasoning for a prediction can be difficult to ascertain. Such understanding would allow users to better detect biases in data, assess the vulnerabilities of a model, and ensure a model meets any regulatory standards (Goodman & Flaxman, 2017) and societal expectations. Explainable AI (XAI) methods aim to increase confidence in AI systems, supporting their acceptance and wider adoption. While the use of XAI terminology varies, we define *explainability* as providing evidence or reasoning for all outputs via an explanation, *interpretability* is the notion that all explanations must be understandable to users, and *faithfulness* is a measure of how accurately an explanation reflects the behaviour of a system.

While some ML models are inherently interpretable (Sudjianto & Zhang, 2021), e.g., decision trees, where the behaviour of a model can implicitly be explained (Breiman, 2017), other ML techniques require explanations to be generated separately. Post-hoc XAI attempts to form explanations after a predictive model has been learnt. Such approaches are often model-agnostic and applicable to a range of ML techniques (Goldstein et al., 2014; Molnar, 2023; Ribeiro et al., 2016b). However, evaluations of such approaches have shown that, just as ML models are adapted to their context, XAI systems should also be adapted to the appropriate deployment domain (Zhang et al., 2019; Sokol et al., 2019).

It is often challenging to interpret context from numerical data representing quantitative information, features and signal values. For example, a value of $0.5$ may represent a probability of $50\%$ or a value in the range $[0, 10]$. This is a common problem in XAI, where feature values are used to explain predictions without contextual knowledge of the data (Sokol et al., 2019; Zhang et al., 2019). Earlier works (Lieberman & Selker, 2000; Selker & Burleson, 2000) discuss the importance of context sensitivity for computer systems, which is crucial for XAI in understanding complex ML models. An XAI framework requires underlying domain knowledge of numerical data, to incorporate the appropriate semantics into the explanation. In this paper, we explore the effects of incorporating contextual domain knowledge into XAI, highlighting its importance when explaining predictions. We demonstrate this by evaluating the interpretability and faithfulness of explanations in an intuitive and quantitative manner. The contributions of this paper are as follows.

[1] Department of Computer Science, University of Warwick, UK [2] TRL Limited, UK. Correspondence to: Saif Anwar <saif.anwar@warwick.ac.uk>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) (b)

*Figure 1.* (a) The global decision space, with red and green representing decision regions for two classes. (b) The local area around an instance, with a set of perturbations shown as orange crosses.

- We present an analysis of an existing XAI framework, namely LIME (Ribeiro et al., 2016b), to illustrate the impact on interpretability and faithfulness of explanations when data context is disregarded.
- We propose a method for incorporating data context into a post-hoc XAI framework when using a proxy model.
- Finally, an algorithm is presented for generating local contextually aware data samples for use when fitting a proxy model.

## 2. Related Work

XAI methods either produce global or local explanations (Mohseni et al., 2021). The former explain model behaviour overall (Wang et al.; Lakkaraju et al., 2016), whereas the latter focus on a small area of the decision space, such as around a particular data instance (Ribeiro et al., 2016b; 2018; Zeiler & Fergus, 2013; Baehrens et al., 2009). This is illustrated in Figure 1, which shows the decision regions of a binary classifier, with the red and green areas representing different classes.

### 2.1. Inherently Interpretable Models

The structure of some ML models is inherently interpretable, e.g., linear regression where feature coefficients can be observed (Murphy, 2023), and decision trees where the decision path can be traced. In these cases, an explanation is the base model itself, which is of course completely faithful to its own behaviour. As a result, such model types are often favoured in high-risk scenarios (Rudin, 2019).

### 2.2. Post-hoc Approaches

It is not always be possible to use an inherently interpretable model, for example if a pre-built model requires explaining. Moreover, for some applications the best performing models are highly complex with large numbers of parameters (Simonyan & Zisserman, 2015) and are consequently not inherently interpretable (Wickramanayake et al., 2021). While there is increasing research into high performing inherently interpretable models (Sudjianto & Zhang, 2021), post-hoc XAI methods have been developed to explain existing models. These are typically model-agnostic, using only input and output data to understand model behaviour. Some methods use counterfactual explanations, by highlighting the consequence of modifying an input on a prediction (Verma et al., 2020), while others, such as Shapley values (Messalas et al., 2019) or proxy models, present contributions of features towards a prediction as an explanation.

Proxy models approximate the behaviour of a base model in an interpretable form, such as a decision tree (Schmitz et al., 1999) or linear regression model (Ribeiro et al., 2016b). The proxy model is used as an explanation without sacrificing predictive performance of the base model. This is achieved by fitting a simpler proxy model to the base model predictions. A global proxy model would approximate the base model behaviour in all areas to increase interpretability, however this may not be sufficiently faithful because of oversimplification. It is generally accepted that there is a trade-off between faithfulness and interpretability (Došilović et al., 2018). Therefore, some XAI approaches use a local proxy model fit to the neighbourhood of an instance being explained. This reduces the coverage of the approximation, and so a more faithful explanation, yet with low complexity may be formed (Wood-Doughty et al., 2021).

### 2.3. Perturbation Based Methods

The Local Interpretable Model-Agnostic Explanations (LIME) method (Ribeiro et al., 2016b) explains the prediction for a given instance by fitting a proxy model in its locality. Since there may not be sufficient training data in a locality to fit a proxy model, algorithms such as LIME fit a proxy model to a set of synthetic perturbations of the instance being explained, as illustrated in Figure 1b.

In LIME, a set of perturbed inputs, $\mathcal{Z}$, is generated in the locality of an input instance, $x$, whose output prediction from some base model, $f$, is being explained. The base model is used to predict a set of target values for the perturbations, $f(\mathcal{Z})$. A local proxy model, $g$, is then fit to this perturbed dataset. Non-categorical features are perturbed in LIME by sampling from a Normal distribution with mean and standard deviation estimated from the training data. Samples are taken from the center of the training data and then scaled around the instance (Garreau & von Luxburg, 2020). Categorical features are perturbed by uniformly sampling from the distribution of feature values in the training data.

When fitting the proxy model according to some loss function, the loss contribution of each perturbation, $z \in \mathcal{Z}$, is weighted by a proximity measure, $\pi_x(z)$, according to some distance function, $D(x, z)$, to ensure the explanation is locally focused around $x$. By default in LIME, Euclidean distance is used and is calculated over all feature dimensions (Ribeiro et al., 2016b). The proximity between two instances $\mathbf{p}$ and $\mathbf{q}$, is calculated as shown in Equation 1, where $\sigma$ is a hyperparameter defining the locality of the explanation.

$$\pi_p(q) = exp\left(\frac{-D^2(\mathbf{p},\mathbf{q})}{\sigma^2}\right) \tag{1}$$

In a review of model-agnostic XAI approaches, Molnar et al. (2021b) observe that perturbation-based methods tend to ignore feature dependencies by extrapolating in areas that are not representative of the original data distribution, and are therefore unknown to the base model (Molnar et al., 2021a). They also suggest that ignoring contextual constraints may lead to unrealistic data. For example, when perturbing a feature representing a person's age, the perturbation method must consider that values cannot be negative or unreasonably large (Molnar et al., 2021b). An explanation fit to such data will therefore not be faithful to the true behaviour of the base model, and so additional feature dependency information should be included (Molnar et al., 2021a).

In this paper, we address the importance of feature dependence and propose a new framework, CHILLI, that incorporates dependency information using prior domain knowledge, and explore the effect on the faithfulness of explanations.

### 2.4. Evaluating XAI Methods

A satisfactory explanation provides transparency, allowing users to understand decisions and how data was used. Quantifying properties such as transparency and interpretability is difficult since the desiderata of XAI include subjective properties relating to trust, ethics and understanding.

In this paper, we quantify the performance of proxy-model XAI approaches through the faithfulness of explanations. The faithfulness of a proxy model, $g$, used to explain a base model, $f$, can be calculated using an error metric to compare the predictions made by $f$ and $g$ on a set of input instances. For a global proxy model, the inputs used for evaluation may be from the training data used for the base model, $f$ (Wood-Doughty et al., 2021), whereas for a local proxy model this may be a set of perturbations, $\mathcal{Z}$, around an instance of interest (Ribeiro et al., 2016b). From a set of possible proxy models, $G$, the proxy model with the lowest error, $g$, is selected as the explanation since it is most faithful.

## 3. Contextually Enhanced Interpretable Local Explainable AI

In this section, we propose Contextually Enhanced Intepretable Local Explainable AI (CHILLI), an XAI framework that combines the contextually aware proximity measures and domain representative perturbation generation method presented below to explain base model behaviour using local proxy models. CHILLI aims to satisfy potential contextual constraints and consider limitations of numerical data. Explanations are fit to perturbed data that is representative of the base model training data and is local to the instance being explained. CHILLI is based on LIME (Ribeiro et al., 2016b), with modifications to the proximity calculations and perturbation generation methods.

### 3.1. Contextually Aware Proximity Measures

Proximity in LIME is based on Euclidean distance (see above, Section 2.3), irrespective of the feature type. However, if for some features the absolute difference between two values does not reflect their distance, this proximity measure becomes invalid. This is the case if, for example, the units are not equidistant, or a feature is not measured linearly, such as magnitude recorded on a logarithmic scale. Such distance measures are also unsuitable for cyclic or temporal features e.g., time of day where raw values for 23:00 and 00:00 appear to be far apart, but domain knowledge informs us that they are consecutive.

We propose that the context of features should be considered independently by incorporating the scale and bounds of each feature to ensure the calculated distance is truly representative. Consider the points $\mathbf{p}$ and $\mathbf{q}$, represented by feature vectors of $d$ dimensions. Instead of using a generic distance function, such as Euclidean distance, the distance between the points, $D(\mathbf{p}, \mathbf{q})$, is calculated individually for each feature dimension, $i$, using a specified distance function, $D_i$. The distance in each feature dimension is normalised, to allow for equal contribution, and averaged across all dimensions to give a single distance value. From this, a proximity measure can be calculated, as shown in Equation 2.

$$\pi_{\mathbf{p}}(\mathbf{q}) = exp\left(\frac{-(\frac{1}{d}\sum_{i\in d} D_i(p_i, q_i))^2}{\sigma^2}\right) \tag{2}$$

Since the proximity measure is used when quantifying the performance of each explanation, $g$, an accurate proximity measure is essential to ensure the most faithful explanation model, $g$, is selected from the set of possible explanations, $G$.

The value of the locality hyperparameter, $\sigma$, may be adjusted to vary the locality of an explanation in the model space. As $\sigma$ increases, the proximity tends to 1, as shown in Figure 2 for a range of distance values. All perturbations for a high enough value of $\sigma$, regardless of distance, will be assigned a proximity of 1 and are considered equally when selecting the best fit proxy model. Conversely, a smaller value of $\sigma$ will result in greater variation between proximities for perturbations of differing distance to the instance being explained. This leads to the selection of a proxy model that performs better on perturbations of closer proximity, thus reducing the locality of the explanation.

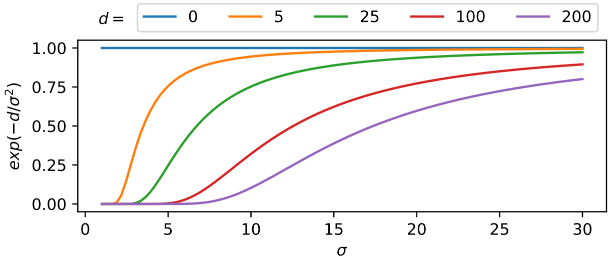


*Figure 2.* A comparison of the effect on proximity between two points of fixed distance, $d$, as the locality parameter, $\sigma$, is varied.

### 3.2. Domain Representative Perturbation Generation

Existing perturbation-based XAI methods, such as LIME, do not consider contextual knowledge when generating the perturbations which are the foundation for creating an explanation (Ribeiro et al., 2016a;b; Zhang et al., 2019; Sokol et al., 2019). Perturbations in LIME are sampled uniformly from the entire feature space, and a proxy model fit to these will focus on all relationships in the feature space. Such an explanation is not local to the instance being explained, but is generalised to the overall model behaviour. Presenting this as a local explanation may be misleading regarding the behaviour of the model.

Such perturbations also ignore any bounds on feature values, such as 'Age' which can only take positive values. Without considering these bounds, perturbations may contain unrealistic values. Moreover, since features are perturbed independently, feature dependencies are ignored. For example, assume that as the population of a city grows, traffic congestion increases. Although these features are correlated, ignoring feature dependancies may result in a perturbation combining lowered congestion with increased population. The omission of such dependencies may result in an unrealistic set of perturbations, leading to an explanation that does not describe the behaviour of the base model (Laugel et al., 2019). An XAI framework should generate perturbations that are representative of real-world data, such as the training data, and are local to the instance being explained.

We present a framework, CHILLI, for generating local and contextually conforming perturbations. Our method takes inspiration from the SMOTE algorithm (Chawla et al., 2002). SMOTE is a well-regarded sampling technique used to generate synthetic data when there is class imbalance (Chawla et al., 2002). A random point is selected from the minority class and its $k$-nearest-neighbours are located, for a predetermined number, $k$. One of these neighbours is selected uniformly at random and a synthetic datapoint is generated by linearly interpolating between the two points and uniformly selecting a random point on the line joining the two. This is repeated for different instances from the minority class until a specified degree of over-sampling has been achieved.

We use this approach to generate perturbations, and to ensure perturbations fall within realistic bounds, they are produced by interpolating between an instance being explained, $x$, and some other randomly selected instance, $x'$, in the training data. Each feature value is interpolated independently. For categorical features, the interpolated value is rounded to the nearest feature value.

To maintain the locality of the perturbations, the selection of $x'$ is from a probability distribution calculated using proximity (Equation 2) which is normalised for all data points such that $P(x' = x_i) = \pi_x(x_i)$ and $\sum_{x_i \in \mathbf{X}} P(x' = x_i) = 1$. As a result, perturbations are more likely to contain values in closer proximity to $x$. The process for generating a set of $N$ perturbations is outlined in Algorithm 1.

**Algorithm 1** Contextual Perturbation Generation

```
Input: Number of perturbations to generate, $N$; Data instance to perturb, $x$; Training dataset, $\mathbf{X}$
Output: Set of perturbations $\mathcal{Z}$
$F$ = Features of $x$
Initialise empty set of perturbations, $\mathcal{Z} = [\,]$
Calculate $\pi_x(x^i)$ for each $x^i \in \mathbf{X}$
Assign a probability to each $x^i$ where $P(x' = x^i) = \frac{\pi_x(x^i)}{max(\pi_x(x^i) \forall x^i \in \mathbf{X})}$
while $|\mathcal{Z}| \leq N$ do
    Uniformly select some value between 0 and 1 $\rightarrow I$
    Select some $x^i \in \mathbf{X}$ based on probability for each $x^i \rightarrow x'$
    for $f$ in $F$ do
        $z_f = x_f + I(x'_f - x_f)$
    end for
    $\mathcal{Z} = \mathcal{Z} \cup \{z\}$
end while
```

## 4. Experimental Setup

We compare the functionality and performance of our proposed method, CHILLI, with that of LIME (Ribeiro et al., 2016b), to explore the effect of incorporating contextual information into XAI frameworks.

### 4.1. Datasets

Our evaluation uses the WebTRIS and MIDAS datasets. WebTRIS (National Highways, 2017), recorded by Highways England, contains traffic data at 15 minute intervals for many motorway sites around England. We restrict the dataset to two sites (M60/9094A and M6/7570A) between 01/01/2016 and 01/01/2017. Each site is modelled individually to monitor performance consistency across the dataset. A Support Vector Regressor (SVR) is trained on a subset of the available features describing date, time, average speed and number of vehicles of various sizes, to predict the 'Total Volume' of traffic per time interval. The data distribution in the individual feature dimensions is shown in Figure 3.

MIDAS (Office, 2022), published by the UK Meteorological Office, records hourly weather observations at multiple locations across the UK. A Recurrent Neural Network (RNN) is trained on data containing various weather features from a station located at Keswick and 3 neighbouring stations (St. Bees Head, Shap and Warcop Firing Range) to predict 'Air Temperature' at Keswick at a given time. It is expected that observations from surrounding areas will be related to the upcoming weather at Keswick, and therefore the data used from neighbouring stations is offset by 1 hour. The distribution of the training data is shown in Figure 4.

We can hypothesise about expected feature importance due to the linearity of the feature relationships against the target variable. Since LIME scales perturbations according to the covariance of the feature against the target variable, we explore the effect on explanation performance of removing generally linear features from the MIDAS data (namely, those describing relative humidity and dewpoint).

### 4.2. Forming Explanations

Explanations containing a set of linear coefficients are produced using CHILLI and LIME for predictions made by a base model. The magnitude of each coefficient indicates the contribution of the corresponding feature towards the base model prediction, whilst the sign indicates the direction of the correlation between the feature and the target variable.

We quantify the performance of an explanation using its error, which represents its faithfulness towards the base model. Explanations for WebTRIS predictions are quantified using RMSE and MAE is used for MIDAS predictions. We compare the error for explanations produced using CHILLI and LIME over 25 instances, selected uniformly at random. The selected instances are shown in Figures 3 and 4.

## 5. Results & Discussion

In this section, we use LIME and CHILLI to fit a local proxy model which is used to explain a prediction produced by a base model for a given instance.

### 5.1. Perturbation Generation

Figure 5 shows explanations produced by CHILLI and LIME alongisde the perturbations used to fit them for a prediction made by the SVR base model, $f$, for a randomly selected instance, $x$, from the WebTRIS M6/7570A test dataset, $\mathcal{X}$. The instance is shown as the black point in each of its feature dimensions against the target 'Total Volume' value. The perturbations of the instance with the prediction of the target feature 'Total Volume' from the base model, $f(z)$, are shown as orange points.

From a visual inspection of Figure 5a, it can be seen that the perturbations of features generated by LIME do not follow the data distribution shown in Figure 3. Moreover, since each feature is perturbed independently, feature values in a single perturbation do not consider feature dependencies, which leads to unrealistic perturbations being generated. For example, a perturbation may have a 'Time Interval' of 03:00, but the value of '0-520cm' may correspond to the number of vehicles that would be observed at rush hour.

The bounds of features have also not been considered, as can be seen in Figure 5a where all non-categorical features exhibit perturbed values which fall outside the normalised range of $[0, 1]$. Negative values of '0-520cm', '521-660cm' and '661-1160cm' imply a negative number of vehicles of the respective sizes passing in the corresponding time interval, which is not possible. This leads to a set of perturbations that do not represent real-world data, and therefore do not represent the training data. The negative impact of such inappropriate perturbations can be observed from the predicted values from the base model, which often predicts a negative volume of traffic flow, which is also not possible. An explanation that is fit on such perturbations will not correctly represent the true behaviour of the base model.

The opacity of each perturbation, shown in Figure 5a, signifies its calculated proximity weighting, $\pi_x(z)$, to the instance being explained. Perturbations which are further from the instance are sometimes assigned a higher weighting than those which are closer. Since this indicates the contribution of each perturbation to the selection of the best fit linear proxy model, the produced explanation will not be locally focused around the instance being explained, and is instead a generalised explanation across all the perturbations.

On the other hand, the perturbations generated by CHILLI not only conform to the distribution of the training data shown in Figure 3, but are also realistic combinations of feature values that fall within the appropriate feature bounds. CHILLI also generates perturbations with greater density around the instance being explained, which can be seen in Figure 5a from the concentration of orange points around

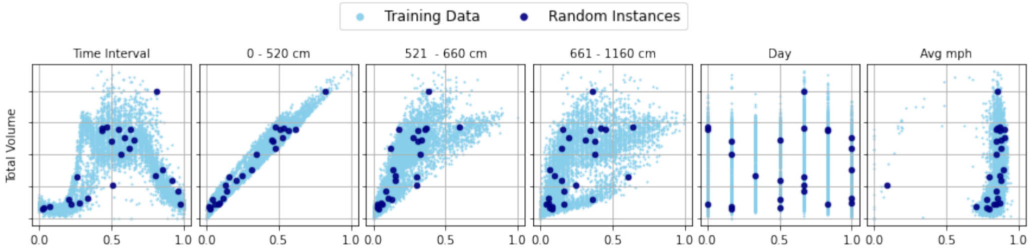


*Figure 3.* Normalised WebTRIS training data shown in each feature dimension against the target ‘Total Volume’ feature. The 25 instances selected uniformly at random for evaluation are shown as the dark blue points.

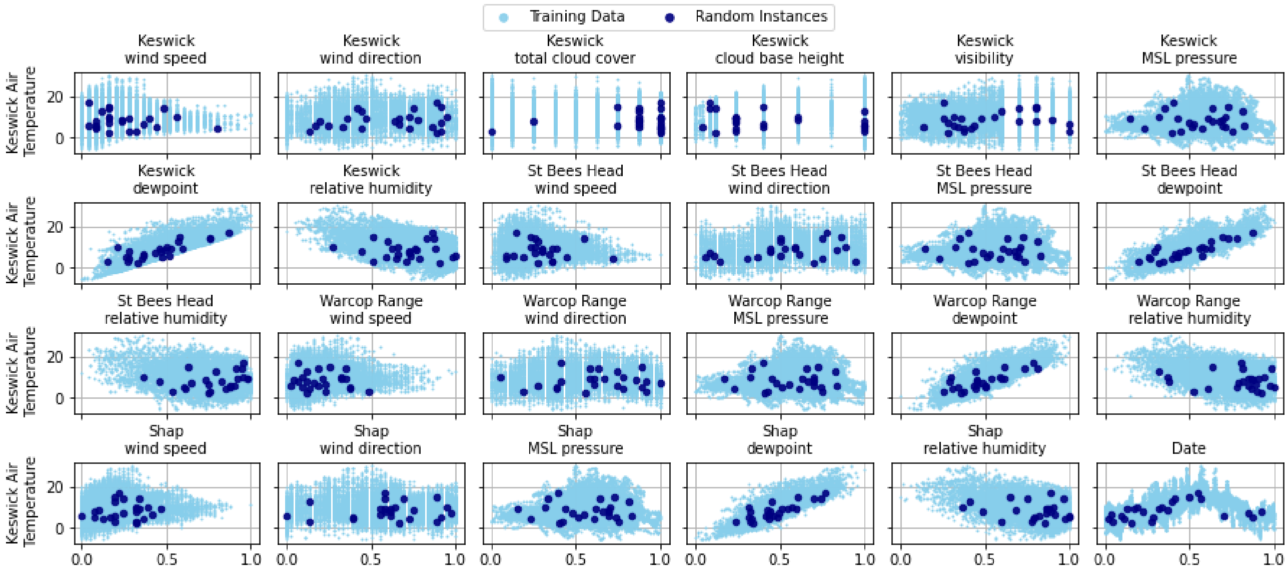
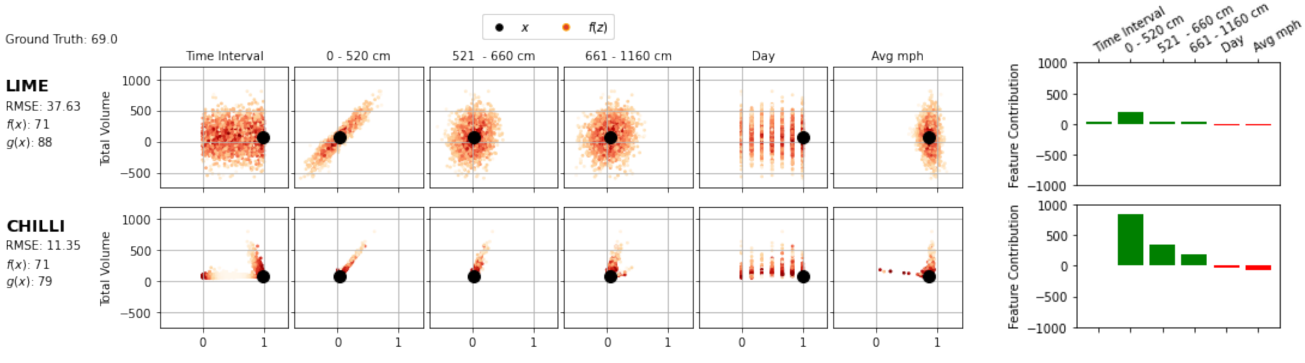


*Figure 4.* Normalised MIDAS training data shown in each feature dimension against the target ‘Keswick Air Temperature’ feature. The 25 instances selected uniformly at random for evaluation are shown as the dark blue points.



(a)

(b)

*Figure 5.* (a) Perturbations (orange points) generated using LIME and CHILLI for a WebTRIS data instance, $x$. The predicted ‘Total Volume’ for each perturbation from the base model, $f(z)$, is shown on the vertical axis. Opacity of perturbations represent proximity to $x$. (b) Explanations produced by CHILLI and LIME, showing the feature coefficients of the linear proxy model fit to the perturbations in (a) representing the contribution of each feature towards the predicted target value, $f(x)$.

the instance. This is also the case for features of a cyclic nature, such as 'Time Interval' in WebTRIS, where $0$ and $1$ are adjacent values. This leads to an explanation that is fit with greater emphasis on perturbations of closer locality.

## 5.2. Feature Contributions

The linear proxy model with the lowest error when fit to the set of perturbations, $\mathcal{Z}$, and base model predictions, $f(\mathcal{Z})$, is selected as the explanation for the instance being explained. The explanations shown in Figure 5b indicate that CHILLI produces explanations with greater disparity between feature coefficients. It is expected for explanations produced by CHILLI to have larger feature coefficients than LIME, since the perturbations generated by LIME are based on a Normal distribution which naturally does not exhibit any linear correlation. Due to the covariance scaling of LIME perturbations towards the training data, some features, such as '0-520cm' in the WebTRIS data, exhibit a strong linear correlation with 'Total Volume' as shown in Figure 3. LIME recognises this and identifies it as the most significant feature contribution in its explanation. Similarly, features in the MIDAS data containing 'dewpoint' and 'relative humidity' have a linear correlation with 'Keswick Air Temperature' across their range of values, as can be seen from Figure 4.

Figure 6 shows the variation in feature contributions in explanations produced for the 25 instances shown in Figure 4. Again, only generally linear features have noticeable contribution in the explanations produced by LIME. This is unsuitable since general feature trends are not relevant in a locally focused explaination. CHILLI produced explanations that are fit to perturbations local to the instance being explained, and recognises general linear trends in cases where the linear relationship is also present locally. However, CHILLI also highlights contributions from other locally impactful features, although they are not as significant as the generally linear features.

Upon removal of generally linear features, there is greater variation in the explanations produced by CHILLI, as shown in Figure 7. Since LIME cannot detect local behaviour, it performs poorly when locality is important, and does not identify any significant feature contributions due to the absence of general trends. CHILLI, on the other hand, is able to detect local trends and achieves significantly lower MAE, indicating that the explanations produced by CHILLI are more faithful to the true behaviour of the base model.

## 5.3. Explanation Faithfulness

As noted in Section 2.4, a lower error indicates a more faithful explanation. The explanation produced by CHILLI achieved a signifcantly lower RMSE than LIME on the perturbations shown in Figure 5. The explanation produced by CHILLI predicted the target variable for the instance to be 79 whilst LIME's explanation predicted 88. The lower error of the CHILLI explanation, combined with a prediction closer to the base model, supports the conclusion that CHILLI produces a more faithful explanation, and is more representative of the base model's true behaviour.

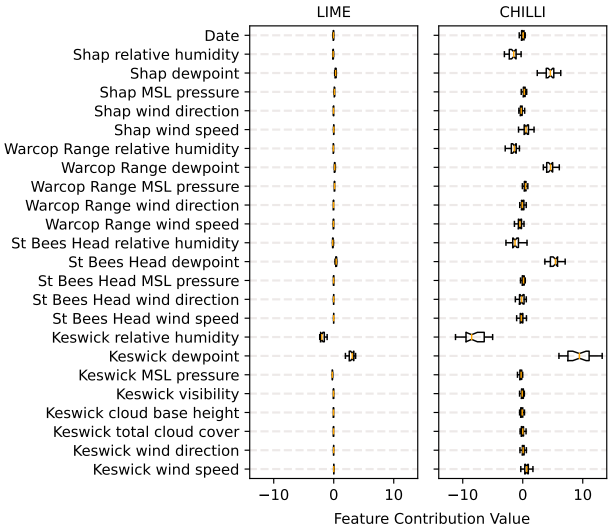


*Figure 6.* Variation in feature contributions presented in explanations produced by LIME and CHILLI across the 25 instances shown in Figure 4. The median value and quartile ranges are shown for each feature.

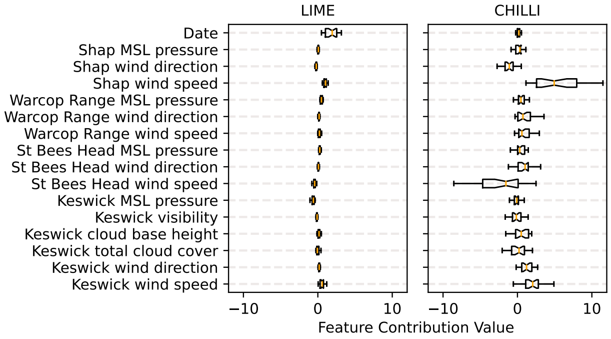


*Figure 7.* Variation in feature contributions presented in explanations produced by LIME and CHILLI for the 25 instances shown in Figure 4 when generally linear features are excluded. The median value and quartile ranges are shown for each feature.

Figure 8 shows a comparison of the error achieved by explanations produced by LIME and CHILLI for both WebTRIS and MIDAS for the 25 instances shown in Figures 3 and 4. Explanations were also produced for the MIDAS instances after removing generally linear features. In all explained instances, CHLLI achieves a lower error than LIME. The average error for each technique across all instances is also indicated in Figure 8. CHILLI leads to an average reduc-

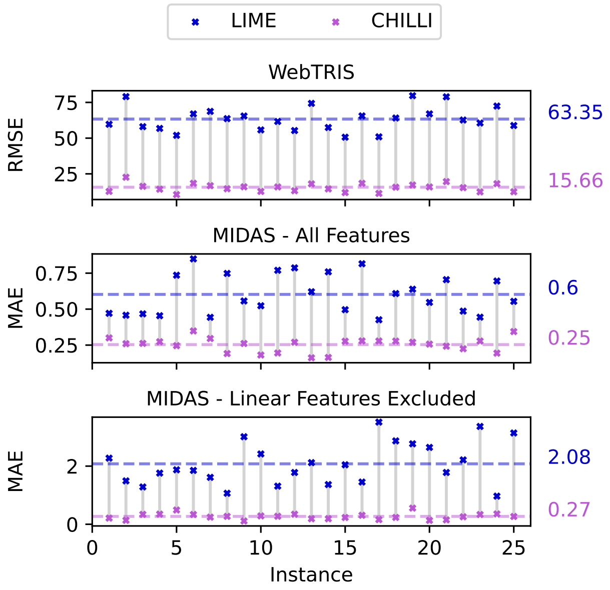


*Figure 8.* Individual (crosses) and average (dashed line) error achieved by explanations from LIME and CHILLI for 25 randomly selected instances from WebTRIS and MIDAS.

tion in error of 75% and 58% on WebTRIS and MIDAS respectively, with all features are included. After removing generally linear features, the error of explanations produced by LIME increases significantly, whereas CHILLI maintains a similar error since it captures other local trends. This leads to an average reduction in error of 87% for CHILLI compared to LIME.

### 5.4. Locality Hyperparameter Exploration

The importance of accurate proximity measurements can be understood by observing the effect of varying $\sigma$ on MAE. Figure 9 shows a comparison of the MAE achieved by LIME and CHILLI for a uniformly randomly selected instance from MIDAS, when using different values of $\sigma$. The MAE achieved by LIME is similar across all values of $\sigma$. Since LIME forms explanations that do not consider the local data context of the instance being explained, it is not expected for them to vary based on locality size. CHILLI achieves lower MAE for lower values of $\sigma$ before stabilising at higher values. As the defined locality increases, perturbations are considered that are further from the instance being explained. Features which do not exhibit linear relationships on a broader scale are difficult to describe using a linear proxy model. This leads to a worse performing explanation, since it attempts to generalise behaviour rather than explaining local trends, as with LIME. While MAE increases when using CHILLI, it still outperforms LIME since the intuition regarding perturbation generation is sound.

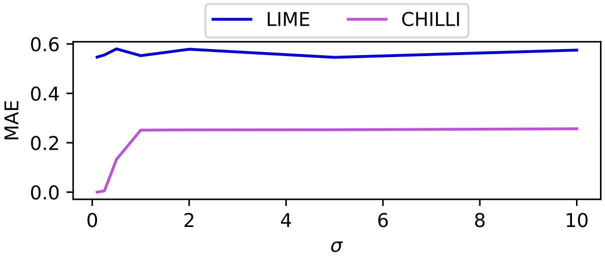


*Figure 9.* A comparison of the MAE achieved by explanations produced using LIME and CHILLI for a single instance of the MIDAS data, whilst varying the locality parameter, $\sigma$.

## 6. Conclusion and Future Work

In this paper, we explored the effect of incorporating contextual domain knowledge into a model-agnostic local perturbation-based XAI approach, namely LIME. We proposed a method for contextually aware proximity measures, to ensure locality is accurately defined and constrained. We also proposed a method for generating perturbations that consider the contextual limitations and dependencies of data features. These methods are combined into a new framework, CHILLI, for generating local explanations for blackbox ML models, and we compared the functionality and performance of CHILLI with LIME.

Using the WebTRIS and MIDAS datasets, we demonstrated that LIME does not appropriately measure proximity between instances, resulting in an explanation which is not local to the instance being explained. Explanations generated by LIME were found to be generalised and only consider features with general linear trends. It was also found that LIME does not generate perturbations that are representative of the training data, and the perturbations contained unrealistic values.

CHILLI was shown to generate perturbations that are representative of the base model training data and are local to the instance being explained. Therefore, CHILLI's explanations had relatively larger feature contributions compared to those produced by LIME. CHILLI consistently achieved a lower error, and therefore produced a more faithful explanation, across all explained instances compared to LIME.

Through empirical and intuitive evaluation of LIME and CHILLI, we conclude that incorporating contextual domain knowledge regarding data features used for generating explanations improves faithfulness, which may ultimately increase trust in both the explanation and explanation framework. In future work we will investigate how improving the performance of local explanations affects the overall trust in a model. We would also like to explore the efficacy of CHILLI when proxy models of a different form are used, such as decision trees or small-order polynomial regressors.

## References

Ala'raj, M. and Abbod, M. F. Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, 104:89–105, July 2016. ISSN 09507051. doi: 10.1016/j.knosys.2016.04.013. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705116300569.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Mueller, K.-R. How to explain individual classification decisions. (arXiv:0912.1128), Dec 2009. URL http://arxiv.org/abs/0912.1128. arXiv:0912.1128 [cs, stat].

Breiman, L. *Classification and Regression Trees*. Routledge, New York, October 2017. ISBN 978-1-315-13947-0. doi: 10.1201/9781315139470.

Byanjankar, A., Heikkila, M., and Mezei, J. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. In *2015 IEEE Symposium Series on Computational Intelligence*, pp. 719–725, Cape Town, December 2015. IEEE. ISBN 978-1-4799-7560-0. doi: 10.1109/SSCI.2015.109. URL http://ieeexplore.ieee.org/document/7376683/.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://arxiv.org/abs/1106.1813. arXiv:1106.1813 [cs].

Došilović, F. K., Brčić, M., and Hlupić, N. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 0210–0215, May 2018. doi: 10.23919/MIPRO.2018.8400040.

Garreau, D. and von Luxburg, U. Explaining the explainer: A first theoretical analysis of lime. (arXiv:2001.03447), Jan 2020. URL http://arxiv.org/abs/2001.03447. arXiv:2001.03447 [cs, stat].

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. Technical Report arXiv:1309.6392, arXiv, March 2014. URL http://arxiv.org/abs/1309.6392. arXiv:1309.6392 [stat] type: article.

Goodman, B. and Flaxman, S. European Union regulations on algorithmic decision-making and a ”right to explanation”. *AI Magazine*, 38(3):50–57, October 2017. ISSN 2371-9621, 0738-4602. doi: 10.1609/aimag.v38i3.2741. URL http://arxiv.org/abs/1606.08813. arXiv:1606.08813 [cs, stat].

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684, San Francisco California USA, Aug 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939874. URL https://dl.acm.org/doi/10.1145/2939672.2939874.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2801–2807, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/388. URL https://www.ijcai.org/proceedings/2019/388.

Lieberman, H. and Selker, T. Out of context: Computer systems that adapt to, and learn from, context. *IBM Systems Journal*, 39(3.4):617–632, 2000. ISSN 0018-8670. doi: 10.1147/sj.393.0617. Conference Name: IBM Systems Journal.

Lodhi, M. K., Ansari, R., Yao, Y., Keenan, G. M., Wilkie, D., and Khokhar, A. A. Predicting Hospital Re-admissions from Nursing Care Data of Hospitalized Patients. *Advances in data mining. Industrial Conference on Data Mining*, 2017:181–193, 2017. doi: 10.1007/978-3-319-62701-4_14. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665368/.

Messalas, A., Kanellopoulos, Y., and Makris, C. Model-Agnostic Interpretability with Shapley Values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–7, July 2019. doi: 10.1109/IISA.2019.8900669.

Mikalsen, K. Ø., Soguero-Ruiz, C., Bianchi, F. M., Revhaug, A., and Jenssen, R. An Unsupervised Multivariate Time Series Kernel Approach for Identifying Patients with Surgical Site Infection from Blood Samples. Technical Report arXiv:1803.07879, arXiv, March 2018. URL http://arxiv.org/abs/1803.07879. arXiv:1803.07879 [cs, stat] type: article.

Mohseni, S., Zarei, N., and Ragan, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4):1–45, Dec 2021. ISSN 2160-6455, 2160-6463. doi: 10.1145/3387166.

Molnar, C. *Interpretable Machine Learning*. Lulu.com, 2 edition, 2023. URL https://christophm.github.io/interpretable-ml-book.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach. Technical Report arXiv:2006.04628, arXiv, June 2021a. URL http://arxiv.org/abs/2006.04628. arXiv:2006.04628 [cs, stat] type: article.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. Technical Report arXiv:2007.04131, arXiv, August 2021b. URL http://arxiv.org/abs/2007.04131. arXiv:2007.04131 [cs, stat] type: article.

Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL probml.ai.

National Highways. WebTRIS , January 2017. http://webtris.highwaysengland.co.uk/.

Office, M. MIDAS Open: UK daily weather observation data, v202207, 2022. URL https://catalogue.ceda.ac.uk/uuid/4b44cec2f9a846f39d5007983b7eaaab.

Reuters. Amazon ditched AI recruiting tool that favored men for technical jobs, October 2018. ISSN 0261-3077. URL https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-Agnostic Interpretability of Machine Learning, June 2016a. URL http://arxiv.org/abs/1606.05386. arXiv:1606.05386 [cs, stat].

Ribeiro, M. T., Singh, S., and Guestrin, C. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier, August 2016b. URL http://arxiv.org/abs/1602.04938. arXiv:1602.04938 [cs, stat].

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Rigano, C. Using Artificial Intelligence to Address Criminal Justice Needs. *US NIJ Journal*, 280, January 2019. www.nij.gov/journals/280/Pages/using-artificial-intelligence-to-address-criminal-justice-needs.aspx.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. (arXiv:1811.10154), Sep 2019. URL http://arxiv.org/abs/1811.10154. arXiv:1811.10154 [cs, stat].

Schmitz, G., Aldrich, C., and Gouws, F. ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10 (6):1392–1401, November 1999. ISSN 1941-0093. doi: 10.1109/72.809084. Conference Name: IEEE Transactions on Neural Networks.

Selker, T. and Burleson, W. Context-aware design and interaction in computer systems. *IBM Systems Journal*, 39(3.4):880–891, 2000. ISSN 0018-8670. doi: 10.1147/sj.393.0880. Conference Name: IBM Systems Journal.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs].

Sokol, K., Hepburn, A., Santos-Rodriguez, R., and Flach, P. bLIMEy: Surrogate Prediction Explanations Beyond LIME, October 2019. URL http://arxiv.org/abs/1910.13016. arXiv:1910.13016 [cs, stat].

Sudjianto, A. and Zhang, A. Designing Inherently Interpretable Machine Learning Models. Technical Report arXiv:2111.01743, arXiv, November 2021. URL http://arxiv.org/abs/2111.01743. arXiv:2111.01743 [cs, stat] type: article.

Sweeney, L. Discrimination in online ad delivery. (arXiv:1301.6822), Jan 2013. URL http://arxiv.org/abs/1301.6822. arXiv:1301.6822 [cs].

Verma, S., Dickerson, J., and Hines, K. Counterfactual Explanations for Machine Learning: A Review. Technical Report arXiv:2010.10596, arXiv, October 2020. URL http://arxiv.org/abs/2010.10596. arXiv:2010.10596 [cs, stat] type: article.

Wang, X., Liu, S., Liu, J., Chen, J., Zhu, J., and Guo, B. Topicpanorama: A full picture of relevant topics. pp. 14.

Wickramanayake, S., Hsu, W., and Lee, M. L. Towards Fully Interpretable Deep Neural Networks: Are We There Yet? Technical Report arXiv:2106.13164, arXiv, June 2021. URL http://arxiv.org/abs/2106.13164. arXiv:2106.13164 [cs] type: article.

Wood-Doughty, Z., Cachola, I., and Dredze, M. Proxy model explanations for time series rnns. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 698–703, Pasadena, CA, USA, Dec 2021. IEEE. ISBN 978-1-66544-337-1. doi:

10.1109/ICMLA52953.2021.00117. URL https://ieeexplore.ieee.org/document/9680082/.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. (arXiv:1311.2901), Nov 2013. URL http://arxiv.org/abs/1311.2901. arXiv:1311.2901 [cs].

Zhang, Y., Song, K., Sun, Y., Tan, S., and Udell, M. ”Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations, June 2019. URL http://arxiv.org/abs/1904.12991. arXiv:1904.12991 [cs, stat].