# Uncertainty Quantification of Surrogate Explanations: an Ordinal Consensus Approach

Jonas Schulz[1,2], Rafael Poyiadzi[2], and Raul Santos-Rodriguez*[2]

[1]TU Dresden
[2]University of Bristol

## Abstract

Explainability of black-box machine learning models is crucial, in particular when deployed in critical applications such as medicine or autonomous cars. Existing approaches produce explanations for the predictions of models, however, how to assess the quality and reliability of such explanations remains an open question. In this paper we take a step further in order to provide the practitioner with tools to judge the trustworthiness of an explanation. To this end, we produce estimates of the uncertainty of a given explanation by measuring the ordinal consensus amongst a set of diverse bootstrapped surrogate explainers. While we encourage diversity by using ensemble techniques, we propose and analyse metrics to aggregate the information contained within the set of explainers through a rating scheme. We empirically illustrate the properties of this approach through experiments on state-of-the-art Convolutional Neural Network ensembles. Furthermore, through tailored visualisations, we show specific examples of situations where uncertainty estimates offer concrete actionable insights to the user beyond those arising from standard surrogate explainers.

## 1 Introduction

Deep learning models are being used in critical applications which demand not only human oversight but for their predictions to be explained as well, if they are to be considered *trustworthy*. Explainability tools have been developed aiming to make black-box classifiers interpretable ([3, 29]). Surrogate explainers, such as Local Interpretable Model-agnostic Explanations (LIME) [25], provide an explanation by fitting an interpretable surrogate model to explain the prediction of an instance.

However, explanations produced by LIME can vary due to the hyperparameters of the procedure. Several papers have looked into the shortcomings of LIME and proposed more robust versions ([28, 15]). In particular, the main sources of uncertainty affecting LIME explanations are studied in [32]. Their work analyses the uncertainty due to LIME's hyperparameters, and also the stochasticity in the process of generating the explanation. Similarly, [10] presents both a theoretical and an empirical analysis of the variability of explanations produced for a single image. These results suggest that the inherent stochasticity of LIME induces diversity among multiple explanations produced for the same instance. The idea of applying LIME multiple times to an instance is proposed in [12], while the robustness of LIME, with regards to changes in the input data, is explored in [1].

In this paper, we take a step back and aim to enrich the explanations by incorporating an estimate of their uncertainty. This allows for a more meaningful interaction, potentially enabling the user to either trust or reject the explanation. Our contributions are as follows:

1. We provide uncertainty estimates for explanations using bootstrapping and ordinal consensus metrics. We showcase these using tailored visualisations that convey this information for the practitioner.

2. Beyond the uncertainty within LIME and the uncertainty induced by the input data, we also

*This work was funded by the UKRI Turing AI Fellowship EP/V024817/1.



https://doi.org/10.7557/18.6294

consider the predictive uncertainty. We do this by considering the model of interest to be an ensemble of black-box models, rather than a single black-box.

## 2 Related Work

The process of deriving surrogate explainers is complex and driven by several interconnected factors and objectives ([24, 23]). In general, this type of explainers can be unstable and lead to varying surrogate coefficients and, in consequence, diverse explanations ([1, 32, 33]). The variability within surrogate coefficients can be seen as uncertainty that surrogate explanations are entailed with. While [12] and [10] highlight the sampling space where the surrogate is fitted as a source of uncertainty, [2] motivates the need of also considering the predictive uncertainty of the black-box to be explained. `BayesLIME` was proposed in [27] to generate surrogate explanations with a measure of uncertainty. There, the uncertainty is quantified by evaluating the probability that surrogate coefficients lie within their 95% credible intervals. The work suggests sampling perturbations that yield most information to the models behaviour, thus reducing the computational complexity. The practitioner is informed about the uncertainty of feature attribution to each explainable component.

In this work we address the quantification of the surrogate explanation uncertainty by aggregating multiple surrogate coefficients and measuring the consensus among the surrogate explainers. The use of a consensus mechanism to obtain explanations that are less sensitive to sampling variance (further discussed in Section 3) has been proposed in [4, 26]. Specifically, [6] and [5] consider aggregating surrogate coefficients in the form of simple ranking schemes inspired from the social sciences and economics. Additionally, we investigate how the estimated consensus behaves as the number of sampled perturbations and surrogates change. As deriving a large number of surrogate on large sets of perturbations is computationally expensive [27], we motivate the aforementioned two factors as key parameters the practitioner needs to fine tune in order to derive explanations that satisfy the desired certainty without being computationally expensive. On top of existing works, we propose to look at explaining uncertainty-aware deep learning ensembles through surrogate explanations to combine surrogate explanations with model uncertainty [2]. We show that the diversity in the ensemble members can be used to induce variability in the surrogate coefficients, allowing the generation of a diverse set of surrogate explainers as in [31, 30]. There, the authors derive multiple diverse explanations for a deep learning ensemble in the form of saliency maps. Whereas in [31] the saliency maps are derived for each ensemble member individually, we follow the approach pursued in [30], where the ensemble is treated as a probabilistic classifier as it does not restrict the number of derived surrogates to the number of members in the ensemble.

## 3 Background

Local-surrogate explanations belong to a category of post-hoc model-agnostic explanation approaches first introduced in [25]. One such approach is LIME, which is an instantiation of the following formulation:

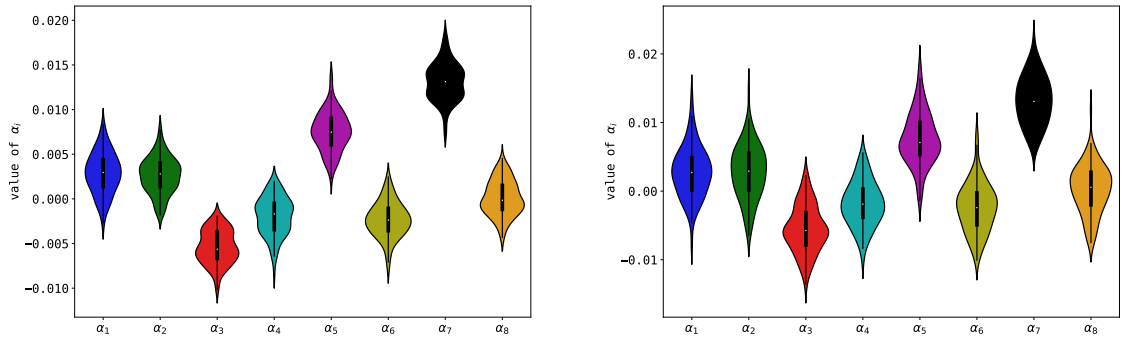
$$\arg\min_{g \in \mathcal{G}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g). \tag{1}$$

The surrogate explainer $g$ is from an interpretable model class $\mathcal{G}$. The locality around the data point $\mathbf{x}$ for which the prediction of a classifier $f$ is to be explained, is controlled by the similarity kernel $\pi_x$. The loss $\mathcal{L}$ characterises how close $g$ is to $f$. The penalty term $\Omega$ represents a complexity measure of $g$. In practice $G$ is the class of linear models: $g(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{x}$. Model fitting is performed on a set of points $\mathcal{P}$ drawn from a Gaussian distribution centred on $\mathbf{x}$, and then the weights $\pi_x$ are computed using the radial basis function kernel.

Previous works have identified the following sources of uncertainty in surrogate explanations:

1. **Sampling variance of $\mathcal{P}$** (examined in [19, 10, 12, 1]).
2. **Implementation of explanation procedure** (highlighted in [32]).
3. **Choice of surrogate structure** (introduced in [32]).

In this paper, assuming sources 2 and 3 are fixed, we focus on the sampling variance as well as the




(a) Variability due to sampling variance

(b) Variability due to predictive uncertainty

Figure 1: (a) Distribution of surrogate coefficients $\alpha_i$ derived by LIME on the image depicted in Fig. 4 (left column) on 100 different perturbation sets $\mathcal{P}_k$. (b) Distribution of surrogate coefficients $\alpha_i$ derived by LIME on the image depicted in Fig. 4 (left column). LIME is run 100 times with a fixed perturbation set $\mathcal{P}$. The classifier's prediction is sampled randomly from the ensemble members.

predictive uncertainty of the model to be explained. We argue that the predictive uncertainty of the model $f$ can be seen as an extra source of variability, adding to the uncertainty of the explanations. To show this, in the examples below we use ensemble models for their nice properties on uncertainty estimation. Details on the architecture of the ensemble are given in Sec. 5.

**Variability of surrogate coefficients due to sampling variance** Following the works of [10] and [12], Fig. 1a shows the variability of surrogate coefficients $\boldsymbol{\alpha}$ due to sampling variance in an image classification task. In images, the explanations are usually based on superpixels given by a segmentation of the image with semantic meaning. Here, LIME is run 100 times (by first drawing 100 distinct sets of points, $\{\mathcal{P}_k\}_{k=1}^{100}$) resulting in 100 surrogates with the default configuration. We generate the predictions by averaging the predictions of the individual ensemble members. Since the sets $\mathcal{P}_k$ of image perturbations are generated randomly, values of $\boldsymbol{\alpha}$ are not deterministic. We see that the mean value of $\alpha_7$ is the highest, suggesting that the superpixel $s_7$ can be more clearly identified as, on average, the most relevant region of the image for the classification purpose. $\alpha_3$ can be identified as the least important. For the rest, the ordering is not clear. This is important if the user is interested in tuning LIME such that the distributions of $\alpha_i$ do not overlap so that the order of importance of the coefficients can be clearly identified.

**Variability of explanations due to predictive uncertainty** The uncertainty of LIME due to the predictive uncertainty of the black-box classifier has not been addressed in previous works. However, this is something that we can study when using ensemble models. In Fig. 1b LIME is run 100 times on a *fixed set of image perturbations* $\mathcal{P}$ (Compared to above, in this experiment we only have one set of points $\mathcal{P}$, as opposed to above where we use 100 $\{\mathcal{P}_k\}_{k=1}^{100}$). Here, a single prediction is obtained from a randomly chosen member of the ensemble. Therefore, differently from the experiment presented in Fig. 1a, the variability of the surrogate coefficients is now solely induced by sampling the predictions $f(\mathbf{x'}_i)$ for image perturbations $\mathbf{x'}_i$ randomly from the individual models contained in the ensemble. Again, for this particular image, the top and bottom coefficients remain the same as before, corroborating the message from the previous example. In Sec. 5 we will further explore this relationship empirically. In Sec. 4, we present a method of deriving multiple diverse surrogates and aggregate their coefficient values through a rating scheme to estimate the uncertainty of the aggregated explanation.

# 4 Uncertainty Quantification via Ordinal Consensus

The coefficients of the surrogate are representative of the behaviour of $f$ locally. A common method for estimating the distribution of surrogate coefficients $\boldsymbol{\alpha}$ is bootstrap ([8]), as the sampling variance of data points around $\mathbf{x}$ naturally induces diversity among bootstrapped surrogates. Here, we propose the use of ensemble techniques to account for the stochasticity of the prediction behaviour of $f$, reinforcing the diversity of the bootstrapped surrogates. We suggest ordinal metrics to aggregate the surrogate coefficients and quantify uncertainty.

## 4.1 Bootstrapping LIME

In our approach, that we refer to as Bootstrapping LIME (BLIME), multiple surrogate models are fitted by bootstrapping the perturbation dataset $\{\mathcal{P}_k\}_{k=1}^{K}$. Since an ensemble model can be treated as a probabilistic classifier, the output $f(\mathbf{x}'_i)$ for a perturbation $\mathbf{x}'_i$ can also be sampled from the ensemble classifier by sampling a base model from the set of models.

**Algorithm 1** BLIME

**Input:** classifier $f$
**Input:** instance $\mathbf{x}$ to be explained
**Input:** class label $l$ to be explained
**Input:** similiarity kernel $\pi_x$
**Input:** number of perturbations $N$
**Input:** number of bootstrap samples $K$
$\mathbf{R} \leftarrow \{\}$
**for** $j \in \{1, 2, 3, ..., K\}$ **do**
  $\mathcal{P}_k \leftarrow \{\}$ **for** $i \in \{1, 2, 3, ..., N\}$ **do**
    $\mathbf{x}'_j \leftarrow sample\_around(\mathbf{x}')$
    $\mathcal{P}_j \leftarrow \mathcal{P}_k \cup \langle \mathbf{x}'_j, f(\mathbf{x}'_j)_k, \pi_x(\mathbf{x}'_j) \rangle$
  **end**
  $\boldsymbol{\alpha} \leftarrow h(\mathcal{P}_k, \pi_x)$
  $r_i \leftarrow get\_ranking(\boldsymbol{\alpha})$
  $\mathbf{R} \leftarrow append(r_i, \mathbf{R})$
**end**
**return R**

The BLIME algorithm is as follows. From every surrogate model, we obtain a coefficient vector $\boldsymbol{\alpha}$. Then, for every coefficient vector, we obtain a ranking $\boldsymbol{r}$, ordering coefficients from smallest to largest in value. In this manner, and continuing with the image classification example, if the procedure is repeated $K$ times, for a total of $M$ superpixels, we can compactly represent these ranking vectors as rows of a ranking matrix $\mathbf{R} \in \mathbb{R}^{K \times M}$. $\mathbf{R}$ is then interpreted as a **rating scheme, where $M$ superpixels are being rated by $K$ surrogates.** The procedure of deriving $\mathbf{R}$ is described in Algorithm 1.

## 4.2 Ordinal Metrics

The reduction of surrogate coefficients to a ranking can be regarded as a normalisation step that makes multiple surrogates comparable. Although this normalisation removes information about the surrogate coefficient scaling, it allows for the use of ordinal statistics. In this way, we can quantify the consensus amongst the surrogates to gain further insights for a given explanation. We discuss the metrics below. These are widely used in machine learning as proxys to the underlying uncertainty. It shall be noted that the user of BLIME is not limited to these metrics as many measures of consensus have been proposed in the literature [7, 13].

**Mean rank** By comparing the mean rank $\overline{r}_j$ of a superpixel $s_j$ to those of all the other superpixels, indicating the relative importance of each of them.

**Ordinal consensus** The ordinal consensus $C_s$ of the ranking of a superpixel $s$, as defined in [20], can be used to evaluate whether there is high agreement among the raters, indicated by $C_s$ being close to 1, whereas $C_s$ closer to 0.5 indicates no agreement. Values of $C_s$ closer to 0 suggest a high polarisation among raters. The ordinal consensus is calculated by first normalising the occurrences $f_j$ of rankings $r_{s,k}$ assigned to a subject $s$ by raters $k$ (with $k = 1, ..., n$). These normalised occurences are then structured as a list of the form $[\frac{f_1}{n}, ..., \frac{f_N}{n}]$ ordered from the lowest to the highest value, whereas $N$ represents the highest rank that can be assigned to a subject. With the cumulative frequencies $F_s$ defined as $F_i = \sum_{j \leq i} f_j$, the differences $d_i$ are calculated using

$$d_i = \begin{cases} F_i, & \text{if } F_j \leq 0.5 \\ 1 - F_i, & \text{otherwise} \end{cases}$$

The ordinal Consensus $C_s$ is then given as $C_s = 1 - \frac{2}{M-1} \sum_{i=1}^{N} d_i$.

**Inter-rater reliability measures** Here we consider reliability measures to address the overall uncertainty among all surrogates regarding all interpretable components, namely `Fleiss' Kappa` $\kappa$ [9] and `Kendall's coefficient of concordance` $W$ [16]. While $W$ measures the agreement among raters specifically for rankings, $\kappa$ estimates the agreement regardless of the similarity of the assigned ranks. Both measures are scaled from 0 and 1, whereas 0 indicates no agreement and 1 means full agreement among the raters.

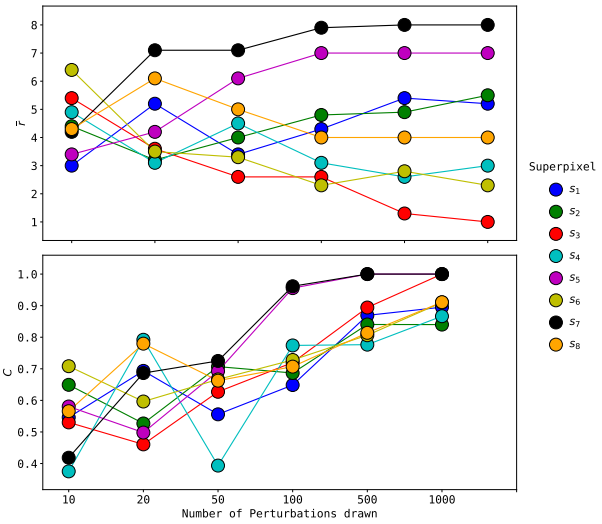
**Kendall's coefficient of concordance** For $N$ subjects (indexed using $s = 1, 2, ..., N$) being rated by $n$ raters (indexed using $k = 1, 2, ..., n$), the total rank $R_s$ is defined as $R_s = \sum_{k=1}^{n} r_{s,k}$, whereas $r_{s,k}$ denotes the rank assigned to subject $s$ by rater $k$. With the mean total rank $\overline{R}$ , Kendall's coefficient is $W = \frac{12 \cdot \sum_{s=1}^{M} (R_s - \overline{R})}{n^2 \cdot (N^3 - N)}$.

**Fleiss' Kappa** For $N$ subjects (index by $s = 1, 2, ..., N$) being assigned to one out of $k$ categories (indexed by $j = 1, 2, ..., k$) by $n$ raters, the quantity $p_j = \frac{1}{N \cdot n} \sum_{s=1}^{N} n_{s,j}$ is defined whereas $n_{s,j}$ denotes the number of raters binning a subject $s$ into category $j$. Fleiss' Kappa estimates the agreement $P_s$ for a subject among all raters as $P_s = \frac{1}{n(n-1)} \sum_{j=1}^{k} (n_{s,j}^2 - n_{s,j})$. With the mean agreement $\overline{P}$ Fleiss Kappa $\kappa$ is $\kappa = \frac{\overline{P} - \sum_{j=1}^{k} p_j^2}{1 - \sum_{j=1}^{k} p_j^2}$.

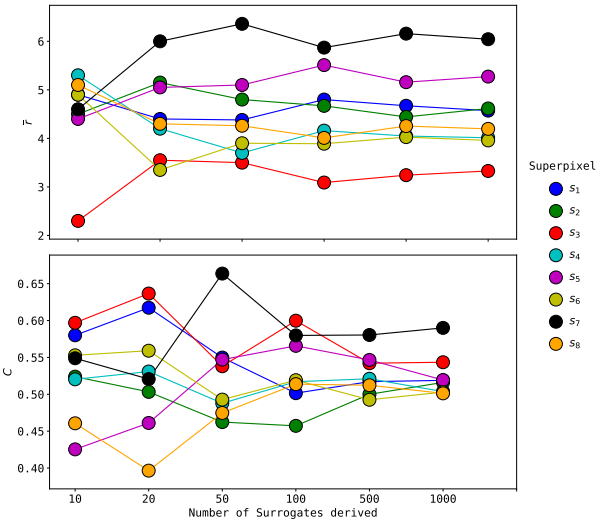
# 5 Experiments

We showcase the approach for both image and text examples.

**Datasets** For the image classification task we use the CIFAR-10 dataset [17] For this work, the data set is split into a training set and a validation set




(a) Influence of the number of perturbations



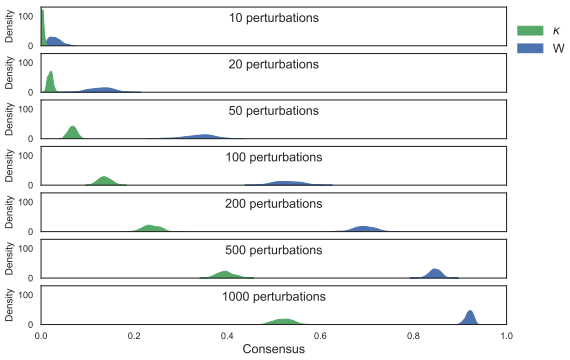

(b) Influence of the number of surrogates

Figure 2: (a) 200 surrogates are derived on bootstrapped perturbations sets. For the superpixels corresponding to the example depicted in figure 4 (top row) mean rank $\overline{r}$ and ordinal consensus $C$ are plotted against the number of perturbations drawn for each surrogate. (b) 200 surrogates are derived on fixed perturbations sets of 200 samples. For the superpixels corresponding to the example depicted in Figure 4 (top row) mean rank $\overline{r}$ and ordinal consensus $C$ are plotted against the number of surrogates derived from the perturbation dataset.

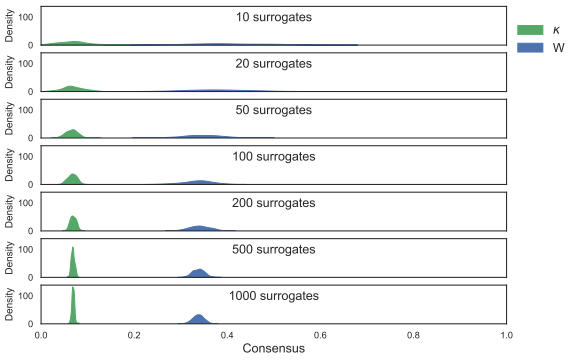
with 50000 and 10000 images respectively. For sentiment classification, we use the movie review dataset IMDB[21]. The task is to classify a textual movie review as positive or negative based. The dataset consists of 50000 labelled reviews.

**Models** For our black-box image classifier, we use an ensemble of 5 CNNs with ResNet architecture as in [14]. The ensemble is created by training all CNNs individually, using random weight initialisation [11] and data shuffling during training to induce diversity [18]. For the text analysis we use an ensemble of fully connected neural networks in combination with GloVe embeddings [22], using random weight initialising and data shuffling during training to induce diversity among the ensemble members. The default configuration of LIME is used with linear regression as surrogates [25].

**Results** In Fig. 2a we see that by increasing the number of perturbations for each bootstrap sample, the mean ranking of the superpixels converges towards values on the full ranking interval 1 to 8, whereas for small numbers of perturbations, the mean ranks are squashed into a rather small interval (top plot). The bottom plot shows that the level of agreement $C$ of the ranking of superpixels increases as the number of image perturbations is also increased. This is expected as the surrogates are trained on datasets that are more similar between them. Therefore, the explanation derived by aggregating multiple surrogates on more perturbations can be considered more certain with regards to the individual superpixels ranking. Examining both plots depicted in Fig. 2a, the highest agreement for the highest and lowest-ranked superpixels ($s_7$ and $s_3$) among the surrogates is maximised. In Figure 2b, the same experiment is run for different numbers of bootstrap samples for a fixed number of perturbations. We see that increasing the number of surrogates does not increase the agreement of the raters assigning ranks to the superpixels (bottom plot). Contrary to Figure 2a, the ranks do not converge to their absolute ranking. The agreement measured by the consensus estimated $C$, however, changes, showing an increasing or decreasing trend. In Figs. 3a and 3b we examine the effects of the number of perturbations drawn and the number of surrogates derived on the uncertainty estimates $\kappa$ and $W$. As shown in Fig. 3a, increasing the number of perturbations drawn shifts the distributions of consensus estimates towards higher values which matches the findings from Fig. 2a. Fig. 3b, however, indicates that the distribution of derived uncertainty estimates become more narrow, resulting in more reliable estimates. Figures 4 and 5 show examples where our method provides additional information to the practitioner about the explanation. In Fig. 4 our method allows the user to




(a) Influence of the number of perturbations




(b) Influence of the number of surrogates

Figure 3: (a) 100 surrogates using bootstrapping on the perturbation dataset $\mathcal{P}$. We report the uncertainty estimates using the consensus $\kappa$ and $W$. The procedure is repeated 100 times with varying numbers of perturbations. (b) 100 sets $\mathcal{P}$ are drawn to fit the surrogates. We report the consensus estimates $\kappa$ and $W$. The procedure is repeated 100 times with varying numbers of surrogates.

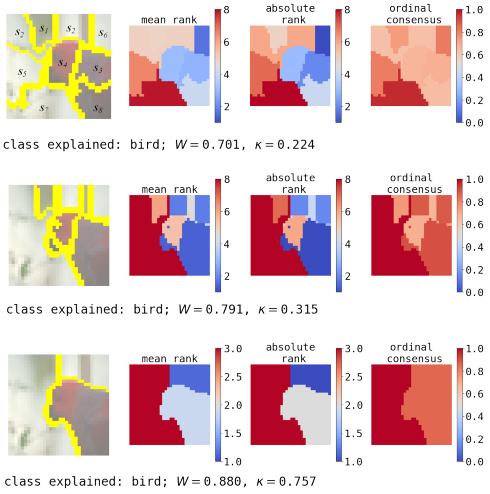



Figure 4: Left: original image $\mathbf{x}$, centre-left: mean ranks of superpixels, centre-right: absolute ranking, right: ordinal consensus. 100 perturbation sets $\mathcal{P}_k$ drawn, 100 data points each to derive surrogates for the predicted class *bird*.

compare different image segmentations for training surrogates. The mean ranks of the superpixels are shown in the second column. The absolute order of rankings according to the mean ranks of superpixels are depicted in the third column. The image in the rightmost column shows the level of agreement amongst the surrogates regarding the ranking of the individual superpixels measured using the ordinal consensus $C$. The original image is segmented into 8 superpixels using different segmentation algorithms. Having access to the uncertainty of the explanation estimated through the values of $W$ and $\kappa$, the practitioner can use this information to compare different image segmentations. The practitioner can conclude that the segmentation depicted in the center row results in an overall more certain explanation, as $W$ and $\kappa$ are higher compared to the segmentation shown the top row. The bottom row of Fig. 4 shows a segmentation of the image segmented into only three superpixels, resulting in a more certain explanation than achieved by the segmentations depicted in the top and center row. The example depicted in Fig. 5 shows our method on a text dataset (IMDB). Here, we highlight how the higher variance of surrogate coefficients is also shown by the ordinal consensus $C$.

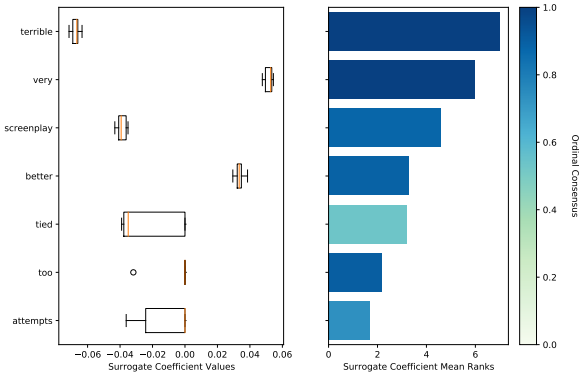



Figure 5: 100 local explanations for a fixed data point using ensemble of 5 fully-connected NNs, 2 hidden layers, IMBD dataset. The mean rank $\overline{r}$ indicates the feature importance.

# 6 Conclusion

In this paper we make the case for the importance of reporting an uncertainty estimated of an explanation together with the explanation when explaining a prediction. This provides the user with the option of rejecting an explanation for being too uncertain. Here, we proposed a procedure where we bootstrap LIME, and then aggregate the outputs using ordinal statistics to measure its uncertainty.

# References

[1] D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[2] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*, 2020.

[3] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[4] U. Bhatt, P. Ravikumar, and J. M. Moura. Towards aggregating weighted feature attributions. *arXiv preprint arXiv:1901.10040*, 2019.

[5] U. Bhatt, P. Ravikumar, et al. Building human-machine trust via interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9919–9920, 2019. URL https://doi.org/10.1609/aaai.v33i01.33019919.

[6] U. Bhatt, A. Weller, and J. M. Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

[7] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[8] B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992. doi:10.1214/aos/1176344552.

[9] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973. doi:https://doi.org/10.1177/001316447303300309.

[10] D. Garreau and U. von Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 1287–1296. PMLR, Aug. 2020. URL http://proceedings.mlr.press/v108/garreau20a.html.

[11] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlrp9.html#GlorotB10.

[12] A. Gosiewska and P. Biecek. ibreakdown: Uncertainty of model explanations for non-additive predictive models. *arXiv preprint arXiv:1903.11420*, 2019.

[13] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

[15] A. Hepburn and R. Santos-Rodriguez. Explainers in the wild: Making surrogate explainers robust to distortions through perception. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3717–3721, 2021. doi:10.1109/ICIP42928.2021.9506711.

[16] M. G. Kendall and B. B. Smith. The problem of m rankings. *The annals of mathematical statistics*, 10(3):275–287, 1939. doi:10.1214/aoms/1177732186.

[17] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[19] E. Lee, D. Braines, M. Stiffler, A. Hudler, and D. Harborne. Developing the sensitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006. International Society for Optics and Photonics, 2019. doi:https://doi.org/10.1117/12.2520149.

[20] R. K. Leik. A measure of ordinal consensus. *Pacific Sociological Review*, 9(2):85–90, 1966. doi:https://doi.org/10.2307/1388242.

[21] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

[22] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. doi:10.3115/v1/D14-1162.

[23] R. Poyiadzi, X. Renard, T. Laugel, R. Santos-Rodriguez, and M. Detyniecki. On the overlooked issue of defining explanation objectives for local-surrogate explainers. *arXiv preprint arXiv:2106.05810*, 2021.

[24] R. Poyiadzi, X. Renard, T. Laugel, R. Santos-Rodriguez, and M. Detyniecki. Understanding surrogate explanations: the interplay between complexity, fidelity and coverage. *arXiv preprint arXiv:2107.04309*, 2021.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. doi:10.18653/v1/N16-3020.

[26] L. Rieger and L. K. Hansen. Aggregating explanation methods for stable and robust explainability. *arXiv preprint arXiv:1903.00519*, 2019.

[27] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *arXiv preprint arXiv:2008.05030*, 2020.

[28] K. Sokol, A. Hepburn, R. Santos-Rodriguez, and P. Flach. blimey: surrogate prediction explanations beyond lime. *arXiv preprint arXiv:1910.13016*, 2019.

[29] K. Sokol, A. Hepburn, R. Poyiadzi, M. Clifford, R. Santos-Rodriguez, and P. Flach. Fat forensics: A python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *Journal of Open Source Software*, 5(49):1904, 2020. doi:10.21105/joss.01904. URL `https://doi.org/10.21105/joss.01904`.

[30] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical image analysis*, 60, 2020. doi:10.1016/j.media.2019.101619.

[31] K. K. Wickstrøm, K. ØyvindMikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen. Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series. *IEEE Journal of Biomedical and Health Informatics*, 2020. doi:https://doi.org/10.1109/jbhi.2020.3042637.

[32] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019.

[33] Z. Zhou, G. Hooker, and F. Wang. S-lime: Stabilized-lime for model explanation. *arXiv preprint arXiv:2106.07875*, 2021.