Unsupervised DeepView: Global Explainability of Uncertainties for High Dimensional Data

1st Carina Newen
Research Center Trustworthy Data Science and Security
TU Dortmund
carina.newen@cs.tu-dortmund.de

2nd Prof. Dr. Emmanuel Müller
Research Center Trustworthy Data Science and Security
TU Dortmund
emmanuel.mueller@cs.tu-dortmund.de

***Abstract*—In recent years, more and more visualization methods for explanations of artificial intelligence have been proposed that focus on untangling black box models for single instances of the data set. While the focus often lies on supervised learning algorithms, the study of uncertainty estimations in the unsupervised domain for high-dimensional data sets in the explainability domain has been neglected so far. As a result, existing visualization methods struggle to visualize global uncertainty patterns over whole datasets.**

We propose Unsupervised DeepView, the first global uncertainty visualization method for high dimensional data based on a novel unsupervised proxy for local uncertainties. In this paper, we exploit the mathematical notion of local intrinsic dimensionality as a measure of local data complexity. As a label-agnostic measure of model uncertainty in unsupervised machine learning, it shows two highly desirable features: It can be used for global structure visualization as well as for the detection of local adversarials. In our empirical evaluation, we demonstrate its ability both in visualizations and quantitative analysis for unsupervised models on multiple datasets.

***Index Terms*—Visualization; Unsupervised Learning; Uncertainty Quantification; Adversarials**

I. Introduction

Visualization of raw data, pre-trained models, and uncertainties of these models are essential methods for explaining machine learning models. While in the area of supervised models, such visualizations of a classification model and its uncertainties are widely studied [14], [16], [23], this has only been tackled less extensively for unsupervised learning methods [21]. However, it has been neglected entirely for the combination of unsupervised models such as clustering [1] or anomaly detection [2] and global explainability. This is due to the fact that unsupervised models do not have labeled training data that allow us to directly evaluate misclassifications or model errors. Similarly, uncertainties are challenging to quantify for a model without known (labeled) patterns. Furthermore, it is easier to depict uncertainties for specific data points rather than reliably approximate uncertainties for the entire model.

Existing visualization methods for high dimensional data [4], [13], [34] rely on labeled data and are not able to visualize the hidden (unknown) patterns of data in case of unsupervised learning. The key challenge is the lack of an objective function that, in the case of supervised learning, describes the separability of one or multiple classes, as depicted in Figure 1 on the left. In contrast to this, unsupervised measures such as cluster-validation [5], clusterability scores [7], or anomaly scores [8] are designed either to evaluate the quality of clustering (globally for the entire dataset) or the detection of rare and exceptional outliers (locally for individual objects). In contrast to unsupervised measures, we aim at both a global visualization of raw data structures and a local uncertainty of specific data areas. Neither the one nor the other should rely on labels but are purely based on the given pre-trained unsupervised model and a proxy for local uncertainty quantification.

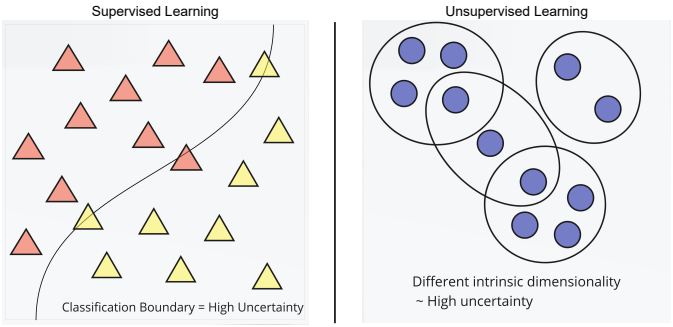

Fig. 1. This Figure shows the difference between uncertainties for supervised models versus uncertainties in the unsupervised domain. Data areas along the classification boundary are most likely misclassified in the supervised domain. In the unsupervised domain, the uncertainty is not as clear. Here, we define our uncertainties by measuring the local intrinsic dimensionality of the data. When this dimensionality varies considerably from its neighbors, we define this as an uncertain area for our visualization scheme.

Our research focuses on unsupervised knowledge discovery and defines novel proxies for uncertainties. In this paper, we exploit the mathematical notion of local intrinsic dimensionality as a proxy for the complexity of the data distribution. As depicted in Figure 1 on the right, we define local divergence in intrinsic dimensionality as the label-agnostic measure of high model uncertainty in unsupervised machine learning. It allows for both local assessments of uncertainty for a specific data

area and, at the same time, for global visualization of data structures vs. adversarial examples in less certain data areas. We have implemented the first unsupervised DeepView approach that visualizes data structures and adversarials without any prior labels. It requires only a pre-trained unsupervised model and its uncertainties. This framework calculates the local intrinsic dimensionality of the data points and creates a mapping of uncertainties using dimensionality reduction [15] on the combination of the data, their prediction uncertainties, and their dimensionality. We then use outlier detection algorithms on this data embedding to depict our uncertainties with a high empirical precision score. Our paper solves the task of global and local uncertainty visualization for unsupervised learning. In contrast, all other known methods either solve global supervised explainability [13], [23], [24] or focus on local explanations [16], [20], [25] with few unsupervised methods [21] only.

II. Related Work

Visualization and Explanation of Machine Learning. In the past, we have seen several distinguished visualization techniques that explain the quality of a model, such as LIME [16], SHAPex [20], Anchor [25], and LIME-based extensions such as EXPLAIN-IT [21] for unsupervised learning. In addition, other explainability methods such as LEMNA [22] extend explainability to Recurrent Neural Networks or Multilayer Perceptrons that are specifically needed for security-critical contexts. Furthermore, countless other approaches, such as causal explanations for supervised learning algorithms, exist [9]. All of these methods focus on single decisions or local explanations while neglecting the global properties of the underlying machine learning model.

DeepView [13] was the first attempt to visualize global decision boundaries of a deep neural network, in contrast to previous methods that explain single examples or their classification features. Other methods produce a projection of the data and generate global explainability; however, they either do not depict decision boundaries [23] or are not applicable to deep neural networks [24] due to their density estimation approach in the input space. Furthermore, these methods cannot detect uncertainties by themselves as they rely on a human evaluation of the algorithm. In contrast, we aim at a completely unsupervised approach that lets the algorithm itself highlight where the uncertain regions are. We enable an intuitive uncertainty estimation using the certainty of the unreliability in our process rather than the certainty of the class label, which is used by supervised approaches, as the background color to depict the robustness of the model in the form of adversarial examples.

Subspace cluster visualization This visualization method essentially distributes the data into two classes, clustering uncertain and certain data points in high dimensional space. Prominent cluster visualization methods include Clustnail [36], the Heidi matrix [37] and Ferdosi's astronomical data subspace clustering algorithm [38]. The prominent difference in contribution between these clustering techniques and our methodology is that we enable the visualization of uncertainties independent of the clustering algorithm used for the method. Furthermore, methods like Ferdosi's subspace clustering provide no direct way of comparing subspaces, whereas UMAP [15], which we use for dimensionality reduction purposes, promises preservations of local and global distances. Instead of employing UMAP [15], any of these clustering techniques could be included in the visualization tool. However, for example, the Heidi matrix's [37] abstract visual mapping would lead to a loss of interpretability of the image produced. VISA [39] provides both a global overview as well as an in-depth-view, but the distance between the clusters might be obscured by the radius of the circles in the visualization scheme, causing the visualization to potentially look cluttered. However, the less complex 2d representation chosen in this paper leads to a more understandable and intuitive graphic, which allows even a less technical audience to understand whether a model is considered less or more uncertain by the visualization tool.

Uncertainty Quantification for Machine Learning. Uncertainty estimates include LUX [26], which is based on decision trees for their local estimates and CLUE [14] and its extensions [27] [28]. LUX [26] assumes the supervised knowledge of the class labels for useful interpretation of their generated estimates. It focuses on local explanations of the model only, while CLUE [14] looks at the smallest change to the input in the latent space to change the model's certainty. Therefore it focuses on local rather than global explanations. There have also been attempts to model uncertainty for existing explainability methods, such as variations of LIME [21] or SHAP [29]. However, in contrast to our work, they are only limited to local explanations and uncertainties. We shift our research focus to global explanation methods that allow a user to choose the globally more reliable model rather than only understand the inner black box of the model using single explanation examples.

Dimensionality Reduction for Machine Learning. As most of the previously mentioned methods work on high dimensional data, we observe dimensionality reduction as an essential related technology. Principal Component Analysis (PCA) [10] and Linear Discriminant Analysis (LDA) [4] are state-of-the-art methods for unsupervised and supervised dimensionality reduction. While t-Distributed Stochastic Neighbor Embedding (t-SNE) [34] and UMAP [15] calculate similarity scores between pairs of data instances in a high dimensional space and try to find an optimal mapping in a lower-dimensional space that sufficiently represents the data points. Although heavily used for raw data visualization, none of these methods can optimize for uncertainties in pre-trained models. As pre-processing steps, they neglect the later model, while in our case, we want to optimize for both. We include local properties of the raw data and an unsupervised model's uncertainty in our visualization optimization.

III. Visualizing the Uncertainties of an Unsupervised Learner

In this work, we consider uncertainties w.r.t. (A) properties of the raw data distribution as well as (B) properties of unsupervised machine learning models that have been pre-trained on these data. We define uncertainties as more general than misclassification of supervised models. In our definition, uncertainties are caused by the complexity of data, e.g., uncertain areas following the empty space phenomenon in high dimensional data [12]. Similarly, based on a pre-trained model, uncertainties may be areas of high probability for adversarial examples. However, in our definition, the adversarial examples are not depicted by wrong class labels. We observe unsupervised adversarial examples in the training of unsupervised models (e.g., adversarial examples in autoencoders [3]).

Hence our methods tackle two main challenges:
First, our unsupervised approach's labels and amount of classes are unknown. We cannot assess the traditional discrepancies between what a model predicts as uncertain and what really is uncertain or incorrect. Without supervision, we are forced to depict uncertainties using novel unsupervised measures and propose a local comparison of this measure with the object's local neighborhood.
Second, existing supervised methods such as DeepView [13] rely on prediction probabilities for the classes, which are then used to visualize the uncertainty of the class prediction. In our case, we assume the knowledge of those prediction uncertainties for our model and use them as input into the visualization method. Formally, we describe our abstract notion of unsupervised uncertainties as follows:

Definition 1
(Unsupervised Quantifiable Uncertainty)

$$UQU(x_i) = p_{adv}(x_i) + p_{model}(x_i)$$

with $p_{adv}(x_i)$ the likelihood of sample x_i being an adversarial and $p_{model}(x_i)$ the likelihood of sample x_i being out-of-distribution for the particular model, suggesting a misclassification.

A key contribution of our work is that we approximate $p_{adv}(x_i)$ by our outlier detection algorithm and its local intrinsic dimensionality measure $p_{lid}(x_i)$. The proxy $p_{lid}(x_i)$ describes the likelihood of the data point being an adversarial because of its Local Intrinsic Dimensionality (LID). In a supervised setting, we would describe $p_{adv}(x_i) = p_{missclass}(x_i|\text{high certainty}(x_i))$. An adversarial example is commonly described in literature as an optimization problem for a classifier $f : \mathbb{R}^m \rightarrow \{1...k\}$, which maps image pixels to a discrete label set [35]. f has a continuous loss function denoted by $loss_f : \mathbb{R}^m \times \{1...k\} \rightarrow \mathbb{R}^+$. An adversarial example solves the following optimization problem for an input $x \in \mathbb{R}^m$ and label $l \in \{1...k\}$:

Minimize $||r||_2$ subject to

1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

So intuitively, the most common definition of an adversarial example is the smallest perturbation to an image that leads to a misclassification of a classifier. In practice, it is very hard to determine whether an image is simply a slight perturbation from another image of a dataset, especially if we generate an adversarial image from an image that is not part of the training dataset. Mostly, adversarial examples differ from normal misclassifications due to their high certainty in the correctness of the prediction, which is why we use this property to determine $p_{adv}(x_i)$. In practice, this means we fail to distinguish between simple misclassifications that happen to have high certainty and real adversarials in the form of input perturbation attacks, but since we aim to approximate the uncertainty of a model as an end goal, this explicit distinction is not necessary. Both are very dangerous for real-world applications and capture the essence of an uncertainty, especially for high-risk scenarios such as autonomous driving.

On the other hand, $p_{model}(x_i)$ is approximated by the certainty of the model itself on the given data points. For our empirical evaluation, we simply compare whether our unsupervised approximation derived from our model is in line with the supervised misclassification (using labels). In real-world use cases, this would not be possible; however, our datasets have given labels that are ignored by our algorithm but useful for external evaluation.

A. Unsupervised DeepView

We propose a generalized visualization technique to depict uncertainties and potential adversarial examples in unsupervised learning. Our framework consists of three components:

1) Local Intrinsic Dimensionality (cf. Section III-C) as our main uncertainty proxy of the data distribution.
2) Dimensionality Reduction (cf. Section III-D) that allows for 2D visualization of high dimensional data.
3) Adversarial Detection (cf. Section III-E) based on an outlier analysis of our uncertainty measure.

In the unsupervised framework, all three of these components can be exchanged in future work. In the following subsections, we present our first instantiation of each of these into the DeepView visualization [13]. This allows for the first time a fully unsupervised analysis and visualization of uncertainties.

Please note that our method is not just showing the quality of the decision boundary in a supervised learner. We extract and measure uncertainties based on unlabeled data. To the best of our knowledge, this is the first uncertainty visualization method that focuses on global interpretability rather than explanations of individual predictions in the unsupervised learning domain that can visualize a smooth two-dimensional manifold of the uncertainties on high-dimensional data such as natural images.

B. Algorithm Overview

Our estimation and visualization algorithm needs to be applicable for any unsupervised task or usable whenever the probabilities of an unsupervised model are given on a randomly sampled data set. As a result of our algorithm, we want to depict uncertain areas with unsupervised adversarial examples and a visualization of certain areas and local uncertainty. To achieve this, we propose the following algorithmic steps:

1) Calculate the local intrinsic dimensionality (LID, cf. Section III-C) of each data point $\forall x_i \in S$ in the sample S compute $LID(x_i)$.
2) Apply dimensionality reduction (UMAP, cf. Section III-D) on the given input of LIDs together with unsupervised model uncertainty. We project these three values appended together as x_i to two dimensions, yielding $y_i = \pi(x_i)$.
3) Create a tight grid of samples r_i in two-dimensional space and map this to high-dimensional space.
4) Outlier detection (cf. Section III-E) algorithm on the uncertainty measures, resulting in binary detection of uncertain and certain areas. We define the uncertain area by a high LID mean.
5) Visualize the outlier scores and interpret them as unsupervised uncertainties.

In contrast to our unsupervised algorithm, the supervised DeepView method [13] is composed of four algorithmic steps which enable visualization of decision boundaries:

1) Application of the dimensionality reduction technique Fisher UMAP [15] to project data points x_i to two dimensions, yielding $y_i = \pi(x_i)$.
2) Creation of a tight grid of samples r_i in two-dimensional space and mapping to high-dimensional space.
3) Application of the network to this mapping to obtain predictions and certainties.
4) Visualization of the labels together with the entropies of the certainties.

Ma et al. [18] already showed that local intrinsic dimensionality measures help characterize adversarial subspaces. While there has also been research on the limitations of these features for the characterization of adversarial examples [30], one of the main critique points was the non-transferability of the LID features to other deep neural networks. This point does not affect the quality of our visualization method, considering it is used on a specific model and does not require transferability to other models. The second criticism expressed in this paper was that the quality of LIDs as features for adversarial attacks with the trained detector algorithm suggested by Ma et al. [18] varies with the confidence parameter, and training of ensembles of adversarials with different confidence levels did not help the detection performance. However, our approach does not solely rely on LIDs. Instead, it uses the dimensionality reduction technique based on UMAP [15] suggested by the DeepView algorithm [13] and searches for outliers on this mapping of the data points LIDs and prediction certainties. Note that the exact outlier detection algorithm is exchangeable in the framework for any method that performs best for your dataset. This means that we take into account the uncertainties as well as the LIDs, giving us a better chance of extracting adversarial or uncertain regions. After all, we do not wish just to identify adversarial examples. Furthermore, there is now no other implementation of a global visualization of the decision boundary of uncertainty in the unsupervised domain. We later show that our approach gives us good approximations of the uncertainties within an unsupervised learner.

C. Local Intrinsic Dimensionality

The intuition behind this metric is to measure the increase of data objects encountered, estimating the dimensionality of the structure of the data. Transferring the idea of expansion of dimensions to distance distributions gives a formal definition of LID [19].

Definition 2 (LOCAL INTRINSIC DIMENSIONALITY)
Given a data sample $x \in X$, let $R > 0$ be a random variable denoting the distance from x to other data samples. If the cumulative distribution function $F(r)$ of R is positive and continuously differentiable at distance $r > 0$, the LID of x at distance r is given by:

$$LID_F(r) \hat{=} \lim_{\epsilon \to 0} \frac{\ln(F((1+\epsilon) * r)/F(r))}{\ln(1+\epsilon)} \tag{1}$$

whenever the limit exists. The maximum likelihood estimator (MLE) of the LID at x given a reference sample drawn from the representation of the data distribution P is defined as follows:

$$\hat{LID}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)}\right)^{-1} \tag{2}$$

Where $r_i(x)$ denotes the distance between x and its $i-$th nearest neighbor within a sample of points drawn from P, and $r_k(x)$ is the maximum of the neighbor distances. While this computation can become computationally expensive with the increase of the neighborhood, Ma et al. [19] showed that discrimination of adversarial and non-adversarial examples turn out to be possible for minibatch sizes of 100 and neighborhood sizes as small as 20, rendering our computational estimation feasible. Therefore, these are the parameters we also use to implement our visualization scheme.

D. Dimensionality Reduction

The goal of dimensionality reduction techniques for visualizations is to find mappings $\pi : (S, d_s) \to \mathbb{R}, d = 2, 3$, where (S, d_s) is a metric space and π ideally preserves the information encoded in a set of data points $x_1, ..., x_n \in S$. The paper is based on a dimension reduction technique called UMAP [15], which performs at least equally well as the state-of-the-art non-linear dimensionality reduction method t-Distributed Stochastic Neighbor Embedding (t-SNE) [34], but allows for the inverse projection suggested by the original DeepView implementation. It allows for our two-dimensional

visualization. Theoretically, this method could be exchangeable as long as inverse mappings can still be established with other algorithms.

E. Adversarial detection

Please note that we can not simply detect unsupervised adversarial examples by applying clustering algorithms because of the imbalance in the data sets. For a good model, we do not find near as many uncertainties as certain data points. After all, we actually look for outliers whose LIDs and prediction scores do not match the distribution of the other data points.

As Ma et al. state in their paper [19], adversarial examples have noticeably higher LID characteristics than normal examples. This means that we can distinguish adversarials using outlier detection algorithms. We follow the intuitive definition of an outlier as given by Hawkins [31] and search for outliers that deviate so much from the other observations as to arouse suspicion that they are generated by a different mechanism.

We detect outliers based on density-based clustering algorithms that allow us to distinguish certain and uncertain areas but also consider that adversarials or uncertainties have to be detected for both single-point outliers and cluster-based outliers. We refer to the definition of cluster-based outliers of LDBSCAN [32] and its extension HDBSCAN [33]. Compared to other clustering-based outlier detection methods, the advantage of these two algorithms is the quantifiable outlier score, which intuitively corresponds to the degree of the outlying object. In the case of HDBSCAN, it is referred to as an outlier score. These scores can be incorporated into our visualization of the decision boundary as the certainty of the outlier. The difference between HDBSCAN and DBSCAN is that HDBSCAN performs a hyperparameter search of the ϵ parameter, namely the radius from at least one cluster point to another, of LDBSCAN without having to preset it, therefore only needing a minimum cluster size as an input parameter. This is why our final implementation uses the HDBSCAN algorithm without comparing the results from LDBSCAN.

F. Runtime Analysis

The runtime of our algorithm, including the three mentioned steps, is similar to the supervised DeepView implementation. However, the extra effort in calculating the LIDs once for all samples in the dataset, approximating it with a neighborhood size of 20 each, and then computing dimensionality reduction and adversarial examples is neglectable compared to the heavy optimization of supervised DeepView. Preliminary experiments have shown very similar runtimes and neglectable differences. Hence, we focus the following on qualitative and quantitative evaluation without deepening the topic of runtime evaluation.

IV. Experiments

In this section, we apply the new Unsupervised DeepView implementation[1] and evaluate how well we capture uncertainties. We measured uncertainties as either supervised misclassifications or adversarial examples and showed exemplary applications of our method to CIFAR-10 and Fashion-MNIST. We ensure not to test on both models used for evaluating DeepView. Still, the model trained on the Fashion-MNIST dataset is a separately trained network relying on the resnet50 convolutional architecture with some additional convolutions and a linear layer. As you can see later in Table I, our algorithm also performs excellently for this model and dataset with very high precision scores over several runs.

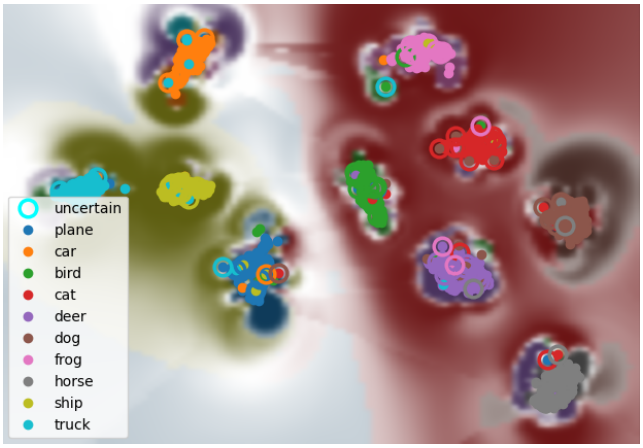

Fig. 2. The original DeepView implementation achieves a two-dimensional visualization of the decision boundary of a supervised classifier. For evaluation purposes, we embedded our as uncertainties recognized points that we calculated without using the known labels into the original DeepView implementation. The blue circles denote points that we recognized as uncertain. As we can see, we were perfectly able to distinguish an adversarial example labeled as a bird, even though it was actually a frog (green isolated dot near the pink cluster), also visualized in Figure 3 from the CIFAR-10 data set. In the same picture, we also see the edges of the truck cluster marked, where a misclassified car and plane are nearly hidden (bright blue cluster with dark blue and orange spots). The other point marked as uncertain is a correctly labeled plane with low uncertainty values.

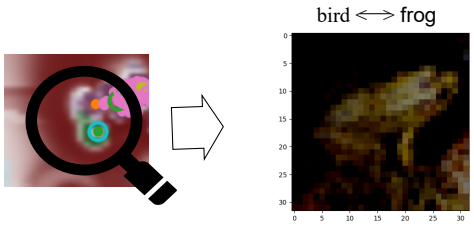

Fig. 3. When clicking on the point instance given as uncertain, the original instance of the image will be shown. On the left-hand side, the model predicted label is shown, and on the right, the actual label is given. For the Unsupervised algorithm, we will only output certain or uncertain labels, for instance. The goal is to allow the user to directly assess the instances visually so as to better understand why the model classified them as uncertain.

To evaluate our visualization scheme, we look at the following questions: (i) Does our combination of outlier detection algorithms, LID features, and prediction probabilities capture the same or similar uncertainties to the supervised version of the DeepView implementation without using labels? (ii)

[1]The code for the project can be found at our chair's collective repository https://github.com/KDD-OpenSource/Unsupervised-Deepview/.

Are the LID features necessary for our implementation to perform well, or could we also just detect outliers using the uncertainties of the prediction alone? (iii) How well are we able to distinguish model uncertainty or adversarials without the knowledge of the label?

Addressing our first question, we compare the original DeepView visualization technique with labels to our unsupervised DeepView method by marking the points recognized as uncertain in the original DeepView plot in Figure 2. Here, we can see that while not all uncertainties were detected, the points marked as uncertain turned out to be either misclassifications, adversarials, or one point with just low model certainty. We were able to quantify misclassifications because we did not use the labels to generate our uncertainty predictions. Still, we knew the actual labels of the dataset because we evaluated over a labeled data set to test the quality of our model. A zoomed view of the uncertainty found, and its original image point can be found in Figure 3. It is visualized when clicking on the data point in the plot given.

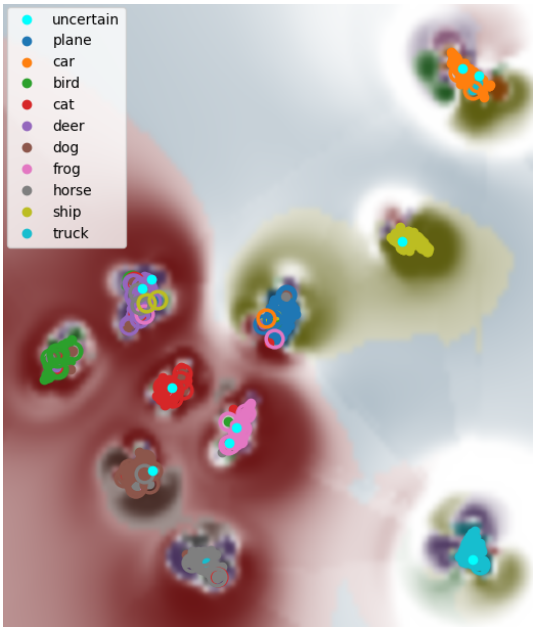

Fig. 4. For evaluation purposes, we also tested whether the LID features were essential for visualization of uncertainty without knowledge of the labels. Here, we just marked the points detected by the HDBSCAN [33] algorithm without the LID features and without knowledge of the labels in the picture generated by the original DeepView algorithm. As we can see, the uncertainties are not captured as well. Often the centers of the class clusters are marked whether they are uncertain or not.

Furthermore, we compare the points detected as outliers using only the data points and prediction uncertainties, without the LIDs as features in Figure 4. This gives us an intuition that the LIDs work as features to compensate for the actual class labels and shows that we can correctly identify actual adversarial examples within the dataset.

Addressing (iii), our precision score is consistently high. While the method does not detect all uncertainties, the uncertainties we do detect are either misclassifications or adversarial examples with the following percentages (cf. Table I).

TABLE I
PRECISION SCORES OVER 10 RUNS

Dataset	Precision
CIFAR-10	0.984
Fashion-MNIST	0.973

The final output of our visualization method can be found in Figure 5. Here, we can see data points denoted as certain and uncertain, as well as our decision boundary colored in by the strength of the blue note. Darker background areas indicate more certain areas in the data for the specific model.

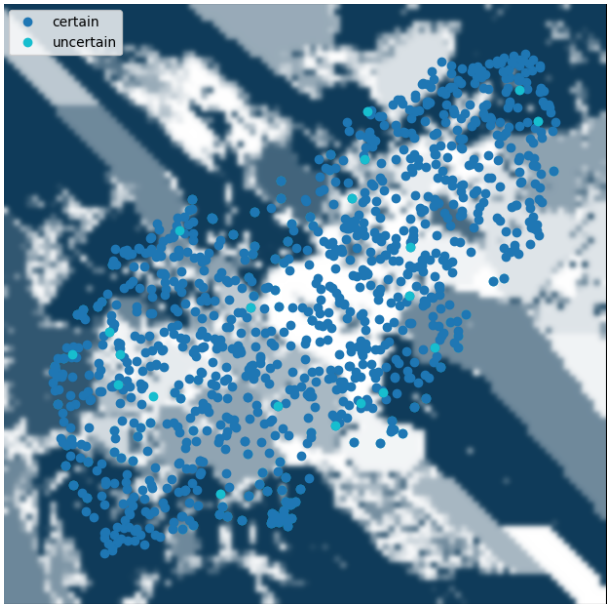

Fig. 5. Our Unsupervised DeepView implementation achieves a two-dimensional visualization of the decision boundary of uncertainties. We do not claim that it identifies all model uncertainties. Still, those it detects are either uncertain in their prediction, misclassifications, or adversarials which we can evaluate with a very high precision score. Furthermore, our visualization method does not aim to identify them directly but rather outputs the areas in which predictions become uncertain or where the classifier performs well. Dark areas denote particularly certain areas, whereas whiter areas are particularly uncertain. Our uncertain data points are colored light blue

V. Conclusion and Future Work

In this paper, we propose *Unsupervised DeepView*, which allows the depiction of a smooth dimensional manifold of uncertainties for high-dimensional data. To the best of our knowledge, it is the first method generally applicable to all unsupervised learning algorithms that provide uncertainties in the unsupervised domain. In contrast to recent methods such as GLAM-CLUE [11] our method does not generate counterfactuals but shows the decision boundary of the uncertainties of the model. We do not require any labeled data as we exploit the mathematical concept of local intrinsic dimensionality as a local proxy. Our global visualization of data structures vs. adversarial examples provides first unsupervised insights into

the uncertainty and vulnerability of machine learning models. It portrays the whole decision boundary rather than the single decisions of the model.

In our empirical evaluation, we present precision for CIFAR-10 and Fashion-MNIST. We evaluate it using the pre-trained residual network with 20 layers discussed in the original paper to show that the algorithms predict similar uncertainties despite the lack of labels. Secondly, we use a different convolutional neural network pre-trained on Fashion-MNIST to confirm our results.

As a first approach in this area, we see future work to extend our method to data beyond high-dimensional vector spaces. For example, graph data that has inherent local and global structures would benefit from our methodology but requires specific graph measures as local proxies of structure and uncertainty. Furthermore, we see improvement potential for the usability of the tool or practicability of trust calibration - Overall, this explainability method is designed not just for a technical audience but should be used to provide a better estimate than just accuracy estimates or other common proxies for the reliability of a model on a specific data set.

Acknowledgements. This work was supported by the Research Center Trustworthy Data Science and Security, an institution of the University Alliance Ruhr.

REFERENCES

[1] Aggarwal, Charu & Reddy, Chandan. (2013). DATA CLUSTERING Algorithms and Applications.

[2] Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." Proceedings of the 2001 ACM SIGMOD international conference on Management of data. 2001.

[3] Benedikt Boing & Rajarshi Roy & Emmanuel Muller & Daniel Neider "Quality Guarantees for Autoencoders via Unsupervised Adversarial Attacks" Proceedings European Conference Machine Learning and Knowledge Discovery in Databases (ECML KDD 2020).

[4] Tharwat, Alaa. (2015). Linear Discriminant Analysis.

[5] Y. Liu, Z. Li, H. Xiong, X. Gao and J. Wu, "Understanding of Internal Clustering Validation Measures," 2010 IEEE International Conference on Data Mining, 2010, pp. 911-916, doi: 10.1109/ICDM.2010.35.

[6] Breunig, Markus & Kriegel, Hans-Peter & Ng, Raymond & Sander, Joerg. (2000). LOF: Identifying Density-Based Local Outliers. ACM Sigmod Record. 29. 93-104. 10.1145/342009.335388.

[7] Ackerman, Margareta, and Shai Ben-David. "Clusterability: A theoretical study." Artificial intelligence and statistics. PMLR, 2009.

[8] Aggarwal, C. C. (2013), Outlier Analysis , Springer .

[9] Schwab, Patrick, and Walter Karlen. "Explain: Causal explanations for model interpretation under uncertainty." Advances in Neural Information Processing Systems 32 (2019).

[10] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.

[11] Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." arXiv preprint arXiv:2112.02646 (2021).

[12] Lee, John A., and Michel Verleysen. Nonlinear dimensionality reduction. Vol. 1. New York: Springer, 2007.

[13] Schulz A, Hinder F, Hammer B. Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. arXiv preprint arXiv:1909.09154. 2019 Sep 19.

[14] Antorán J, Bhatt U, Adel T, Weller A, Hernández-Lobato JM. Getting a clue: A method for explaining uncertainty estimates. arXiv preprint arXiv:2006.06848. 2020 Jun 11.

[15] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018 Feb 9.

[16] Zhang, Yujia, et al. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations." arXiv preprint arXiv:1904.12991 (2019).

[17] Galil, Ido, and Ran El-Yaniv. "Disrupting Deep Uncertainty Estimation Without Harming Accuracy." Advances in Neural Information Processing Systems 34 (2021).

[18] Ma, Xingjun, et al. "Characterizing adversarial subspaces using local intrinsic dimensionality." arXiv preprint arXiv:1801.02613 (2018).

[19] Michael E. Houle. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In SISAP, pp. 64–79, 2017a.

[20] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[21] Morichetta, Andrea, Pedro Casas, and Marco Mellia. "EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis." Proceedings of the 3rd ACM CoNEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks. 2019.

[22] Guo, Wenbo, et al. "Lemna: Explaining deep learning based security applications." proceedings of the 2018 ACM SIGSAC conference on computer and communications security. 2018.

[23] Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." Nature communications 10.1 (2019): 1-8.

[24] Schulz, Alexander, Andrej Gisbrecht, and Barbara Hammer. "Using discriminative dimensionality reduction to visualize classifiers." Neural Processing Letters 42.1 (2015): 27-54.

[25] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.

[26] Bobek, Szymon, and Grzegorz J. Nalepa. "Introducing uncertainty into explainable ai methods." International Conference on Computational Science. Springer, Cham, 2021.

[27] Ley, Dan, Umang Bhatt, and Adrian Weller. "δ-CLUE: Diverse Sets of Explanations for Uncertainty Estimates." arXiv preprint arXiv:2104.06323 (2021).

[28] Ley, Dan, Umang Bhatt, and Adrian Weller. "Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates." arXiv preprint arXiv:2112.02646 (2021).

[29] Slack, Dylan, et al. "Reliable post hoc explanations: Modeling uncertainty in explainability." Advances in Neural Information Processing Systems 34 (2021).

[30] Lu, Pei-Hsuan, Pin-Yu Chen, and Chia-Mu Yu. "On the limitation of local intrinsic dimensionality for characterizing the subspaces of adversarial examples." arXiv preprint arXiv:1803.09638 (2018).

[31] Hawkins, Douglas M. Identification of outliers. Vol. 11. London: Chapman and Hall, 1980.

[32] Duan, Lian, et al. "Cluster-based outlier detection." Annals of Operations Research 168.1 (2009): 151-168.

[33] McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." J. Open Source Softw. 2.11 (2017): 205

[34] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[35] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).

[36] Tatu, Andrada, et al. "Clustnails: Visual analysis of subspace clusters." Tsinghua Science and Technology 17.4 (2012): 419-428.

[37] Vadapalli, Soujanya, and Kamalakar Karlapalem. "Heidi matrix: nearest neighbor driven high dimensional data visualization." Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. 2009.

[38] Ferdosi, Bilkis J., et al. "Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators." 2010 IEEE Symposium on Visual Analytics Science and Technology. IEEE, 2010.

[39] Assent, Ira, et al. "VISA: visual subspace clustering analysis." ACM SIGKDD Explorations Newsletter 9.2 (2007): 5-12.