Bounding open space risk with decoupling autoencoders in open set recognition

Max Lübbering · Michael Gebauer · Rajkumar Ramamurthy · Christian Bauckhage · Rafet Sifa

Received: 18 January 2022 / Accepted: 18 June 2022 / Published online: 16 July 2022
© The Author(s) 2022

Abstract

One-vs-Rest (OVR) classification aims to distinguish a single class of interest (COI) from other classes. The concept of novelty detection and robustness to dataset shift becomes crucial in OVR when the scope of the rest class is extended from the classes observed during training to unseen and possibly unrelated classes, a setting referred to as open set recognition (OSR). In this work, we propose a novel architecture, namely decoupling autoencoder (DAE), which provides a proven upper bound on the open space risk and minimizes open space risk via a dedicated training routine. Our method is benchmarked within three different scenarios, each isolating different aspects of OSR, namely plain classification, outlier detection, and dataset shift. The results conclusively show that DAE achieves robust performance across all three tasks. This level of cross-task robustness is not observed for any of the seven potent baselines from the OSR, OVR, outlier detection, and ensembling domain which, apart from ATA (Lübbering et al., From imbalanced classification to supervised outlier detection problems: adversarially trained auto encoders. In: Artificial neural networks and machine learning—ICANN 2020, 2020), tend to fail on either one of the tasks. Similar to DAE, ATA is based on autoencoders and facilitates the reconstruction error to predict the inlierness of a sample. However unlike DAE, it does not provide any uncertainty scores and therefore lacks rudimentary means of interpretation. Our adversarial robustness and local stability results further support DAE’s superiority in the OSR setting, emphasizing its applicability in safety-critical systems.

Keywords Open set recognition · Open space risk · Outlier detection · Dataset shift

1 Introduction

Deep neural networks (DNNs) have achieved state-of-the-art classification results in diverse domains such as natural language understanding [73], computer vision [27], and speech recognition [39]. Despite their success, DNNs are generally not well equipped for real-world applications [3] and tend to fail when exposed to data from unseen distributions. This issue often goes unnoticed, as state-of-the-art classification results are primarily obtained in extremely controlled benchmark environments with an inherent closed world assumption, raising the question of applicability to real-world scenarios [28,29,37,53]. Specifically, DNNs tend to generalize only well within the concepts they were trained on but tend to provide incorrect predictions with exaggerated confidence when exposed to samples from unseen distributions [3,21,28,29,37,51], jeopardizing model robustness. As visualized in Fig. 1b, the multi-layer perceptron (MLP) learns to separate the two half-moons and XOR circles but fails to generalize to the uniform noise. Note that the MLP and MiMo [24] models were trained via empirical risk minimization (ERM) to separate the two classes, so there is no preference to which class the outliers should be attributed. However, we would expect a well-generalizing model to be uncertain about these samples and not assign them to one of the classes with high confidence.
Fig. 1  Class probabilities of DAE, MLP and MiMo [24] visualized as contours: in contrast to MiMo and MLP, DAE learns a hull around the red class of interest (COI) for the three datasets, enabling the method to not only separate the two inlier classes but also to reject the unseen uniform noise. MLP and MiMo only wrap the inliers if the rest samples encourage such a decision boundary, as shown for the Bounding Gaussians dataset

In this work, we consider open set recognition (OSR) as the generalized version of one-vs-rest (OVR) classification [5], in which the rest samples not only stem from known classes but also from different unknown sources of outlier generating processes [6,62]. Therefore, the aforementioned deficiencies of DNNs also pose a severe threat to OSR. As suggested by Scheirer et al. [62], these deficiencies can be alleviated by framing the optimization problem in OSR as a combination of ERM and open space risk minimization [62]. The open space is defined as the set of points that have at least distance \(d\) to any of the inlier samples, i.e., the closed space. The open space risk is then defined as the relative measure of the open space volume of false inliers over the closed space volume classified as inliers (i.e., true inliers), as illustrated in Fig. 2. Consequently, minimizing the open space risk forces the model to learn a hull around the inliers leading to increased model robustness.

The difference between ERM and ERM in conjunction with open space risk minimization is illustrated in Fig. 1. Due to its bounded open space risk DAE, learns a hull around the inlier data, whereas the MLP assigns an infinite space to the inlier class resulting in infinite open space risk. In practice, this problem is often mitigated by the long-established background class setup [41], in which all rest classes are subsumed within a single background class [19,57]. This approach can be an effective measure to learn a working OSR model, as illustrated in Fig. 1h. On the downside, this approach is also highly data-intensive, rendering it often infeasible and error-prone while still having an unbounded open space risk.

In OSR, we usually deal with significant class imbalance. When the focus lies on filtering a single, narrow class of interest (COI), similar to the needle in the haystack problem, then this inlier class tends to be underrepresented. When the focus shifts toward outlier detection, for instance, in the case of computer virus detection, inlier samples outnumber instances of the rest class. Further, only some classes of the problem domain are usually known for the set of rest classes (RCs), and outliers or dataset shifts are only witnessed at test time, making OSR a highly imbalanced semi-supervised setting. Throughout the paper, we refer to COI samples as
we provide a novel framework for evaluating OSR w.r.t. classification, outlier detection and dataset shift, partially based on earlier work by Hendrycks and Gimpel [28]. With this setup, we are able to show that the three subtasks $T_c$, $T_o$, and $T_d$ of OSR pose different challenges toward deep learning methods.

- We propose the new method DAE that provides robust results across all subtasks of OSR. This kind of robustness across all three tasks of OSR is not observed for any of the four baselines MLP, MiMo, ATA, and OCA. In contrast to the autoencoder-based baselines ATA and OCA, DAE can be trained in an end-to-end fashion by alleviating the vanishing gradient problem [67]. Additionally, ATA provides probability scores for COI, which is an important feature for deployment in safety-critical systems.

**Fig. 2** Conceptual illustration of OSR: the closed set $S_c$ is visualized by the green area comprising all inliers. There are two different rest samples types: red points resemble a known rest class, whereas blue and pink samples reflect unobserved outliers. The objective is to learn an indicator function $f$ (decision boundary shown as blue line) that filters inliers and rejects all rest samples. The open space risk $R_O(f)$ can be interpreted as the ratio of the blue area over the green area that is bounded by the decision boundary of $f$ (color figure online).
• While our findings suggest that DAE, MLP, and MiMo become uncalibrated w.r.t. the \( T_d \) and \( T_o \) settings, we find indications that calibration improvements on \( T_c \) will generalize better to \( T_d \) and \( T_o \) for DAE than for MiMo and MLP.

This paper comprises the following additional contributions: While the previous paper focused on OVR, we examine DAE in light of the OSR framework in this paper and prove its upper bound on the open space risk. The benefits of the bound and DAE’s ability to minimize open space risk are empirically verified on a range of text and image classification datasets with a varying number of inlier classes. These experiments reveal the resulting robustness gain and pinpoint DNNs’ deficiencies with their lack of a bounded open space risk. We additionally benchmark DAE against recent OSR methods C2AE and OpenMax and show its superiority from an intuitive and empirical point of view.

Further, we visually compare the reconstruction differences on two image datasets between DAE and the autoencoder-based baselines and explain the benefits of the adversarial loss function. Moreover, we show DAE’s robustness w.r.t. local stability in feature space and adversarial robustness using the FGSM method [21]. Finally, we provide an ablation study empirically verifying that only the combination of all loss terms yields the desired classification and robustness properties.

With its added theoretical foundation and empirical verification, DAE becomes a promising candidate for deployment in safety-critical systems.

2 Related work
The OVR classification strategy is often applied to binary models in order to extend them to multi-class classification [5]. Naturally, there is no necessity for OVR classification for DNNs, since they already support multi-class classification by design. However, there are common situations where OVR becomes relevant, e.g., if faced with only positively labeled samples and all remaining samples with potentially unknown sources are assigned to a single negative class [16,31,66] or if the goal, as in this paper, is to filter normal samples from abnormal ones [6]. As motivated previously, vanilla MLPs are unsuitable in this case due to their infinite open space risk and consequential robustness deficiencies toward outliers [6]. While in modern architectures, researchers often try to circumvent this problem by subsuming all rest classes within a single background class [19,57], this effectively only reduces but does not solve the unbounded open space risk [6]. However, there have been few attempts to bound the open space risk in DNNs. For instance, Bendal and Boult [4] and Rudd et al. [59] leverage extreme value theory (EVT) to determine a compact abating probability model based on the deep features of the full network outputs. Noteworthy, both approaches are offline and are therefore not involved in training the network. The autoencoder-based approach C2AE proposed by Oza and Patel [52] also uses EVT to determine the decision boundary in reconstruction error space and requires at least two inlier classes. Other approaches try to bound open space risk by using tent activation functions [58]. They show that it increases robustness to adversarial attacks while potentially compromising classification performance. In conclusion, as stated in a recent survey, OSR is still largely unsolved within the deep learning domain [6].

In the related OCC setting with its focus on outlier detection, DNN-based approaches have been researched from three angles: (1) combining kernel methods [65] with DNN methods [14,60,71], (2) generative models (e.g., generative adversarial networks [22] or variational autoencoders [35])-based outlier detectors [50,64,72], and (3) based on (semi-) supervised autoencoders [8,10,26,44,45,47]. Here, the key idea is to learn a representation of the inlier distribution and subsequently to estimate the outlierness of a sample via its reconstruction error.

Other contributions focus on the calibration, as DNNs tend to provide wrong predictions with overly high confidence estimates for out-of-distribution samples [23,37]. Since OVR classification incorporates outliers when extended to OCC, this issue is also prominent in OVR classification. Several methods have been proposed: either by adding a calibration task to the model, which aligns it with the target probability distribution [12] or by incorporating diverse predictions from ensemble methods such as deep ensembles [37] and MiMo [24]. These methods combine multiple neural networks as weak classifiers, whose diverse outputs are aggregated to well-calibrated predictions, thus compensating for overly confident predictions. Conversely, as pointed out by Van Amersfoort et al. [67] and also supported by our findings, the diversity of the weak classifiers within the ensemble methods is not strong enough to generalize well to out-of-distribution samples. Instead, Amersfoort et al. propose the kernel-based method DUQ, which learns centroids of classes in a lower-dimensional space [67]. Uncertainty is measured as a distance from the class centroids to out-of-distribution points.

In contrast to Van Amersfoort et al. [67], our DAE method incorporates radial basis functions to estimate the outlierness of a sample via the distance to its reconstruction and therefore does not require any centroid updating routines. Further, ATA [45] optimizes the decision boundary w.r.t. F1 score in an offline brute-force line search in reconstruction error space, which does not actively minimize the open space risk. For DAE, we designed a customized loss function that allows learning the decision boundary end-to-end and minimizes open space risk. Furthermore, it forces the classes to

\[ TD \] and \( T_c \) settings, we find indications that calibration improvements on \( T_c \) will generalize better to \( T_d \) and \( T_o \) for DAE than for MiMo and MLP.

This paper comprises the following additional contributions: While the previous paper focused on OVR, we examine DAE in light of the OSR framework in this paper and prove its upper bound on the open space risk. The benefits of the bound and DAE’s ability to minimize open space risk are empirically verified on a range of text and image classification datasets with a varying number of inlier classes. These experiments reveal the resulting robustness gain and pinpoint DNNs’ deficiencies with their lack of a bounded open space risk. We additionally benchmark DAE against recent OSR methods C2AE and OpenMax and show its superiority from an intuitive and empirical point of view.

Further, we visually compare the reconstruction differences on two image datasets between DAE and the autoencoder-based baselines and explain the benefits of the adversarial loss function. Moreover, we show DAE’s robustness w.r.t. local stability in feature space and adversarial robustness using the FGSM method [21]. Finally, we provide an ablation study empirically verifying that only the combination of all loss terms yields the desired classification and robustness properties.

With its added theoretical foundation and empirical verification, DAE becomes a promising candidate for deployment in safety-critical systems.

2 Related work
The OVR classification strategy is often applied to binary models in order to extend them to multi-class classification [5]. Naturally, there is no necessity for OVR classification for DNNs, since they already support multi-class classification by design. However, there are common situations where OVR becomes relevant, e.g., if faced with only positively labeled samples and all remaining samples with potentially unknown sources are assigned to a single negative class [16,31,66] or if the goal, as in this paper, is to filter normal samples from abnormal ones [6]. As motivated previously, vanilla MLPs are unsuitable in this case due to their infinite open space risk and consequential robustness deficiencies toward outliers [6]. While in modern architectures, researchers often try to circumvent this problem by subsuming all rest classes within a single background class [19,57], this effectively only reduces but does not solve the unbounded open space risk [6]. However, there have been few attempts to bound the open space risk in DNNs. For instance, Bendal and Boult [4] and Rudd et al. [59] leverage extreme value theory (EVT) to determine a compact abating probability model based on the deep features of the full network outputs. Noteworthy, both approaches are offline and are therefore not involved in training the network. The autoencoder-based approach C2AE proposed by Oza and Patel [52] also uses EVT to determine the decision boundary in reconstruction error space and requires at least two inlier classes. Other approaches try to bound open space risk by using tent activation functions [58]. They show that it increases robustness to adversarial attacks while potentially compromising classification performance. In conclusion, as stated in a recent survey, OSR is still largely unsolved within the deep learning domain [6].

In the related OCC setting with its focus on outlier detection, DNN-based approaches have been researched from three angles: (1) combining kernel methods [65] with DNN methods [14,60,71], (2) generative models (e.g., generative adversarial networks [22] or variational autoencoders [35])-based outlier detectors [50,64,72], and (3) based on (semi-) supervised autoencoders [8,10,26,44,45,47]. Here, the key idea is to learn a representation of the inlier distribution and subsequently to estimate the outlierness of a sample via its reconstruction error.

Other contributions focus on the calibration, as DNNs tend to provide wrong predictions with overly high confidence estimates for out-of-distribution samples [23,37]. Since OVR classification incorporates outliers when extended to OCC, this issue is also prominent in OVR classification. Several methods have been proposed: either by adding a calibration task to the model, which aligns it with the target probability distribution [12] or by incorporating diverse predictions from ensemble methods such as deep ensembles [37] and MiMo [24]. These methods combine multiple neural networks as weak classifiers, whose diverse outputs are aggregated to well-calibrated predictions, thus compensating for overly confident predictions. Conversely, as pointed out by Van Amersfoort et al. [67] and also supported by our findings, the diversity of the weak classifiers within the ensemble methods is not strong enough to generalize well to out-of-distribution samples. Instead, Amersfoort et al. propose the kernel-based method DUQ, which learns centroids of classes in a lower-dimensional space [67]. Uncertainty is measured as a distance from the class centroids to out-of-distribution points.

In contrast to Van Amersfoort et al. [67], our DAE method incorporates radial basis functions to estimate the outlier-
be more separated in reconstruction error space than ATA’s adversarial loss function.

### 3 Decoupling autoencoders

Similar to existing autoencoder-based approaches for outlier detection [8, 26, 45], the decoupling autoencoder (DAE) method learns the outlierness of a sample via its reconstruction error. Existing approaches estimate the decision boundary via brute-force algorithms [1, 45] or learn the decision boundary via a subsequent downstream layer [44]. In contrast, DAE learns the decision boundary end-to-end while optimizing for a pessimistic decision boundary that is most close to the inlier samples without compromising generalization performance. Thus, the decision boundary’s open space risk is actively minimized, a favorable setting in safety-critical systems.

From an architectural point of view, as displayed in Fig. 3, the network reconstructs a sample \( x \in \mathbb{R}^n \) using the autoencoder \( \varphi(x) = d(e(x)) \) consisting of an encoder \( e(\cdot) \) and a decoder \( d(\cdot) \). The reconstruction error \( e_{MSE} \) between the original and reconstructed sample \( \hat{x} \in \mathbb{R}^n \) computes to

\[
e_{MSE}(x, \hat{x}) = \frac{1}{n} \sum^n_i (x_i - \hat{x}_i)^2.
\]

The reconstruction error is mapped to the inlier probability via Gaussian \( g: z \rightarrow e^{\frac{-(z^2)}{2\sigma^2}} \) with a mean of zero. Thus, the higher the reconstruction error, the smaller the inlier probability. More formally, the full network \( f \) is given by

\[
f(x) = g(e_{MSE}(x, d(e(x))), \sigma_1).
\]

Note that the standard deviation \( \sigma_1 \) is directly coupled with the decision boundary \( t \) via

\[
t = \sqrt{-2\sigma^2} \ln \frac{1}{2},
\]

as the threshold is fixed at the 0.5 level of function \( g \), also shown in Fig. 4.

During training, the network has three objectives:

(a) To minimize inlier reconstruction errors and maximize rest sample reconstruction errors, such that the inlier samples are easily distinguishable from rest samples within the one-dimensional reconstruction error space.

(b) To classify samples correctly based on the decision boundary \( t \).

(c) To reduce the open space risk by minimizing the decision boundary \( t \), such that the model is sensitive to unseen outliers.

The overall loss function \( \hat{L} \) incorporates these three training objectives by combining the adversarial loss function \( L_R \), binary cross-entropy (BCE) classification loss \( L_{BCE}(\hat{y}, y) = -[y \ln \hat{y} + (1 - y) \ln(1 - \hat{y})] \) and a regularizer term \( |t| \), as follows:

\[
\hat{L}(x, \hat{x}, y) = L_R(x, \hat{x}, y) + \lambda_2 L_{BCE}(f(x), y) + \lambda_3 |t| \quad (2)
\]

where \( y \) and \( \hat{y} \) denote the target label and the model’s predicted COI probability of sample \( x \), respectively. Factors \( \lambda_2 \) and \( \lambda_3 \) scale the classification loss term and \( |t| \) regularization.

The adversarial reconstruction loss \( L_R \) comprises the minimization and maximization of reconstruction errors w.r.t. inliers and rest samples, as defined by:

\[
L_{R1}(x, \hat{x}, y) = L_{MSE}(x, \hat{x}) w_1(L_{MSE}(x, \hat{x})) \quad (3)
\]

\[
L_{R2}(x, \hat{x}, y) = L_{MSE}(x, \hat{x}) w_0(L_{MSE}(x, \hat{x})) \quad (4)
\]

\[
L_R(x, \hat{x}, y) = \begin{cases} 
\lambda_0 L_{R1}(x, \hat{x}, y), & y \in \text{inliers} \\
\lambda_1 L_{R2}(x, \hat{x}, y), & \text{otherwise,}
\end{cases} \quad (5)
\]

**Fig. 3** Decoupling autoencoder (DAE) architecture: a joint architecture composed of an autoencoder \( \varphi(x) = d(e(x)) \) for sample reconstruction, a reconstruction error module \( e_{MSE} \) for outlierness estimation and a RBF kernel \( g \) with standard deviation \( \sigma_1 \) for classification. Thus, the entire network is given by

\[
f(x) = g(e_{MSE}(x, d(e(x))), \sigma_1)
\]

**Fig. 4** Combination of classification objective \( L_{BCE} \) and \( |t| \) regularization to optimize for a decision boundary \( t \), which is minimal but does not compromise classification of inliers (red points) and rest samples (blue points) (color figure online)
where scaling factors $\lambda_0 \in \mathbb{R}^+$ and $\lambda_1 \in \mathbb{R}^+$ determine the minimization/maximization magnitude, respectively. Within $L_R$, the mean squared error $L_{\text{MSE}}(\hat{x}, \tilde{x}) = \frac{1}{n} \sum^o_i (\tilde{x}_i - \hat{x}_i)^2$ is weighted by $w_i : L_{\text{MSE}} \rightarrow \mathbb{R}$ for inliers and by $w_o : L_{\text{MSE}} \rightarrow \mathbb{R}$ for rest samples. These two functions, given by Eqs. (6) and (7), push the reconstruction errors of inliers and rest samples away from the decision boundary $t$ toward the origin and $\infty$, respectively, thereby providing a clear class separation:

$$w_i(l) = \begin{cases} \frac{1}{e^{-\frac{\|l\|_2^2}{2}\lambda_1}} & l > t \\ e^{-\frac{\|l\|_2^2}{2\lambda_1}} & \text{otherwise} \end{cases}$$  \hspace{1cm} (6)$$

$$w_o(l) = \begin{cases} -1 & l < t \\ -e^{-\frac{\|l\|_2^2}{2\lambda_1}} & \text{otherwise} \end{cases}$$  \hspace{1cm} (7)

Note that in both cases, the standard deviation $\sigma_2$ of the Gaussian is a hyperparameter and determines how far the two classes are being separated. $\sigma_2$ is not to be confused with $\sigma_1$, which is coupled with the decision boundary $t$. The reconstruction error maximization of rest samples is achieved in Eq. (7) by the negation which is equal to flipping the loss gradients [45] and thus corresponds to gradient ascend. The reconstruction error objective $L_R$ is conceptually visualized in Fig. 5.

Due to the spacious separation of the inlier set and rest sample set in reconstruction loss space, there is a wide range of possible decision boundaries $t$. We argue that the best $t$ is as close as possible to the inlier set without compromising the classification performance. Thus, offering a reasonable trade-off between classification ($T_c$) and outlier detection ($T_o$)/dataset shift ($T_d$), while reducing the open space risk. The trade-off is modeled via the second and third addends of $\hat{L}$ in Eq. (2). The $\|t\|_1$ regularizer minimizes the decision boundary $t$ toward 0, eventually leading to an impractical classifier always predicting RC independently of $x$. This impractical solution is prevented by the classification loss term $L_{\text{BCE}}$ acting as a stopping criterion, as visualized in Fig. 4. Note that the four scaling factors $\lambda_0, \ldots, \lambda_3$ trade off these three objectives of reconstruction minimization/maximization, classification performance and out-of-distribution robustness.

The combination of $L_R$ and $L_{\text{BCE}}$ also solves the vanishing gradient problem, which would occur for large $e_{\text{MSE}}$, due to the Gaussian output activation function. When solely minimizing the classification loss $L_{\text{BCE}}$ jointly with the regularizer term, i.e., $\hat{L}^* = L_{\text{BCE}} + \|t\|_1$, we can show that

$$\lim_{e_{\text{MSE}}(\hat{x}, f(x)) \to \infty} \frac{\partial L^*_\text{total}}{\partial \Theta} = 0,$$  \hspace{1cm} (8)

where $\Theta$ are the network weights of autoencoder $\varphi$. The total derivative of $L^*_\text{total}$ computes to

$$\frac{\partial L^*_\text{total}}{\partial \Theta} = \frac{\partial L^*_\text{total}}{\partial f} \frac{\partial f}{\partial g} \frac{\partial g}{\partial e_{\text{MSE}}} \frac{\partial e_{\text{MSE}}}{\partial \varphi} \frac{\partial \varphi}{\partial \Theta}.$$  \hspace{1cm} (9)

As $x$ tends to infinity, Gaussian $g$ becomes a horizontal line, resulting in gradients equal to 0:

$$\lim_{e_{\text{MSE}}(\hat{x}, f(x)) \to \infty} \frac{\partial g}{\partial e_{\text{MSE}}} = 0.$$  \hspace{1cm} (10)

Thus, the expression in Eq. (8) zeros out, which is why the gradient updates become ineffective for inliers with large $e_{\text{MSE}}$. As $L_R$ is independent of Gaussian $g$, it is not affected by this problem and enforces convergence by minimizing these inliers.

### 4 Classification concern conflicting with robustness

During training of DNNs, we minimize a surrogate loss function $L$, such as negative log-likelihood (NLL) instead of the non-differentiable 0-1 loss, over a given empirical data distribution $P_{\text{data}}(x, y)$, as the true $P_{\text{data}}(x, y)$ is
unknown [20]. This corresponds to minimizing the empirical risk which is given by
\[ E(x, y \sim p_{data}(x, y)) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i, \Theta), y_i), \]
where \( N \) is the training set size and \( f \) the model with parameters \( \Theta \). This procedure optimizes for a discriminating function which correctly separates the classes in the training set, i.e., optimization for classification performance. Under the assumption that the empirical data generating distribution \( \hat{p}_{data}(x, y) \) is similar to the true data generating distribution \( p_{data}(x, y) \), then within the problem domain the discriminatory function will generalize to unseen data \( x, y \sim p_{data}(x, y) \).

A problem arises when the model is exposed to samples that are highly unlikely according to \( \hat{p}_{data}(x, y) \), i.e., outliers because the model was not optimized w.r.t. such samples. This issue is visualized in Fig. 1 for MLP and MiMo on the Half-Moon and XOR Circles datasets. Both methods learn to separate the observed data (i.e., \( \hat{p}_{data}(x, y) \)). When we consider only the training data (red and blue samples), the learned model indeed minimizes the empirical risk. However, for out-of-distribution data (orange points), we would like to observe high uncertainty. Since this is not reflected in the training objective, the model often predicts one of the two classes with high confidence, even though the out-of-distribution data cannot be attributed to any of the two classes [28].

A simple solution would be to facilitate reconstructive representation learning with, e.g., autoencoders by forcing the reconstruction error to capture the outlierness of a sample. As shown in Fig. 1 for DAE and in Fig. 6 for ATA, each model has learned a representation of COI and rejects any major deviation from it as RC. While this indeed increases the robustness [44,45], it can also harm the classification performance since a representation for all input features needs to be learned [67]. Uninformative, non-causal features can therefore also have a diminishing effect on classification performance. Models trained within the ERM framework do not suffer from this problem, as the feature extractor would neglect these features [67].

In conclusion, there is a trade-off between classification performance and robustness to outliers/dataset shift which DAE aims to alleviate within the OSR framework.

### 5 Achieving bounded open set recognition with autoencoders

In recent years, novel deep learning algorithms have advanced state-of-the-art in many classification tasks. However, as noted in the previous section, it has also been shown that these algorithms, when solely optimized for empirical risk, often give wrong predictions with high confidence when exposed to dataset shift and outliers. In this section, we formalize this issue in line with Scheirer et al. [63] and prove that our approach has an upper bound on the open space risk, a primary criterion for robust OSR.

OSR was first defined by Scheirer et al. [63] and recently surveyed by Geng et al. [17] and Boult et al. [6] and is still a largely unsolved topic within the deep learning domain.

---

**Fig. 6** Reconstruction error landscape of ATA and OCA visualized on the Half-Moon, XOR Circles, and Bounding Gaussians dataset, equivalent to Fig. 1

(a) ATA on Half-Moon  
(b) ATA on XOR Circles  
(c) ATA on Bounding Gaussians  
(d) OCA on Half-Moon  
(e) OCA on XOR Circles  
(f) OCA on Bounding Gaussians
OSR formalizes the problem of distinguishing a class of interest (i.e., samples originating from an observed set of classes) from samples derived from, e.g., outlier generating processes, dataset shifts, or other possibly unobserved but related classes. As deep learning classifiers are generally trained based on ERM by leveraging a surrogate loss such as cross-entropy, they only learn to differentiate the observed classes. This can be viewed as closed set classification, which is illustrated in Fig. 1: the MLP successfully learns to distinguish the two half-moons resembling the closed set. A problem arises when we zoom out of the problem domain and consider samples from the open set (i.e., outside of the closed set), then these samples that are far away from the closed set are still assigned to one of the two half-moons.

As a solution, OSR proposes an indicator function \( f \) over input space \( \chi \), that maps inliers to 1 and rest samples to 0. Partially following the notation of Scheirer et al. [63], let \( \hat{\chi} \) be the COI and \( S^c_\chi = \{ x \in \chi \mid \min_{x \in \hat{\chi}} |x - s| < d \} \) be the corresponding closed set, i.e., the set of all points within \( \chi \) that are in \( d \) proximity to at least one of the inlier samples \( s \in \hat{\chi} \). Let \( \mathcal{O} = \chi - S^c_\chi \) be the open space, then the open space risk is defined as

\[
R_O(f) = \frac{\int_{\mathcal{O}} f(x)dx}{\int_{S^c_\chi} f(x)dx},
\]

which yields the ratio of the false inlier area over the true inlier area. Thus, \( R_O \) can be minimized by reducing the volume of the indicator function \( f \). In Fig. 2, \( R_O \) can be interpreted as the ratio of the blue area (positive region of \( f \) outside \( S^c_\chi \), i.e., false positives) over the area of correctly predicted inliers in \( S^c_\chi \) (i.e., true positives).

Note that a trivial solution for \( R_O \) minimization could be achieved by predicting all samples (except for one inlier to prevent division by zero) as RC independently of their true class, i.e., resulting in a volume of 0 of the indicator function \( f \) over the open space. That is why it is crucial to counteract this solution by framing OSR as a two-fold problem with the two objectives: a) Minimization of \( R_O(f) \) and b) ERM as regularization.

As shown by the decision boundary of the MLP in Fig. 1 and denoted in prior research, vanilla MLPs do not provide an upper bound on the open space risk and in practice are generally unbounded [4,58,63]. Therefore, we have to turn toward deep learning architectures, such as autoencoders, that are capable of learning manifolds on the input space and by design learn a hull around the inliers.

**Lemma 1** Any autoencoder with saturating activation functions (e.g., sigmoid) within at least one of its layers and a reconstruction error output module acting as a manifold learner has a bounded open space risk \( R_O \).

**Proof** Let \( f(x) = \mathbb{1}(|\varphi(x) - x| < d) \) be the recognition function. \( f(x) \) is an indicator function which maps the reconstruction error function of autoencoder \( \varphi : \mathbb{R}^n \to \mathbb{R}^n \) to \{0, 1\} based on threshold \( d \in \mathbb{R} \).

Let \( \varphi \) comprise at least one layer with an activation function that is saturating toward both tails (here, sigmoid). Assuming layer \( \varphi^l \) to be the last layer with sigmoid activation and \( \varphi^l \) to have \( m \) neurons, then the image this layer maps to is fixed within hypercube \((0, 1)^m\). Therefore, the image of all subsequent layers \( \varphi^i, \forall i > s \), is also fixed. It follows that \( \lim_{x \in \mathbb{R}^n \to \infty} \varphi(x) - x = \infty \), as the image of \( \varphi(x) \) is bounded. In conclusion, when starting from a sample classified as inlier and moving in any but fixed direction in feature space, then the reconstruction error will approach infinity. This enforces to cross the decision boundary of the recognition function \( f \), proving the existence of a bound on the open space risk.

**Lemma 2** The open space risk can be approximated solely based on the weights of the layers succeeding the last layer with saturating activations.

**Proof** As previously defined, let \( f(x) \) be the recognition function based on the autoencoder \( \varphi \) with \( \varphi^l \) being the last layer with saturating activation function \( a^l \) and \( a^l \) be the corresponding activations. Neurons in the subsequent layers can have a monotonic, non-saturating activation function that is saturating toward both tails (here, sigmoid). Therefore, the image of \( x^* \) can be approximated as the supremum of activation \( a^l_j+1 \) and neuron \( \varphi^l_j+1 \) can be bounded by the following inequalities, respectively:

\[
\sup_{x \in \chi} \varphi^l_j+1(x^*) \leq \sup_{x^* \in \mathbb{R}} \left| w^{l+1} \sup_{x^* \in \mathbb{R}} a^l(x^*) \right| + |b^{l+1}| \quad (12)
\]

\[
\sup_{x \in \chi} \varphi^l_j+1(x) \leq a^{l+1} \left( \sup_{x^* \in \mathbb{R}} \alpha^l_j+1(x^*) \right), \quad (13)
\]

where \( w^k_i \) denotes the weight between neuron \( i \) of layer \( k - 1 \) to neuron \( j \) of layer \( k \) and \( b^k \) denotes the bias of neuron \( j \) in layer \( k \). It follows for the subsequent layers \( s + 1 \),

\[
\sup_{x \in \chi} \varphi^{s+1+l}(x) \leq a^{s+1+l} \left( \sup_{x^* \in \mathbb{R}} \alpha^l_j+1(x^*) \right). \quad (14)
\]

Therefore, the image of \( \varphi \) can be bounded by a hypercube with its center at the origin. It follows that the recognition function’s open space \( \int_{\mathcal{O}} f(x)dx \) can be approximated as the union of the set of points that are within the hypercube and those that are in less than \( d \) proximity to the hypercube.

**Lemmas 1 and 2** have multiple practical implications. Firstly, our autoencoder methods are capable of bounding
the open space risk and are therefore by design superior to MLP-based architectures in the OSR setting. Secondly, approximating the open space risk enables us to filter models with higher robustness during the model selection process. Thirdly, given that the approximated bound is a hypercube with its center in the origin of the feature space, it is recommendable to perform feature transformation such that the inlier class is located close to the origin, thus allowing for a smaller hypercube. Furthermore, Lemma 1 depicts the dependency of the open space risk on the weights after the last layer with saturating activation functions. By regularizing the weights in conjunction with the centering of the inlier samples, it is possible to actively minimize the bound on the open space risk. While this idea is out-of-scope of this contribution, it is a promising future research direction.

For instance, it is possible to handcraft the connections in the first layer of $\varphi$ such that the weights perform translation and scaling of the feature space.

### 6 Toward adversarial robustness and local stability

Adversarial perturbations offer an effective way of measuring a model’s robustness locally as well as globally. As initially described by Goodfellow et al. [21], the idea of gradient-based adversarial attacks is to confuse a neural network $f$ with parameters $\Theta$ by adding an imperceptible perturbation to the original sample $x$ with target $y$. The perturbation itself is not arbitrary but maximizes the loss $L$ of the network. A common methodology of calculating these adversarial examples is the fast gradient sign method (FGSM) [21], which determines the perturbation by taking the sign of the gradients w.r.t. the sample:

$$\eta = \epsilon \text{sign}(\nabla_x L(\Theta, x, y)), \quad (15)$$

where scaling factor $\epsilon$ determines the volume of change. Note that since $\epsilon$ is fixed, also the perturbation’s volume is fixed for all steps and across models.

While in practice, adversarial examples are often used to improve model robustness via adversarial training [21], in this work, we use the FGSM framework for robustness estimation of trained models. By perturbing a sample in a step-wise adversarial fashion, the model confidence development can be tracked which provides deep insights into local stability and with an increasing number of steps also into global robustness of the model. Technically, at a given step $i$, a sample $x^{i-1}$ is perturbed according to

$$x^i = x^{i-1} + \epsilon \text{sign}(\nabla_{x^{i-1}} L(\Theta, x^{i-1}, y)) \quad (16)$$

and the difference in confidence $\Delta c_i$ at step $i$ w.r.t. the original sample is defined as

$$\Delta c_i(x^0) = f(x^i) - f(x^0), \quad (17)$$

where $x^0$ denotes the original sample. By varying the step-size $\epsilon$ and the number of steps, the aforementioned local stability and global robustness can be easily estimated. Further, there also exists another adversarial robustness metric [74], which is defined as:

$$\Psi(x^0) = \min_{\tilde{x} \in [x^0]} \frac{1}{D_{KL}(f(x^0), f(\tilde{x}))} \quad (18)$$

and computes the Kullback–Leibler divergence $D_{KL}$ in the denominator as the relative entropy between the two confidence estimates. However, we did not consider this measure, as $D_{KL}$ rapidly changes around $(0, 0)$ and $(1, 1)$ and small changes in this area therefore lead to overly pessimistic robustness scores.

Generally, we hypothesize that DAE, as a representative for an OSR architecture, is more robust than the MLP, as the latter one rather slices the spaces in discriminative hyperplanes w.r.t. the two classes. Therefore, we can expect a given dataset shift sample to be closer to a hyperplane than DAE’s hull which is learned directly around the inlier class. We would assume similar adversarial robustness between MLP and DAE for an inlier sample.

### 7 Experiments and results

In this section, we discuss the experiment setup and compare the performance of DAE to the aforementioned baselines. The approaches are compared on an algorithmic level by applying nested cross-validation (CV) [32,56,68]. The best models are selected by the highest area under the precision-recall curve (AUPR) score and are reported along with area under the receiver operating characteristics (AUROC) and F1 score with respective confidence measurements.

#### 7.1 Selected baselines

Since OSR touches upon multiple machine learning areas such as binary classification and outlier detection, we decided to consider the two most prominent methods published under the OSR framework and seven potent baselines from the outlier detection, OVR classification and ensembling domain.

**OpenMax:** this is an offline OSR method that replaces the softmax layer within a fully trained multi-class DNN [4]. It applies extreme value theory to the network activations, which yields a final layer that outputs a probability score for the outlierness of a sample and closed-set probability
scores. While the method’s theoretical foundations are sound, it assumes that outlier activations differ significantly from inlier activations. In practice, this requirement is not always fulfilled [67], leading to robustness scores similar to those of the SoftMax baseline. By design, this method is limited to OSR problems with multiple COIs.

C2AE: this OSR approach [52] is derived from OCA which is used in outlier detection. The C2AE autoencoder is trained in a two-step fashion: (1) during closed set training, the encoder is trained jointly with a downstream classifier to perform closed set classification. (2) For decoder training, the latent vector is conditioned on an inlier-class-specific vector. During inference, it is assumed that an inlier has a lower reconstruction error compared to rest samples when the inlier’s latent vector is conditioned on the respective inlier-class-specific vector. Similar to OCA, this requires rest samples to be uncorrelated with inlier samples. By design, this method is limited to OSR problems with multiple COIs and requires #closed-set classes many inference steps for each sample.

OCA: classic semi-supervised, autoencoder-based method from the outlier and novelty detection domain that is trained to reconstruct inliers. The reconstruction error is assumed to be lower for inliers than for outliers, thus rendering the reconstruction error predictive of the inlierness of a sample. As shown by Lübbering et al. [46], this method requires rest samples to be true outliers, as rest samples correlated with inliers tend to get reconstructed accurately.

ATA: this is a recent autoencoder-based method from supervised outlier detection, which, in contrast to OCA, actively maximizes the reconstruction error of rest samples [45]. Therefore, rest samples that are correlated with the COI also get maximized, thus alleviating OCA’s aforementioned deficiencies.

MLP: OVR classification DNN which subsumes the set of rest classed within a single background class [19, 41, 57]. This binary neural network requires the rest samples to wrap the COI in feature space such that the model learns a decision boundary hull around the COI, making this approach highly data-intensive.

SoftMax: this DNN has a softmax layer for inlier probabilities as its final layer. The outlierness of a sample is estimated offline by the entropy of the sample’s predicted inlier class probabilities (softmax output). Since the softmax predictions are generally overly confident for outlier data [28], the utility of this method is limited. By design, this method is restricted to problems with multiple COIs.

MiMo: this ensemble method [24] combines several weak classifiers into a single DNN by weight sharing, making this method more efficient compared to traditional ensemble DNNs. Similar to the OVR setup in MLP, this baseline subsumes non-COI classes into a single rest class. As shown by Lakshminarayanan et al. [37], ensemble methods can yield accurate predictive uncertainty estimates, which could ultimately improve OSR performance.

7.2 Evaluation approach

Evaluating OSR models w.r.t. classification performance and robustness toward outliers and dataset shifts are not straightforward. Firstly, the OSR setting is often highly imbalanced with observed rest samples significantly outnumbering the COI samples, whereas outliers/corruptions generally have a low prevalence in practice. Secondly, depending on the application’s deployment domain, the focus between precision and recall can shift: For instance, in medical diagnosis systems, recall is often of the uttermost importance, whereas precision is generally favorable in equity trading precision. Due to this, threshold-dependent metrics such as F1 score or accuracy can be misleading and fail to capture the big picture.

As a viable solution to this issue, there are threshold-independent metrics such as AUPR and AUROC, which evaluate the model at all possible thresholds [11]. The AUROC metric calculates the area under the receiver operating characteristic (ROC) curve, which maps the false positive rate (FPR) onto the true positive rate (TPR) for each threshold. The two rates are defined by $\text{FPR} = \frac{FP}{FP+TN}$ and $\text{TPR} = \frac{TP}{TP+FN}$, where TP, FP, TN and FN depict the number of true positives, false positives, true negatives and false negatives at a certain threshold, respectively. Note that TPR is also often referred to as recall in the literature. From an interpretation point of view, AUROC yields the probability of a random rest sample being ranked higher than a random inlier [15, 28] and therefore is invariant to class imbalance. This invariance to class imbalance allows for interpretable results across datasets since a random classifier achieves an expected AUROC score of 50% and a perfect classifier a score of 100%.

Since AUROC in isolation is insufficient for OVR classification with its prevalent class imbalance, we additionally consider AUPR. This metric is generally employed when faced with the “needle in the haystack problem”, as AUPR takes the different class base rates into account [28]. Similar to AUROC, the PR curve maps the recall onto the precision $\frac{TP}{TP+FP}$ for each threshold. Since the AUPR metric is base rate dependent, there is no fixed baseline AUPR score for a random classifier as there is for AUROC. In fact, given a random classifier, the AUPR score is roughly equal to the random classifier’s precision, which is equal to the rate of the positive class [28, 61]. Assuming a random classifier predicts sample $i$ as the positive class with confidence $p_i \sim U[0, 1]$, then for any given but fixed threshold $q$ we get two sample sets: (1) set of positive predicted samples, (2) set of negative predicted samples. Naturally, any random subsample set has an expected positive rate of $POS/N$, where
POS and $N$ are the number of positive samples and the number of all samples in the entire dataset, respectively. Therefore for an arbitrary but fixed threshold $q \in [0, 1]$, the subsample set of positive predictions contains an expected rate of true positives of $\text{POS}/N$, which is equivalent to the precision. Similar consideration hold for recall. Given any threshold $q$, we sample approx. $q \cdot N$ many samples of which approx. $q \cdot \text{POS}$ are positive, i.e., recall is approx. equal to $q$. This is why the random classifier has a constant PR curve at the precision level of number of positive samples/total number of samples and an AUPR score of the same value. In conclusion, it is essential to communicate the AUPR scores always w.r.t. the random classifier performance and the pre-defined positive class; otherwise, these metrics lack interpretation means.

Additionally, F1 score is also reported to show if each algorithm can learn reasonable decision boundaries, especially w.r.t. the classification task $T_c$. In contrast to the former mentioned AUROC and AUPR metric, this metric evaluates the model at a fixed 0.5 threshold level, which is reasonable for task $T_c$, but is less conclusive for tasks $T_d$ and $T_s$. Similar to AUPR, the F1 score metric is also affected by the base rate of the positive class. Following the argumentation of the AUPR baseline, the F1 score of a random classifier calculates to

$$\text{F1 score}_{\text{base}} = \frac{\frac{\text{POS}}{N} \cdot 0.5}{\frac{\text{POS}}{N} + 0.5}. \quad (19)$$

Further, the algorithms are evaluated based on the correctness of their subjective uncertainty estimation (calibration) in terms of the class-wise expected calibration error (CECE) [23,36,49] and Brier score [7]. While calibration is a crucial criterion for the trustworthiness of machine learning models, it is noteworthy that it is an orthogonal concern to model accuracy, e.g., in the trivial case, a uniformly random classifier is perfectly calibrated on a balanced dataset but inaccurate [37].

The CECE metric is defined as

$$\text{bin-CE}_{i,j} = |y_j(B_{i,j}) - \hat{p}_j(B_{i,j})| \quad (20)$$

$$\text{CECE} = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=1}^{m} \frac{|B_{i,j}|}{N} \text{bin-CE}_{i,j} \quad (21)$$

where parameters $K, m, N$ denote the number of classes, number of bins and dataset size, respectively. The set $B_{i,j}$ contains the samples whose confidence prediction w.r.t. class $j$ falls into the $i^{th}$ bin. The actual ratio of class $j$ samples and average predicted confidence of samples in the bin is denoted by $y(B_{i,j})$ and $\hat{p}_j(B_{i,j})$, respectively. Therefore, bin-CE$_{i,j}$ denotes the calibration error for class $j$ within bin $i$ and CECE is computed as the weighted average over all bin-CE$_{i,j}$.

The Brier score for binary classification is defined as

$$BS = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i; \Theta))^2, \quad (22)$$

where $N$ is the number of samples, $y_i$ is the target of sample $x_i$ and $f(x_i; \Theta)$ is the respective prediction (probability) of the model.

For a comprehensive robustness study, we also evaluate the algorithms w.r.t. their robustness to adversarial attacks and the related concern of local stability via the difference in confidence $\Delta c_i$ [see Eq. (17)]. We empirically determined that a perturbation scaling factor $\epsilon = 0.001$ and $\#\text{steps} = 300$ captures both adversarial robustness and local stability. While for MLP, it is straightforward to calculate the sample’s perturbation w.r.t. the binary cross-entropy loss, DAE’s training loss yields gradients that are almost 0 for small and large reconstruction errors due to the Gaussian’s flatness at its mean and limits. Similar to Eq. (8), this results in dead updates within FGSM. As a solution, we calculate the gradients directly on the reconstruction error $\epsilon_{\text{MSE}}$.

7.3 Datasets

To evaluate the models on OSR with its classification subtask $T_c$ and the more challenging subtasks of contextual outlier detection $T_d$ and dataset shift $T_s$, we extended four textual datasets and three image datasets:

**Reuters dataset**: this multi-label dataset is a standard benchmark for document classification and outlier detection, which contains 10,788 news documents from 90 different categories published by the news outlet Reuters. Since multilabel classification is out of the scope of this work, we only consider documents that are assigned to a single class.

**ATIS dataset**: this dataset comprises 5871 transcribed queries that passengers requested to the air travel information system (ATIS) for flight related information. These queries were grouped into 17 categories.

**Newsgroups dataset**: this dataset contains 18,000 news-group posts from 20 different topics and is a standard dataset for text classification and text clustering.

**TREC dataset**: question classification dataset contains 5500 questions not limited to any particular topic domain [40]. This makes the dataset compelling for dataset shift evaluation.

**MNIST dataset**: image classification dataset containing 70,000 images of handwritten digits [38]. Each of the $28 \times 28$ grayscale images shows a single digit ($0$–$9$).

**FMNIST dataset**: image classification dataset comprising 70,000 fashion items [70]. Each of the $28 \times 28$ grayscale images shows one of the ten fashion items t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, or ankle boot.
EMNIST-letter dataset: dataset for image classification providing 145,600 grayscale images of alphabetic characters. Similar to TREC, we use this dataset solely for evaluation.

Due to the high number of different categories and their large size, each dataset is a viable benchmark dataset for OSR model evaluation. As shown in Table 1, we evaluated DAE and the baselines in seven different experiment setups. For each experiment setup, there is a single train split St and up to five test splits. All splits share the same set of inlier classes, which depending on the derived dataset vary from a narrow to a broader range of topics. The rest class covers a wide range of topics with an increasing scope from Tc, To to Td.

Specifically, classification split Sc shares all training classes and therefore resembles classic OVR classification Tc. Split So increases the scope by incorporating contextual outliers from the same dataset, as defined by the outlier detection task To. Finally, Sd1, Sd2 and Sd3 have maximum RC scope by providing rest samples from a completely new dataset, representative of dataset shift task Td. We vectorized the samples of the textual datasets as pooled 100-dimensional dense Glove embeddings [54] and z-transformed the image samples.

### 7.4 Results

To benchmark DAE against the aforementioned baselines, we train all models with approximately analogical complexity in terms of the number of trainable parameters. For text classification, the models MiMo, OpenMax, SoftMax, and MLP have three hidden layers of sizes 50, 25, and 12. The autoencoder-based approaches have three hidden layers of size 50, 25, and 12 for the encoder and the decoder in reverse order. For image classification, OpenMax, SoftMax and MLP have hidden layers of sizes 410, 256, 128, 64, and 43; MiMo has hidden layer sizes of 120, 32, and 16. The autoencoder-based image classifiers have hidden layers of size 256 and 128 for the encoder and the decoder in reverse order. C2AE has additional classification downstream layers of sizes 128, 32, 5. MiMo has five ensemble components and, therefore, an input size five times the input size of the other approaches. All neurons have sigmoid activation functions.

Within the nested CV, we performed a hyperparameter search concerning lr, balanced sampling and weight decay for all approaches. Specifically for DAE, we optimized for the initial decision boundary t and σ2 and the loss scaling factors λi as defined in Eqs. (2) and (7). ATA was optimized w.r.t. outlier weighting factor and bin range. We found that across various experiments all baselines showed best performances when optimized with Adam [34].

Tables 3 and 4 show the results for each task with the best performing approaches on each dataset highlighted in boldface. To make model robustness comparable, a model is counted as weak when its AUROC score drops at least 5
| Dataset   | Abbr. | Rest labels                                                                 |
|-----------|-------|-----------------------------------------------------------------------------|
| Reuters   | rc    | Carcass, cotton, cpi, crude, gnp, heat, housing, interest, ipi, jobs, livestock, lumber, money-fx, money-supply, nat-gas, oilseed, orange, pet-chem, reserves, retail, rubber, ship, tin, wpi, wpi |
|           | ro    | Alum, bop, cocoa, coconut, coffee, copper, fuel, gas, gold, grain, groundnut, income, iron-steel, lei, meal-feed, nickel, potato, rice, sugar, tea, veg-oil, zinc |
|           | rd    | Alum, cocoa, coffee, copper, cotton, cpi, gold, grain, interest, ipi, jobs, money-fx, money-supply, nat-gas, reserves, rubber, ship, sugar, tin, trade, yen |
| ATIS      | ac    | Airfare, airline, ground_service                                            |
|           | ao    | Abbreviation, restriction, airport, quantity, meal, city, flight_no, ground_fare, flight_time, distance, aircraft, capacity |
|           | ad    | Abbreviation, aircraft, airport, capacity, city, distance,flight_no, ground_fare, flight_time, ground_fare, meal, quantity, restriction |
| News groups | nc    | Comp.os.ms-windows.msc, misc.forsale, rec.sport.baseball, sci.crypt, sci.med, soc.religion.christian, talk.politics.guns, talk.politics.misc |
|           | no    | Rec.sport.hockey, comp.sys.ibm.pc.hardware, talk.religion.misc, talk.politics.mideast, comp.sys.mac.hardware, sci.electronics, alt.atheism, rec.motorcycles, rec.autos, comp.windows.x, comp.graphics |
|           | nd    | Alt.atheism, comp.graphics, comp.os.ms-windows.msc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc |
| TREC      | td    | HUM, NUM, LOC, ABBR                                                        |
| MNIST     | m₁    | 0, 1, 2, 3, 4, 9                                                           |
|           | mₑ    | 0, 1, 3, 4, 9                                                              |
|           | m₀    | 5, 6, 8                                                                    |
|           | mₙ    | 0, 1, 2, 3, 4, 5, 6, 7, 8, 9                                               |
| FMNIST    | f₁    | T-shirt, trouser, pullover, dress, coat, ankle boot                         |
|           | fₑ    | T-shirt, trouser, pullover, coat, ankle boot                               |
|           | f₀    | Coat, ankle boot                                                           |
|           | fₙ    | Sandal, shirt, bag                                                         |
|           | f₆    | T-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot |
| EMNIST-letter | eₒ   | a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z |
Table 3 Performance of DAE and the baselines ATA, OCA, MLP and MiMo on splits $S_n$, $S_{n+1}$, $S_{n+2}$ and $S_{n,0}$ of the textual and image datasets with a single COI. The AUROC results are aggregated in the first column for each split, counting best and weak performances. The best and weak scores are highlighted in bold face and by underline, respectively. Specifically, when a model's AUROC score drops at least 5% points below the best AUROC score, this model is counted as weak performing. For reference, BASE denotes the performance of a random classifier that predicts

| AUROC Agg. | ATIS | NEWSGROUFS | MNIST$_7$ | FMNIST$_7$ |
|------------|------|------------|----------|------------|
|            | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score |
| $S_n$ DAE 0 0 | 98.6 ±0.8 | 99.6 ±0.2 | 97.9 ±1.1 | 96.6 ±0.6 | 86.0 ±2.0 | 60.4 ±26.4 | 99.3 ±0.1 | 98.3 ±0.2 | 96.8 ±1.0 |
| ATA 0 0 | 95.4 ±1.0 | 98.9 ±0.2 | 92.7 ±1.4 | 94.8 ±1.6 | 86.7 ±2.1 | 74.7 ±1.6 | 99.1 ±0.2 | 98.7 ±0.2 | 97.9 ±0.1 |
| OCA 0 2 | 72.0 ±2.1 | 91.4 ±1.1 | 64.1 ±1.7 | 67.4 ±1.8 | 28.4 ±13.0 | 30.8 ±3.2 | 94.9 ±0.2 | 82.2 ±0.5 | 70.1 ±1.3 |
| MLP 4 0 | 99.9 ±0.1 | 99.8 ±0.0 | 98.0 ±0.4 | 97.5 ±0.4 | 91.1 ±0.9 | 79.6 ±4.8 | 99.8 ±0.1 | 99.3 ±0.3 | 97.8 ±0.1 |
| MiMo 0 0 | 98.9 ±0.1 | 99.8 ±0.0 | 97.6 ±0.4 | 97.1 ±0.7 | 89.4 ±1.3 | 81.9 ±1.6 | 99.7 ±0.0 | 98.7 ±0.1 | 95.4 ±0.5 |
| BASE | 50.0 | 81.9 | 31.0 | 50.0 | 11.4 | 9.3 | 50.0 | 14.6 | 11.3 |
| $S_{n+1}$ DAE 1 0 | 89.3 ±0.5 | 91.4 ±0.2 | 84.1 ±1.4 | 94.6 ±1.2 | 50.3 ±4.0 | 16.3 ±9.6 | 99.3 ±0.1 | 98.7 ±0.1 | 96.7 ±0.9 |
| ATA 0 0 | 87.4 ±0.5 | 89.9 ±0.5 | 85.6 ±0.8 | 91.8 ±1.8 | 59.8 ±2.4 | 18.5 ±0.9 | 99.1 ±0.2 | 98.8 ±0.2 | 96.7 ±0.4 |
| OCA 1 1 | 80.4 ±0.5 | 81.9 ±0.6 | 62.1 ±1.5 | 63.6 ±1.9 | 6.6 ±0.8 | 8.9 ±0.9 | 99.7 ±0.0 | 98.9 ±0.1 | 73.9 ±1.8 |
| MLP 2 1 | 85.3 ±1.7 | 86.2 ±1.7 | 85.3 ±0.8 | 95.6 ±0.6 | 95.6 ±3.3 | 25.0 ±5.1 | 99.8 ±0.0 | 98.7 ±0.3 | 96.8 ±0.1 |
| MiMo 0 0 | 87.6 ±1.9 | 90.0 ±1.9 | 85.1 ±0.8 | 95.1 ±1.0 | 58.7 ±4.6 | 48.0 ±2.9 | 99.7 ±0.0 | 98.9 ±0.0 | 95.4 ±0.4 |
| BASE | 50.0 | 61.8 | 27.6 | 50.0 | 1.9 | 1.8 | 50.0 | 12.7 | 10.1 |
| $S_{n+2}$ DAE 1 0 | 99.0 ±0.1 | 98.4 ±0.3 | 42.3 ±2.2 | 97.7 ±1.8 | 95.5 ±2.7 | 46.5 ±24.7 | 99.3 ±0.1 | 99.0 ±0.2 | 99.0 ±0.2 |
| ATA 0 0 | 97.9 ±1.8 | 95.3 ±3.1 | 85.2 ±1.0 | 97.2 ±6.2 | 55.3 ±1.8 | 70.3 ±1.7 | 98.6 ±0.1 | 98.5 ±0.2 | 60.8 ±1.2 |
| OCA 3 0 | 97.9 ±0.2 | 96.5 ±0.2 | 65.8 ±1.6 | 99.3 ±0.6 | 98.9 ±0.7 | 45.1 ±7.2 | 99.9 ±0.0 | 99.9 ±0.0 | 73.9 ±1.8 |
| MLP 0 3 | 73.1 ±8.2 | 73.4 ±7.7 | 46.3 ±3.3 | 91.5 ±3.0 | 31.1 ±5.5 | 41.9 ±9.6 | 99.8 ±0.0 | 99.8 ±0.1 | 98.5 ±0.2 |
| MiMo 0 0 | 80.5 ±1.1 | 39.1 ±1.5 | 44.3 ±0.9 | 84.6 ±1.7 | 26.7 ±4.9 | 28.1 ±2.3 | 99.8 ±0.1 | 99.5 ±0.1 | 97.1 ±0.5 |
| BASE | 50.0 | 20.9 | 14.7 | 50.0 | 6.1 | 5.4 | 50.0 | 22.6 | 15.6 |
| $S_{n+3}$ DAE 0 0 | 98.3 ±0.2 | 98.1 ±0.2 | 59.7 ±3.6 | 97.1 ±2.5 | 95.3 ±3.1 | 35.9 ±25.7 | 99.3 ±0.1 | 98.7 ±0.2 | 97.1 ±0.4 |
| ATA 1 1 | 99.0 ±0.3 | 98.7 ±0.5 | 92.7 ±0.6 | 90.2 ±2.0 | 38.8 ±4.0 | 26.9 ±3.2 | 98.6 ±1.3 | 98.3 ±2.2 | 71.7 ±12.6 |
| OCA 3 0 | 94.3 ±0.4 | 94.4 ±0.4 | 65.8 ±1.6 | 99.4 ±0.5 | 99.1 ±0.6 | 45.1 ±7.2 | 99.6 ±0.1 | 97.5 ±0.3 | 73.7 ±1.9 |
| MLP 0 3 | 82.0 ±10.0 | 68.1 ±16.3 | 64.5 ±5.5 | 81.4 ±11.1 | 20.3 ±18.8 | 19.0 ±5.9 | 99.4 ±0.1 | 94.2 ±1.9 | 93.2 ±1.4 |
| MiMo 0 1 | 95.7 ±0.8 | 92.1 ±1.6 | 74.0 ±1.6 | 60.2 ±3.9 | 9.3 ±5.0 | 8.1 ±0.1 | 99.4 ±0.1 | 96.6 ±0.5 | 90.5 ±0.9 |
| BASE | 50.0 | 34.6 | 20.5 | 50.0 | 4.2 | 3.9 | 50.0 | 12.3 | 9.9 |
| $S_{n+4}$ DAE 1 0 | 96.1 ±0.5 | 87.8 ±1.8 | 8.0 ±0.1 | 96.8 ±1.4 | 93.4 ±2.6 | 42.7 ±16.9 |
| ATA 0 0 | 93.6 ±2.0 | 81.7 ±5.1 | 45.3 ±20.7 | 96.2 ±1.0 | 78.7 ±2.8 | 71.8 ±2.1 |
| OCA 1 0 | 92.3 ±0.5 | 83.3 ±0.9 | 65.8 ±1.6 | 98.9 ±0.7 | 98.2 ±0.9 | 45.1 ±7.2 |
| MLP 0 1 | 69.9 ±8.1 | 12.7 ±12.9 | 9.2 ±1.3 | 95.9 ±2.2 | 75.0 ±10.0 | 70.2 ±10.1 |
| MiMo 0 1 | 92.2 ±1.7 | 54.4 ±4.3 | 11.1 ±0.9 | 88.0 ±2.3 | 57.4 ±11.5 | 54.1 ±3.9 |
| BASE | 50.0 | 4.1 | 3.8 | 50.0 | 11.5 | 9.3 |
Table 4  Performance of DAE and the baselines on the derived datasets with more than one COI. Across all tasks DAE and ATA show high robustness, whereas the remaining baselines perform poorly on at least a single task. Baseline results on the Reuters dataset partially adopted from Lübbering et al. [46]

| #best | AUROC Agg. | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score | AUROC | AUPR | F1 score |
|-------|------------|-------|------|----------|-------|------|----------|-------|------|----------|-------|------|----------|-------|------|----------|
| S_1   | DAE 0      | 99.2  | 0.3  | 99.7     | 0.1  | 99.0 | 0.0     | 99.0 | 0.1 | 97.8     | 0.1 | 96.8 | 0.4     | 97.5 | 0.3 | 93.2     | 0.9 | 91.2 | 0.7     | 97.1 | 0.2 | 98.6     | 0.1 | 95.9 | 0.1     |
| ATA   | 0          | 99.4  | 0.2  | 99.8     | 0.1  | 97.8 | 0.6     | 99.4 | 0.2 | 99.2     | 0.2 | 98.4 | 0.1     | 97.0 | 0.3 | 95.9     | 0.2 | 93.5 | 0.4     | 97.2 | 0.5 | 98.9     | 0.2 | 96.3 | 0.3     |
| OCA   | 0 4        | 78.3  | 5.3  | 92.7     | 2.2  | 73.5 | 3.8     | 73.8 | 1.9 | 58.6     | 2.5 | 34.9 | 3.1     | 91.3 | 0.2 | 82.4     | 1.1 | 73.0 | 0.5     | 81.2 | 0.6 | 91.7     | 0.4 | 76.5 | 1.2     |
| C2AE  | 0 4        | 83.3  | 0.7  | 94.6     | 0.3  | 66.2 | 2.3     | 79.7 | 1.5 | 65.8     | 1.7 | 53.2 | 0.3     | 90.7 | 0.6 | 80.8     | 1.2 | 64.3 | 3.1     | 82.6 | 0.4 | 92.7     | 0.3 | 83.8 | 0.8     |
| OpenMax | 0 4     | 74.5  | 2.2  | 90.3     | 2.4  | 87.2 | 0.0     | 84.3 | 1.0 | 69.4     | 2.0 | 44.4 | 0.0     | 53.8 | 9.7 | 31.9     | 6.6 | 44.4 | 0.0     | 66.6 | 12.7 | 82.8     | 7.3 | 83.3 | 0.0     |
| SoftMax | 0 4     | 74.4  | 2.1  | 90.3     | 2.4  | 87.2 | 0.0     | 84.3 | 1.0 | 69.4     | 2.0 | 44.4 | 0.0     | 56.2 | 11.8 | 34.5     | 7.3 | 44.5 | 0.0     | 63.0 | 2.0 | 75.6     | 1.0 | 83.3 | 0.0     |
| MLP   | 4 0        | 99.7  | 0.1  | 99.9     | 0.1  | 99.9 | 0.3     | 99.9 | 0.1 | 99.8     | 0.1 | 98.9 | 0.2     | 99.4 | 0.1 | 98.6     | 0.2 | 93.8 | 0.4     | 98.7 | 0.1 | 99.4     | 0.1 | 96.5 | 0.1     |
| MiMo  | 1          | 99.6  | 0.1  | 99.9     | 0.0  | 98.6 | 0.1     | 99.7 | 0.0 | 99.3     | 0.1 | 96.8 | 0.3     | 98.7 | 0.2 | 97.1     | 0.4 | 91.2 | 1.0     | 98.2 | 0.1 | 99.2     | 0.1 | 95.8 | 0.2     |
| BASE  | 50.0       | 77.3  | 30.4 | 50.0     | 28.6 | 18.2 | 50.0     | 28.6 | 18.2 | 50.0     | 50.0 | 71.4 | 29.4     | 50.0 | 28.6 | 18.2     | 50.0 | 71.4 | 29.4     | 50.0 | 28.6 | 18.2     | 50.0 | 71.4 | 29.4     |

S_2: DAE 2 0 98.9 0.2 99.2 0.3 97.8 0.2 98.7 0.1 97.1 0.2 81.2 3.0 95.2 0.2 87.4 0.3 64.3 1.2 89.2 0.3 84.4 0.4 67.9 0.8
ATA 1 0 99.4 0.0 99.7 0.0 97.1 0.6 99.3 0.1 98.9 0.2 90.0 2.9 93.9 0.4 87.2 0.6 56.9 0.9 88.0 0.8 83.8 0.4 64.0 0.4
OCA 0 2 75.1 4.1 86.1 2.7 70.8 3.7 95.4 1.1 90.8 2.2 36.4 3.4 92.6 0.3 78.7 1.5 71.3 0.7 83.4 0.3 72.9 0.8 69.3 0.8
C2AE 0 2 76.6 1.4 87.0 0.8 64.8 2.1 95.6 0.9 91.0 1.8 67.5 2.7 87.6 0.5 66.6 1.7 57.2 3.4 84.8 1.2 78.5 2.2 70.3 3.2
OpenMax 0 4 81.0 2.4 87.3 3.1 77.4 0.0 76.2 1.6 46.1 2.4 36.3 0.0 72.7 12.8 41.0 13.9 34.8 0.0 77.0 7.9 65.9 13.7 57.1 0.0
SoftMax 0 4 81.0 2.3 87.3 3.1 77.4 0.0 76.2 1.6 46.1 2.4 36.3 0.0 66.1 12.3 34.4 14.3 34.8 0.1 77.4 1.1 54.4 2.7 57.1 0.0
MLP 0 2 99.3 0.1 99.3 0.2 97.5 0.2 99.1 0.2 97.7 0.4 79.3 2.4 89.3 0.7 63.9 1.8 62.8 1.4 74.8 5.2 60.0 6.1 64.2 0.7
MiMo 1 1 99.6 0.1 99.8 0.1 97.5 0.3 98.5 0.2 96.9 0.3 83.3 1.9 91.9 1.1 77.9 3.2 63.2 1.6 84.3 0.4 75.8 0.7 65.9 1.1
BASE 50.0 63.1 27.9 50.0 22.2 15.4 50.0 21.1 14.8 50.0 40.0 22.2
Table 4 continued

| AUROC Agg. | RETIERS | MNIST$_{2,7}$ | FMNIST$_{3,7}$ | FMNIST$_{0,1,2,3,7}$ |
|------------|----------|---------------|----------------|---------------------|
|            | AUROC ↑ | AUPR ↑ | F1 score ↑ | AUROC ↑ | AUPR ↑ | F1 score ↑ | AUROC ↑ | AUPR ↑ | F1 score ↑ | AUROC ↑ | AUPR ↑ | F1 score ↑ |
| $S_{d1}$   |          |               |               |               |               |               |               |               |               |               |               |               |               |
| DAE        | 1 0      | 99.5 ± 0.2    | 99.3 ± 0.2    | 63.3 ± 3.6    | 98.8 ± 0.5    | 98.8 ± 0.4    | 82.0 ± 6.1    | 98.1 ± 0.2    | 98.2 ± 0.2    | 81.3 ± 3.3    | 98.0 ± 0.3    | 99.0 ± 0.2    | **93.9 ± 1.2** |
| ATA        | 0 0      | 97.0 ± 0.7    | 96.2 ± 0.8    | 64.4 ± 1.1    | 98.9 ± 0.3    | 99.0 ± 0.2    | 65.4 ± 2.6    | 96.1 ± 0.2    | 96.5 ± 0.2    | 63.2 ± 1.2    | 94.7 ± 1.5    | 97.5 ± 0.7    | **76.2 ± 0.9** |
| OCA        | 3 0      | 98.6 ± 0.2    | 98.1 ± 0.4    | **75.5 ± 3.3** | 99.6 ± 0.3    | **99.2 ± 0.7** | 36.3 ± 3.3    | **99.8 ± 0.0** | **99.8 ± 0.0** | 90.1 ± 0.1    | **99.5 ± 0.1** | **99.7 ± 0.1** | **80.8 ± 1.6** |
| C2AE       | 0 1      | 98.7 ± 0.1    | 98.4 ± 0.1    | 66.7 ± 2.3    | 98.5 ± 0.4    | 97.3 ± 0.8    | 67.5 ± 2.7    | **94.0 ± 1.5** | 92.9 ± 2.0    | 68.8 ± 3.0    | 97.6 ± 0.5    | 98.6 ± 0.3    | **89.9 ± 3.4** |
| OpenMax    | 0 4      | **75.0 ± 4.9** | 70.4 ± 5.1    | 45.0 ± 0.0    | **74.3 ± 1.7** | 56.1 ± 3.5    | 53.3 ± 0.0    | **84.5 ± 16.9** | 80.0 ± 22.3  | 52.9 ± 0.1    | **83.7 ± 10.8** | 84.6 ± 10.6  | 73.7 ± 0.0 |
| SoftMax    | 0 4      | **74.4 ± 5.1** | 69.4 ± 5.4    | 45.0 ± 0.0    | **74.3 ± 1.8** | 56.1 ± 3.5    | 53.3 ± 0.0    | **71.3 ± 15.3** | 63.5 ± 20.1  | 52.9 ± 0.1    | **86.4 ± 1.6** | 83.6 ± 3.4   | 73.7 ± 0.0 |
| MLP        | 0 3      | **90.1 ± 1.8** | 68.0 ± 6.3    | 67.0 ± 1.8    | **84.7 ± 3.3** | 67.2 ± 4.2    | 58.3 ± 2.0    | 94.9 ± 1.3    | 90.3 ± 2.6    | **83.7 ± 2.3** | **66.6 ± 6.5** | 69.5 ± 6.0   | 75.7 ± 1.2 |
| MiMo       | 0 4      | **92.8 ± 2.1** | 81.9 ± 4.6    | 59.6 ± 3.1    | **93.0 ± 1.8** | 91.0 ± 1.2    | 67.5 ± 8.6    | **90.6 ± 3.5** | 87.1 ± 5.8    | 73.4 ± 4.7    | **80.9 ± 1.9** | 85.5 ± 2.8   | 77.0 ± 0.6 |
| BASE       |          | 50.0         | 29.0         | 18.4         | 50.0          | 36.4         | 21.1         | 50.0          | 35.9         | 20.9         | 50.0         | 58.3         | 26.9         |
| $S_{d2}$   |          |               |               |               |               |               |               |               |               |               |               |               |               |
| DAE        | 1 0      | **99.3 ± 0.1**| **98.9 ± 0.3**| 49.9 ± 7.3    | 97.1 ± 0.6    | 86.2 ± 3.1    | **76.5 ± 4.8**| 98.4 ± 0.2    | 97.5 ± 0.3    | 78.1 ± 3.5    | 97.8 ± 0.3    | 98.1 ± 0.2    | **86.8 ± 2.6** |
| ATA        | 1 0      | 98.1 ± 0.5    | 96.6 ± 0.8    | 66.1 ± 2.7    | **97.8 ± 0.2**| **93.0 ± 1.2**| 55.0 ± 6.2    | 96.6 ± 0.3    | 95.4 ± 0.4    | 51.0 ± 2.2    | 94.5 ± 1.6    | 96.0 ± 1.1    | 60.9 ± 1.4 |
| OCA        | 2 0      | 98.4 ± 0.3    | 97.7 ± 0.6    | **75.5 ± 3.3**| 88.7 ± 2.8    | 56.4 ± 9.5    | 32.1 ± 4.5    | **99.8 ± 0.0**| **99.6 ± 0.1**| **80.1 ± 1.0**| **99.4 ± 0.1**| **99.3 ± 0.2**| 80.7 ± 1.6 |
| C2AE       | 0 4      | 98.5 ± 0.1    | 98.0 ± 0.2    | 66.7 ± 2.3    | 97.2 ± 0.2    | 91.2 ± 0.5    | 66.7 ± 2.7    | 97.1 ± 0.5    | 93.6 ± 1.3    | 68.8 ± 3.3    | 98.6 ± 0.2    | 98.5 ± 0.3    | **89.3 ± 4.1** |
| OpenMax    | 0 4      | **85.5 ± 1.7**| 72.7 ± 3.0    | 35.5 ± 0.0    | **74.4 ± 1.4**| **35.9 ± 1.5**| 35.5 ± 0.0    | **82.1 ± 15.3**| 66.1 ± 24.9  | 35.0 ± 0.0    | **80.4 ± 10.0**| 70.2 ± 14.5  | 57.4 ± 0.0 |
| SoftMax    | 0 4      | **83.3 ± 1.7**| 72.6 ± 2.9    | 35.5 ± 0.0    | **74.4 ± 1.4**| **35.9 ± 1.5**| 35.5 ± 0.0    | **71.5 ± 13.5**| 48.9 ± 24.4  | 35.1 ± 0.1    | **81.7 ± 2.0**| 61.8 ± 3.9   | 57.4 ± 0.0 |
| MLP        | 0 3      | **84.2 ± 5.6**| 49.2 ± 16.0   | 47.5 ± 4.4    | 94.8 ± 0.7    | 76.5 ± 2.5    | 63.9 ± 1.8    | **93.8 ± 1.7**| 77.9 ± 6.3    | 70.9 ± 4.7    | **65.3 ± 3.8**| 51.5 ± 4.2   | 60.6 ± 0.6 |
| MiMo       | 0 3      | **82.1 ± 8.0**| 58.4 ± 17.2   | 39.3 ± 1.8    | 96.3 ± 0.6    | 88.2 ± 1.4    | 70.8 ± 2.4    | **93.6 ± 2.2**| 84.3 ± 5.4    | 68.2 ± 6.3    | **81.3 ± 1.5**| 76.6 ± 3.6   | 62.3 ± 0.7 |
| BASE       |          | 50.0         | 21.6         | 15.1         | 50.0          | 21.6         | 15.1         | 50.0          | 14.9         | 50.0         | 50.0         | 40.2         | 22.3         |

Score highlighting applied from Table 3.
percentage points below the best performing model. These scores are highlighted with underline within the result tables. Each results table aggregates the best and weakly performing models in the first column. Since the class base rates fluctuate significantly across datasets and splits, AUPR and F1 score as base rate-dependent metrics were not considered for the model robustness evaluation. Tables 3 and 4 present the results on datasets with a single COI and multiple COIs, respectively.

On the classification task $T_c$, represented by split $S_c$, MLP and MiMo yield the best results on all datasets. DAE and ATA provide competitive results on this task, whereas OCA, C2AE, OpenMax, and SoftMax fail on almost all $S_c$ splits in terms of AUROC.

For the contextual outlier detection task $T_o$, we see that DAE and ATA outperform MLP and MiMo in terms of AUPR and AUROC scores on the multi-COI datasets. DAE and ATA provide similar results to MiMo and MLP on the single-COI datasets. As expected, semi-supervised methods OCA, SoftMax, OpenMax, and C2AE do not achieve any performance gains.

Concerning dataset shift task $T_d$, the autoencoder-based methods yield by far the strongest results with DAE being the only method with a zero weak count and OCA providing the most top scores. In contrast, the performance of MLP and MiMo diminishes further from $T_o$ to $T_d$. The results of OpenMax, C2AE, and SoftMax improve on $T_d$ compared to $T_o$. On average, the autoencoder-based methods have a weak performance rate of 6%, whereas the remaining baselines have a weak performance rate of 80%, clearly depicting the superiority of the autoencoder-based methods on the task $T_d$.

Taking the architectural properties of each method into account, we can conclude the following: The OVR baselines MLP and MiMo require the one-vs-rest relationship to be reflected within the data, similar to the Bounding Gaussians dataset example in Fig. 1. Only in this case, the ERM objective encourages for a hull around the inlier data, which generalizes to unseen rest classes. However, with an increasing number of COI classes (see FMNIST0,1,2,3,7 results), there is no data-inherent rest preference for unseen classes, leading to poor robustness scores on the outlier detection and dataset shift task. Compliant with earlier research [28,29], MLP and MiMo reveal the most severe robustness deficiencies when facing dataset shift. While in practice OSR is often approached by subsuming all the rest samples within a single background class, our results display its insufficiency.

Conversely, semi-supervised OCA and C2AE are not able to learn the OVR relationship in the problem domain, indicating that inliers and rest samples within $T_c$ and $T_o$ are too correlated in features space. When the scope of OVR is extended to dataset shift, OCA outperforms all baselines, and C2AE becomes competitive. While the results for SoftMax and OpenMax express the same behavior, the underlying reasons are different: ERM has no intrinsic mechanism that prevents outliers from being mapped to inlier feature representations, a problem described as feature collapse that is prevented via, e.g., two-sided Lipschitz constant regularization [67]. OpenMax and SoftMax both suffer from this effect since both of them apply offline uncertainty estimates solely based on the activations. Anecdotally, we also replaced the sigmoid activation functions with RELU within OpenMax, since the network becomes piece-wise linear with possibly more expressive activations. However, this did not lead to any robustness improvements.

In contrast to aforementioned baselines, DAE and ATA do not express any robustness deficiencies. In fact, they provide competitive results on all three OSR subtasks, showing that they are able to distinguish the two OVR classes in the problem domain while being highly robust to dataset shift. Nevertheless, we find that there is still a small trade-off between accuracy and robustness, which has also been reported in previous research for various deep learning methods [30,42]. Both methods use an adversarial loss function that minimizes/maximizes inlier and rest sample reconstruction errors, respectively. Therefore, these methods resolve OCA’s issue of correlated rest and inlier samples within $T_c$ and $T_o$. Additionally, due to the bounded open space risk, they suffer less remote artifact areas that map outliers to inlier data, as seen for MLP and MiMo. While DAE and ATA are the most robust models, DAE is the best performing model in 7 cases compared to ATA, which performs best in only 3 cases. Since both methods mainly differ in terms of decision boundary estimation, the results suggest that DAE’s loss function with its BCE term and $t$ regularization has a positive effect.

The robustness properties of DAE and the baselines can also be seen in Figs. 1 and 6. DAE, ATA, MiMo, and MLP are capable of separating the red inliers from the blue rest samples, however, in a fundamentally different way. While DAE and ATA learn a representation for the inlier class, MiMo and MLP learn a separating line, which does not generalize to the unobserved orange outliers. As shown in Fig. 1h, the background class setup enables the ERM models MLP and MiMo to learn a proper hull around the inlier samples only if the observed rest samples bias the ERM toward such a decision boundary. Finally, OCA learns to separate the inliers from the orange outliers but passively minimizes the rest samples along with the inliers. This explains the poor classification performance of OCA on $T_c$ and $T_o$, but high robustness to dataset shift.

Similar conclusions on the autoencoder-based methods can be drawn from Fig. 7 which displays samples from each split of the MNIST7 and FMNIST3,7 datasets. DAE and ATA can reconstruct inliers and distort rest samples, resulting in a reconstruction error that is highly predictive of the inlierness of a sample. In contrast, OCA not only learns to
reconstruct inlier samples but also implicitly learns to reconstruct rest samples that originate from the same problem domain, explaining its low AUROC scores on $T_c$ and $T_o$. Similarly, C2AE constantly reconstructs a sample as one of the two inlier classes. The overall reconstruction quality is much lower compared to the other two approaches. This could be caused by the joint encoder/downstream layer training which aims for classification instead of reconstruction. While the autoencoder-based methods by design have a bounded open space risk, the adversarial training within DAE and ATA forces the rest samples to be in the open space, far away from the decision boundary, as indicated by the rest sample distortions. Therefore, DAE and ATA are superior in classification settings $T_c$ compared to the semi-supervised autoencoder methods.

With adversarial robustness and local stability, there are two additional crucial aspects of robustness, which can be measured by the change in model confidence after step-wise applying adversarial perturbations, as defined in Eq. (17). As shown in Fig. 8, both DAE and MLP are similarly stable when exposed to adversarially perturbed inliers of MNIST7. Regarding rest samples, there is a substantial robustness gain from $S_c$ to $S_d$ on MNIST7 with DAE being significantly more robust than MLP. On FMNIST_{3,7}, the increased diversity of inlier samples diminishes the MLP’s adversarial robustness. This supports our presumption that the MLP’s recognition function has artifact areas far from the inliers that erroneously map to the inlier class. Conversely, the increase in inlier diversity enhances the inlier adversarial robustness of DAE. Allegedly, this forces DAE to learn a more voluminous decision boundary hull that is less susceptible to inlier perturbations.

As shown in Table 5, we also investigated the well-known problem of poorly calibrated neural networks on out-of-distribution examples after being trained via ERM [24,37]. Specifically, we compared DAE to MLP and MiMo to check if the DAE architecture improves model calibration. The results indicate that the three methods are similarly well-calibrated on task $T_c$ in terms of Brier score and CECE. On tasks $T_o$ and $T_d$, we found that each method provided poor calibration with such a high variance across datasets that we omitted these inconclusive results. Instead, we analyzed the similarity of calibration between $T_c$ compared to the semi-supervised autoencoder methods.

Technically, we measured the calibration error within each bin $\text{bin-CE}_{i,j}$ [see Eq. (20)], to then calculate the average per-bin calibration error difference $\Delta \text{CE}$ between split $S_c$ and the $S_o$/$S_d$ splits. Across all three datasets, DAE has a significantly lower $\Delta \text{CE}$ compared to MLP and MiMo. This is an interesting insight, as it suggests that for DAE, calibration deficiencies on $T_c$ are more similar to deficiencies on $T_o$/$T_d$, compared to the other approaches and calibration improvement on $T_c$ could generalize to calibration improvements on $T_o$ and $T_d$. Multiple postprocessing-based approaches for calibration such as histogram binning [75] and temperature scaling [23] have
Fig. 8  Adversarial robustness comparison between DAE and MLP: for each of the splits $S_c$, $S_0$, and $S_d$ in MNIST$_7$ and FMNIST$_{3,7}$ datasets, the samples (true positives with more than 95% confidence) were perturbed via FGSM in 300 steps with a step size of 0.001 resolving the full range from local stability to global adversarial robustness. The results clearly show that DAE is more robust to adversarial perturbations applied to inliers and especially rest samples. On the two datasets, DAE is most robust w.r.t. adversarial perturbations on dataset shift samples, demonstrating the advantage of DAE being a true OSR method with bounded open space risk.

Table 5  Calibration of DAE, MLP and MiMo on split $S_c$ in terms of CECE and Brier score and average per-bin calibration error difference $\Delta CE$ between $T_r$ and $T_o/T_d$. All models are similarly well-calibrated on the classification task. The calibration similarity across splits higher for DAE compared to the other methods. Metrics and confidence are measured in %.

|                  | ATIS    | REUTERS | Newsgroups |
|------------------|---------|---------|------------|
|                  | CECE $\downarrow$ | Brier $\downarrow$ | $\Delta CE$ $\downarrow$ |
|                  | CECE $\downarrow$ | Brier $\downarrow$ | $\Delta CE$ $\downarrow$ |
|                  | CECE $\downarrow$ | Brier $\downarrow$ | $\Delta CE$ $\downarrow$ |
| DAE              | 3.1 ± 2.2 | 3.2 ± 1.9 | 24.8 ± 1.2 |
| MLP              | 2.8 ± 0.7 | 2.9 ± 0.7 | 28.6 ± 1.3 |
| MiMo             | 1.9 ± 0.6 | 2.9 ± 0.1 | 25.0 ± 1.7 |

Best scores highlighted in bold face.

Table 6  Ablation study on MNIST$_{3,7}$ w.r.t. the loss terms in $L$ controlled by hyperparameters $\lambda_i$ [see Eqs. (2) and (5)]. The respective loss histograms are shown in Fig. 9. The results clearly indicate that the combination of all loss terms yields the highest dataset shift robustness with slight degradation in classification performance. Without the adversarial loss term (i.e., $\lambda_1 = 0$), the models express significant robustness deficiencies. The two cases without $L_{REC}$ lead to unusable results as $t$ becomes 0. Note that the F1 scores can deviate from previous results in Table 4 for which the best models were selected based on AUROC scores on $S_c$. If F1 score is a concern, we suggest filtering models whose decision boundary $t$ has converged to a constant value and subsequently select the best model based on AUROC.

| Hyperparameters | Interpretation | Figures | $S_c$ AUROC | $S_0$ AUROC | $S_d$ AUROC | $S_d$ AUROC | $S_d$ AUROC | $S_d$ AUROC |
|-----------------|----------------|---------|-------------|-------------|-------------|-------------|-------------|-------------|
| $\lambda_0$    | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $S_c$ AUROC | $S_0$ AUROC | $S_d$ AUROC | $S_d$ AUROC | $S_d$ AUROC |
| 1 0.01 0.01 1   | $\hat{L}$     | Figure 9a–d | 98.9 | 97.5 | 98.9 | 96.9 | 99.0 | 98.4 | 99.0 | 97.2 |
| 1 0.01 0.01 1   | $L_R + |t|$               | Figure 9e–h | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 |
| 1 0 0 1         | inlier and $|t|$ minimization | Figure 9f–i | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 | 0.5 | 0 |
| 0 0 1 0         | $L_{REC}$     | Figure 9m–p | 99.2 | 97.1 | 98.2 | 91.7 | 71.0 | 66.9 | 91.7 | 76.3 |
| 0 0 1 1         | $L_{REC} + |t|$    | Figure 9q–t | $99.7$ | $97.9$ | 98.3 | 92.5 | 89.1 | 77.9 | 96.9 | 85.0 |

Best scores highlighted in bold face
Fig. 9 Inlier and outlier loss histograms regarding the different cases within the ablation study in Table 6. The model output $g(\text{emSE})$ mapping reconstruction errors to probabilities is plotted in green. The decision boundary $t$ is plotted as dotted red line. Note that in some cases, e.g., in Fig. 9a–d, the maximum reconstruction errors are capped to highlight the accuracy of the decision boundary. The results show that the combination of loss terms $\hat{L}$ leads to inliers and outliers being minimized/maximized, respectively, and the decision boundary being as close as possible to the inliers to maximize robustness without jeopardizing classification performance. All the other combinations lead to poor classification performance or lack of outlier robustness (color figure online).
been proposed which could exploit the calibration similarity across tasks.

As shown in Table 6, we performed an ablation study w.r.t. the different loss terms in $\hat{L}$ to show that only the specific combination in $\hat{L}$ leads to the desired classification and robustness properties. The results clearly show that the loss function $\hat{L}$ has the highest robustness with a minor classification degradation. If we remove the classification loss term $L_{BCE}$ from $\hat{L}$, then the decision boundary $t$ converges to 0 which classifies all samples as outliers irrespective of their true class. If the model is solely trained on $L_{BCE}$ or $L_{BCE} + |t|$, then the classification performance increases slightly, however, accompanied by severe robustness deficiencies to outliers and dataset shift.

These findings can be explained by Fig. 9 which plots the loss histograms for each hyperparameter combination in Table 6. The strong robustness performance of $\hat{L}$ can be attributed to the wide inlier/outlier separation and small decision boundary $t$ which allows to reject outliers effectively, as shown in Fig. 9a–d. Interestingly, $L_{BCE}$ (Fig. 9m–p) and $L_{BCE} + |t|$ (Fig. 9q–t) can separate inliers from rest classes on $S_c$ but fail to generalize to unseen classes. Especially without $|t|$ regularization, the inlier reconstruction errors are less minimized, leading to dataset shift samples becoming indistinctive from inliers (see Fig. 9o). If $L_{BCE}$ is jointly optimized with $|t|$ regularization, then the minimization of inliers improves but outliers get less maximized due to the vanishing gradient problem, as derived in Eq. (8), resulting in poor OOD data robustness (see Fig. 9s). The adversarial component within $L_{B2}$ does not suffer from the vanishing gradient limitation and enforces the maximization of inliers which becomes apparent when comparing Figs. 9a–9q. If the adversarial component is missing in $L_R$, then only inliers are minimized which causes a significant overlap of inliers and rest classes especially on $S_c$ (see Fig. 9i). In conclusion, the combination of all loss terms in $\hat{L}$ yields the best separation of inliers and outliers due to effective minimization of inliers and maximization of outliers and solves the vanishing gradient problems. Moreover, the minimization of the decision boundary $t$ with $L_{BCE}$ acting as an antipole enables the model to robustly reject outliers without jeopardizing classification performance.

From this extensive analysis, we can conclude that DAE, as an OSR method with a bounded open space risk, clearly shows its superiority compared to the potent baselines ranging from OSR, OVR to outlier detection methods. Apart from ATA, every baseline consistently failed at more than one of the three subtasks of OSR, questioning their applicability in safety-critical systems. The consistent classification performance across all three tasks $T_c$, $T_o$ and $T_d$ combined with an increased (adversarial) robustness shows the benefits of DAE’s reduced and bounded open space risk and exposes the deficiencies of the ERM and semi-supervised baselines.

8 Conclusion

Open set recognition (OSR) is a common setup in machine learning applications. Whenever the objective is to distinguish at least one class of interest (COI) from all remaining, possibly unknown classes (RC), e.g., ordinary internet traffic from novel intrusion attempts or general discussions from all types of hate speech, OSR methods seem natural. Our analysis revealed that when extending the scope of RC, OSR poses severe challenges of outlier detection and dataset shift to deep neural networks solely optimized via empirical risk minimization. We provide an effective solution to these deficiencies with our proposed decoupling autoencoder (DAE) architecture. We proved the existence of a bounded open space risk for DAE and supported its classification and (adversarial) robustness benefits across three different sub-tasks of OSR. Specifically, we benchmarked DAE against capable baselines from various domains (DNNs, ensemble methods, outlier detection, and OSR) w.r.t. the OSR subtasks of classification, outlier detection, and dataset shift. DAE showed superior robustness across all subtasks throughout all experiments compared to the baselines, which failed on at least one of the tasks apart from ATA. In comparison with ATA, DAE can actively minimize the open space risk and does not require an offline brute-force line search for decision boundary estimation.

For future work, it would be interesting to extend DAE toward multi-class classification with a bounded open space risk, which would allow for robust multi-class classification under extreme dataset shift conditions. Finally, a promising idea could be the development of feature extractors that prevent the model from learning representations of noisy or uninformative features, thereby further alleviating the trade-off between classification and robustness performance.

Acknowledgements In parts, the authors of this work were supported by the Fraunhofer Research Center for Machine Learning (RCML) within the Fraunhofer Cluster of Excellence Cognitive Internet Technologies (CCIT) and by the Competence Center for Machine Learning Rhine Ruhr (ML2R) which is funded by the Federal Ministry of Education and Research of Germany (grant no. 01IS18038B).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indi-
cated if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Aggarwal, C.C.: Outlier analysis. In: Data Mining. Springer (2015)
2. Aggarwal, C.C., Sathe, S.: Outlier Ensembles: An Introduction. Springer (2017)
3. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D.: Concrete Problems in AI Safety, Paul Christiano (2016)
4. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1563–1572 (2016)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
6. Boult, T.E., Cruz, S., Dhamija, A.R., Gunther, M., Henrydoss, J., Scheirer, W.J.: Learning and the unknown: surveying steps toward open world recognition. Proc. AAAI Conf. Artif. Intell. 33, 9801–9807 (2019)
7. Brier, G.W.: Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78(1), 1–3 (1950)
8. Chalapathy, R., Menon, A.K., Chawla, S.: Anomaly detection using one-class neural networks. (2018). arXiv:1802.06360
9. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
10. Dau, H.A., Ciesielski, V., Song, A.: Anomaly detection using replicator neural networks trained on examples of one class. In: Proc. of the 10th Int. Conf. on Simulated Evolution and Learning (2014)
11. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)
12. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04665 (2018)
13. Domingues, R.: Probabilistic modeling for novelty detection with applications to fraud identification. arXiv preprint arXiv:1903.01730 (2019)
14. Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recogn. 58, 121–134 (2016)
15. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)
16. Fusilier, D.H., Cabrera, R.G., Montes, M., Rosso, P.: Using pu-learning to detect deceptive opinion spam. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 38–45 (2013)
17. Geng, C., Huang, S., Chen, S.: Recent advances in open set recognition: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 43(10), 3614–3631 (2020)
18. Gharib, M., Mohammadi, B., Dastgerdi, S.H., Sabokrou, M.: AutoTDS: auto-encoder based method for intrusion detection system. arXiv preprint arXiv:1911.03306 (2019)
19. Girshick, R.: Fast r-CNN. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
22. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
23. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR (2017)
24. Havasi, M., Jenatton, R., Fort, S., Liu, J.Z., Snoek, J., Lakshminarayanan, B., Dai, A.M., Tran, D.: Training independent subnetworks for robust prediction. In: International Conference on Learning Representations (2021)
25. Hawkins, D.M.: Identification of Outliers, vol. 11. Springer (1980)
26. Hawkins, S., He, H., Williams, G., Baxter, R.: Outlier detection using replicator neural networks. In: Proc. of Int. Conf. on Data Warehousing and Knowledge Discovery (2002)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
28. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings (2017)
29. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure (2019)
30. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
31. Hsieh, C.J., Natarajan, N., Dhillon, I.: Pu learning for matrix completion. In: International Conference on Machine Learning, pp. 2445–2453. PMLR (2015)
32. Iizuka, N., et al.: Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet 361, 923–929 (2003)
33. Kaynar, O., Yüksel, A.G., Görmüz, Y., Isik, Y.E.: Intrusion detection with autoencoder based deep learning machine. In: Proc. of 25th Signal Processing and Communications Applications Conference (SIU) (2017)
34. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015)
35. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. (2017)
36. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2019)
37. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 30 (2017)
38. LeCun, Y.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
39. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
40. Li, X., Roth, D.: Learning question classifiers. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)
41. Linden, A., Kindermann, J.: Inversion of multilayer nets. In: Proc. Int. Joint Conf. Neural Networks, vol. 2, pp. 425–430 (1989)
42. Lopes, R.G., Yin, D., Poole, B., Gilmer, J., Cubuk, E.D.: Improving robustness without sacrificing accuracy with patch Gaussian augmentation. arXiv preprint arXiv:1906.02611 (2019)

43. Lübbering, M., Gebauer, M., Ramamurthy, R., Bauckhage, C., Sifa, R.: Decoupling autoencoders for robust one-vs-rest classification. In 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. IEEE (2021)

44. Lübbering, M., Gebauer, M., Ramamurthy, R., Sifa, R., Bauckhage, C.: Supervised autoencoder variants for end to end anomaly detection. In: Pattern Recognition, ICPR International Workshops and Challenges (2021)

45. Lübbering, M., Ramamurthy, R., Gebauer, M., Bell, T., Sifa, R. and Bauckhage, C.: From imbalanced classification to supervised outlier detection problems: adversarially trained auto encoders. In: Artificial Neural Networks and Machine Learning—ICANN 2020 (2020)

46. Lübbering, M., Gebauer, M., Ramamurthy, R., Pielka, M., Bauckhage, C., Sifa, R. Utilizing representation learning for robust text classification under dataset shift. In: Proceedings of the Conference “Lernen, Wissen, Daten, Analysen”, CEUR Workshop Proceedings (in press) (2021)

47. Lübbering, M., Pielka, M., Das, K., Gebauer, M., Ramamurthy, R., Bauckhage, C., Sifa, R.: Toxicity detection in online comments with limited data: a comparative analysis. In: ESANN (in press) (2021)

48. Moya, M.M., Koch, M.W., Hostetler, L.D.: One-class classifier networks for target recognition applications. NASA STI/Recon. Tech. Rep. N. 93, 24043 (1993)

49. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2015)

50. Nedelkoski, S., Cardoso, J., Kao, O.: Anomaly detection and classification using distributed tracing and deep learning. In: 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (2019)

51. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks Are easily fooled: high confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

52. Oza, P., Patel, V.M.: C2ae: class conditioned auto-encoder for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2307–2316 (2019)

53. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis)contents: a survey of dataset development and use in machine learning research. ArXiv (2020)

54. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)

55. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010)

56. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)

57. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28, 91–99 (2015)

58. Rosza, A., Boul, T.E.: Improved adversarial robustness by reducing open space risk via tent activations. arXiv preprint arXiv:1908.02435 (2019)

59. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boul, T.E.: The extreme value machine. IEEE Trans. Pattern Anal. Mach. Intell. 40(3), 762–768 (2017)

60. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 4393–4402. PMLR (2018)

61. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3), e0118432 (2015)

62. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boul, T.E.: Toward open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35(7), 1757–1772 (2012)

63. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boul, T.E.: Toward open set recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35(7), 1757–1772 (2013)

64. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide Marker discovery. In: Proc. Int. Conf. on Information Processing in Medical Imaging (2017)

65. Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In: Proc. of Advances in Neural Information Processing Systems (2000)

66. Tanielian, U., Vashil, F.: Relaxed softmax for pu learning. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 119–127 (2019)

67. Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119. PMLR (2020)

68. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 7, 1–8 (2006)

69. Wang, N., Chen, C., Xie, Y., Ma, L.: Brain tumor anomaly detection via latent regularized adversarial network. CoRR, arXiv:2007.04734 (2020)

70. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)

71. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015)

72. Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y. and Chen, J.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In: Proceedings of the 2018 World Wide Web Conference (2018)

73. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems (2019)

74. Yu, Fei, Qin, Z., Liu, C., Zhao, L., Wang, Y., Chen, X.: Interpreting and evaluating neural network robustness. arXiv preprint arXiv:1905.04270 (2019)

75. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Icml, vol. 1, pp. 609–616. Citeseer (2001)

76. Zou, J., Zhang, J. and Jiang, P.: Credit card fraud detection using autoencoder neural network. arXiv preprint arXiv:1908.11553 (2019)