Friedrich-Alexander-Universität
Department Artificial Intelligence
in Biomedical Engineering

IDEA Lab
Image Data Exploration
and Analysis Lab

Deep One-Class Classification

Lukas Ruff[1] Robert A. Vandermeulen[2] Nico Görnitz[3] Lucas Deecke[4] Shoaib A. Siddiqui[2 5]
Alexander Binder[6] Emmanuel Müller[1] Marius Kloft[2]

A Critical Scientific Review

By – Mirza Shaheen Iqubal
Matriculation Number - 22998316

Presentation Outline

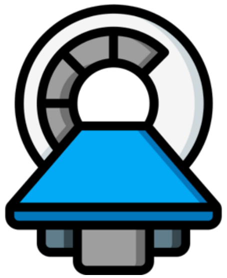
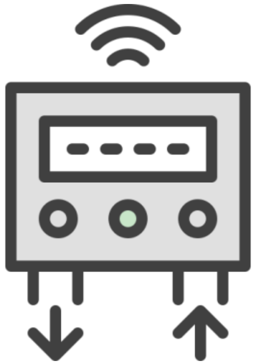
1. Motivation
2. Prior Art
3. The Deep SVDD Paradigm
4. Connecting Theory to Experimental Design
5. Empirical Evaluation & In-Depth Results Analysis
6. Critique of the Reference Implementation
7. Conclusion & Future Research Directions

Motivation: The Anomaly Detection Problem

Definition: Identifying samples that do not conform to expected patterns

Paradigm: Unsupervised learning with clean training data

Applications:


The Sub-Optimality of Reconstruction-Based Deep AD

Challenge: Most deep AD methods use reconstruction-based objectives

The Flaw: Autoencoders may generalize well and reconstruct anomalies

Missing Pressure: No explicit penalty for reconstructing anomalies

Result: Misalignment between training objective and anomaly detection goal

Prior Art: Kernel Methods vs. Deep Approaches

Kernel SVDD: Finds minimal-volume hypersphere in RKHS

- Limitation: Kernel choice, feature engineering, $O(n^2)$ scaling

Deep Autoencoders: Dominant deep approach

- Limitation: Compactness is indirect, difficult hyperparameter tuning

The Deep SVDD Paradigm: A Shift in Objective I

Deep Support Vector Data Description (Deep SVDD)

Innovation: From reconstructive to compressive objective

Core Principle: Compress normal data representations into minimal-volume hypersphere

Mechanism: Network learns to discard intra-class variance

Advantage: Direct pressure on representation compactness

The Deep SVDD Paradigm: A Shift in Objective II

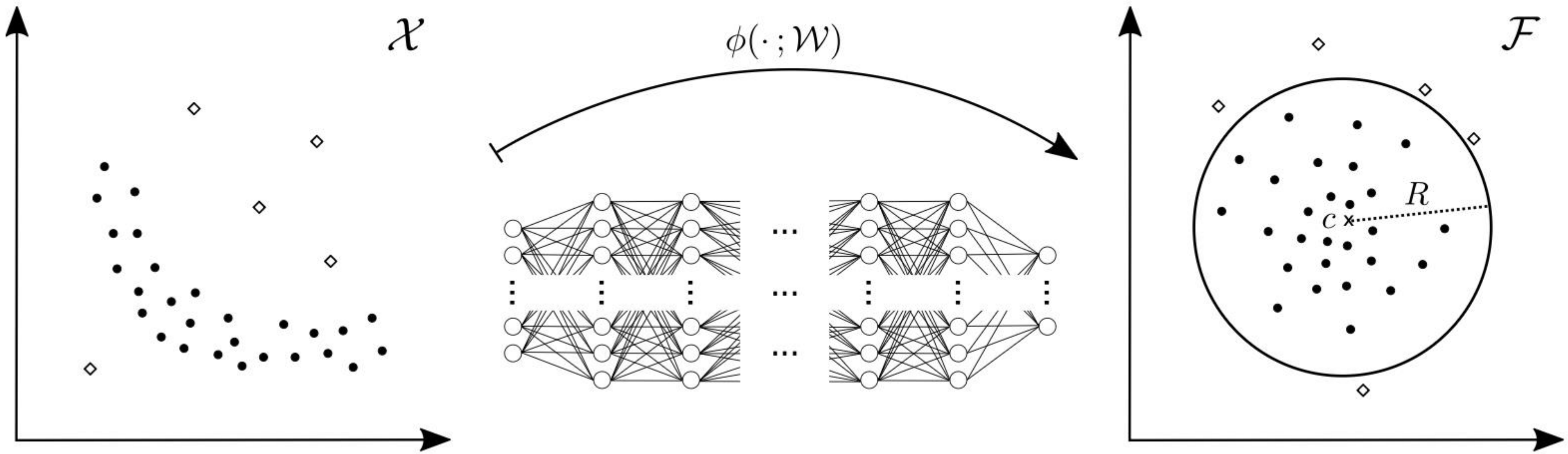


Figure1: Deep Support Vector Data Description (Deep SVDD)

Mathematical Formulation I: Soft-Boundary Deep SVDD

Objective Function:

$$\min_{R,\mathcal{W}} \quad R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\max\{0, \|\phi(\boldsymbol{x}_i;\mathcal{W}) - \mathbf{c}\|^2 - R^2\}$$
$$+ \frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^{\ell}\|_F^2.$$

R^2: Directly minimizes hypersphere volume

Hinge-Loss Term: Penalizes samples outside the sphere

v: Trade-off hyperparameter (upper bound on outlier fraction)

Mathematical Formulation II: One-Class Deep SVDD

For a "clean" training set, the objective function simplifies to:

$$\min_{\mathcal{W}} \quad \frac{1}{n}\sum_{i=1}^{n}\|\phi(\boldsymbol{x}_i;\mathcal{W})-\boldsymbol{c}\|^2+\frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^{\ell}\|_F^2.$$

Goal: Minimize mean squared distance from center c

Interpretation: Find 'center of mass' in learned feature space

Anomaly Score: $s(x) = ||\phi(x; W^*) - c||^2$ (simple and efficient)

Critical Theoretical Constraints

Proposition 1: Center c must be fixed (prevent co-adaptation)

Proposition 2: No bias terms (prevent trivial constant function)

Proposition 3: Unbounded activations required (ReLU, not tanh)

Purpose: Prevent 'hypersphere collapse' (trivial solution)

Critical Link: From Theory to Experimental Design

Architecture: LeNet-style CNN with no bias, ReLU activations

Pre-training: Autoencoder initialization ensures meaningful starting point

Inductive Bias: CNN's local feature bias is crucial for interpretation

Result: Theory directly informs practical implementation

Empirical Evaluation: Setup & Baselines

Datasets:

1. **MNIST & CIFAR-10**
2. **GTSRB**

Competing Methods:

Shallow**: Kernel SVDD, Kernel Density Estimation (KDE), Isolation Forest (IF).**

Deep**: Deep Convolutional Autoencoder (DCAE), AnoGAN.**

Metric**: Area Under the Receiver Operating Characteristic Curve (AUC), averaged over 10 seeds.**

Detailed Results: MNIST and CIFAR-10

Figure 2. Most normal (left) and most anomalous (right) in-class examples determined by One-Class Deep SVDD for selected MNIST (top) and CIFAR-10 (bottom) one-class experiments.

Detailed Results: MNIST and CIFAR-10

Table 1. Average AUCs in % with StdDevs (over 10 seeds) per method and one-class experiment on MNIST and CIFAR-10.

NORMAL CLASS	OC-SVM/ SVDD	KDE	IF	DCAE	ANOGAN	SOFT-BOUND. DEEP SVDD	ONE-CLASS DEEP SVDD
0	**98.6**±0.0	97.1±0.0	98.0±0.3	97.6±0.7	96.6±1.3	97.8±0.7	98.0±0.7
1	99.5±0.0	98.9±0.0	97.3±0.4	98.3±0.6	99.2±0.6	99.6±0.1	**99.7**±0.1
2	82.5±0.1	79.0±0.0	88.6±0.5	85.4±2.4	85.0±2.9	89.5±1.2	**91.7**±0.8
3	88.1±0.0	86.2±0.0	89.9±0.4	86.7±0.9	88.7±2.1	90.3±2.1	**91.9**±1.5
4	**94.9**±0.0	87.9±0.0	92.7±0.6	86.5±2.0	89.4±1.3	93.8±1.5	**94.9**±0.8
5	77.1±0.0	73.8±0.0	85.5±0.8	78.2±2.7	88.3±2.9	85.8±2.5	**88.5**±0.9
6	96.5±0.0	87.6±0.0	95.6±0.3	94.6±0.5	94.7±2.7	98.0±0.4	**98.3**±0.5
7	93.7±0.0	91.4±0.0	92.0±0.4	92.3±1.0	93.5±1.8	92.7±1.4	**94.6**±0.9
8	88.9±0.0	79.2±0.0	89.9±0.4	86.5±1.6	84.9±2.1	92.9±1.4	**93.9**±1.6
9	93.1±0.0	88.2±0.0	93.5±0.3	90.4±1.8	92.4±1.1	94.9±0.6	**96.5**±0.3
AIRPLANE	61.6±0.9	61.2±0.0	60.1±0.7	59.1±5.1	**67.1**±2.5	61.7±4.2	61.7±4.1
AUTOMOBILE	63.8±0.6	64.0±0.0	50.8±0.6	57.4±2.9	54.7±3.4	64.8±1.4	**65.9**±2.1
BIRD	50.0±0.5	50.1±0.0	49.2±0.4	48.9±2.4	**52.9**±3.0	49.5±1.4	50.8±0.8
CAT	55.9±1.3	56.4±0.0	55.1±0.4	58.4±1.2	54.5±1.9	56.0±1.1	**59.1**±1.4
DEER	66.0±0.7	**66.2**±0.0	49.8±0.4	54.0±1.3	65.1±3.2	59.1±1.1	60.9±1.1
DOG	62.4±0.8	62.4±0.0	58.5±0.4	62.2±1.8	60.3±2.6	62.1±2.4	**65.7**±2.5
FROG	74.7±0.3	**74.9**±0.0	42.9±0.6	51.2±5.2	58.5±1.4	67.8±2.4	67.7±2.6
HORSE	62.6±0.6	62.6±0.0	55.1±0.7	58.6±2.9	62.5±0.8	65.2±1.0	**67.3**±0.9
SHIP	74.9±0.4	75.1±0.0	74.2±0.6	**76.8**±1.4	75.8±4.1	75.6±1.7	75.9±1.2
TRUCK	75.9±0.3	**76.0**±0.0	58.9±0.7	67.3±3.0	66.5±2.8	71.0±1.1	73.1±1.2

Detailed Results: GTSRB

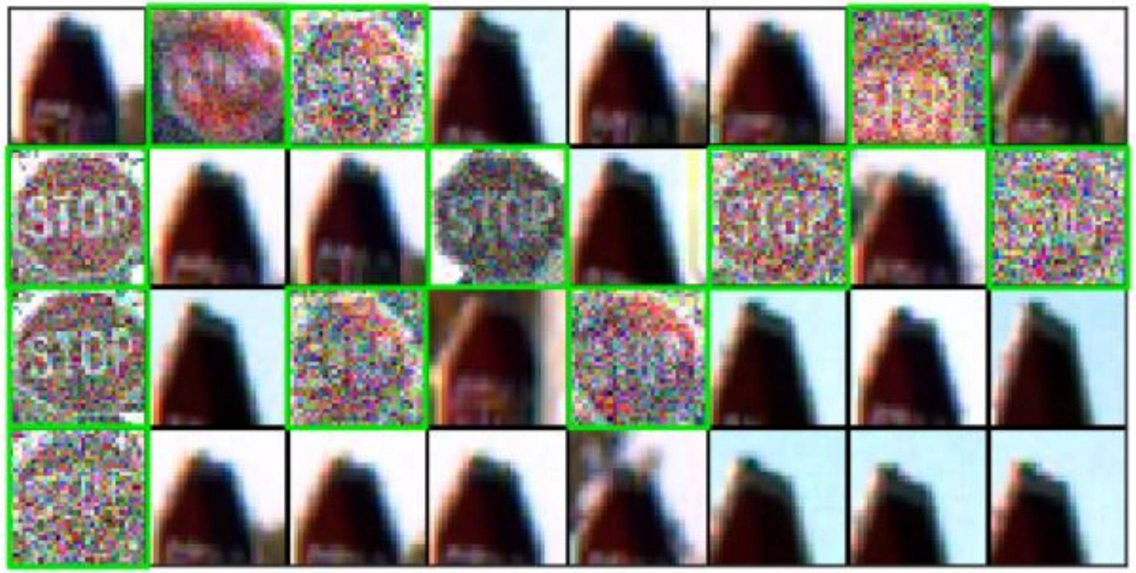


Figure 4. Most anomalous stop signs detected by One-Class Deep SVDD. Adversarial examples are highlighted in green.

Detailed Results: GTSRB

Table 2. Average AUCs in % with StdDevs (over 10 seeds) per method on GTSRB stop signs with adversarial attacks.

METHOD	AUC
OC-SVM/SVDD	67.5 ± 1.2
KDE	60.5 ± 1.7
IF	73.8 ± 0.9
DCAE	79.1 ± 3.0
ANOGAN	–
SOFT-BOUND. DEEP SVDD	77.8 ± 4.9
ONE-CLASS DEEP SVDD	$\mathbf{80.3} \pm 2.8$

Adversarial & Qualitative Findings

Adversarial Detection (GTSRB): Deep SVDD achieves 80.3% AUC

- Outperforms DCAE (79.1%)
- Tight boundary detects malicious perturbations

Qualitative Analysis: Model identifies unusual in-class samples

- Learns meaningful, prototype-centric representation

Critique of the Original Code

Issue 1: Lack of version pinning in requirements.txt

Issue 2: No automated test suite for verification

Issue 3: Monolithic, tightly coupled to specific datasets

Impact: Reproducibility challenges and limited extensibility

Proposed Enhancements

1. **Generalize Dataset Handling**.
2. **Implement Advanced Objectives**: Such as a **hybrid objective**

Hybrid Deep SVDD: Loss = μ1 * SVDD_Loss + μ2 * Reconstruction_Loss

1. **Integrate Model Explainability like Grad-CAM**.
2. **Bolster Robustness Testing approach** like adding data augmentation capabilities, such as injecting **Gaussian noise.**

Conclusion & Future Research Directions

Conclusion: Deep SVDD provides theoretically-grounded paradigm

Strengths: End-to-end learning, superior performance on MNIST

Limitations: Architectural sensitivity, pre-training dependency

Future Work: Non-image modalities, advanced self-supervised learning

Thank You

Open for Discussion