Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Integrating Statistical Depth Functions with Deep Learning for Explainable Multivariate Outlier Detection

Hedeel Kamil Habeeb

Faculty of Nursing, University of Al-Qadisiya, Al-Qadisiyah, Iraq. Email: hadeel.kamil@qu.edu.iq.

ARTICLE INFO

Article history:
Received: 26 /04/2026
Rrevised form: 15/05/2026
Accepted : 17/06/2026
Available online: 30 /06/2026

Keywords:

Statistical Depth Functions; Deep Learning; Multivariate Outlier Detection; Explainable AI; Anomaly Detection; Tukey Depth; Mahalanobis Distance; Robust Statistics; Feature Augmentation; Neural Networks.

ABSTRACT

The process of identifying outliers in multiple variables stays as a major obstacle for data-driven modeling because researchers must handle expanding data dimensions and growing data complexities. The paper presents a new framework called Statistical Outlier Detection with Depth (SODD) which combines Statistical Depth Functions (SDFs) with Deep Learning systems to create an explainable multivariate outlier detection system that shows strong performance. The statistical depth functions establish a formal system which determines how far multivariate data points exist from their center point while maintaining their resistance to extreme values and their ability to show geometric characteristics. The proposed scheme utilizes a combination of four major approaches where depth scores are incorporated within deep neural networks to generate depth-enhanced feature extraction processes, depth-modulated losses, depth-guided regularizations, and depth-dependent architectures that generate better detection results along with higher levels of interpretability in models. The SODD framework utilizes the application of Tukey depth, Mahalanobis depth, Projection Depth, and Spatial depth in order to measure the extent to which an observation is an outlier under different data distributions. Experiments have been carried out on synthetic data as well as some standard real-world datasets, and the framework was implemented through a Python development environment. The results obtained are quite promising, with scores for AUC, Precision, and Recall of 0.94, 0.89, and 0.87, respectively, making the model perform comparably well relative to the baselines examined. Moreover, the inclusion of depth-based explanations enhances the interpretability aspect of the model.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.22960

1.Introduction

A core challenge of data-driven modelling is learning an accurate representation of the target variable of interest. Disentangling outliers from regular observations, and reconstructing observations from corrupted inputs, remain two classics yet widely-studied tasks across numerous fields. Outlier detection, for example, refers to identifying a set of unmixed components given only a collection of observations, whereas anomaly detection refers to recovering original observations from faulty ones in the presence of noise or occlusions. Given the specific and distinct information contained in outlier, missing or corrupted observations, the effective modelling of such observations, together with the incorporation of appropriate domain knowledge to accurately estimate their representative structure, is essential for successful completion of either task [1].

An integrated setting is then proposed, where statistical depth functions are incorporated into the learning of the above tasks in the deep-learning paradigm; the objective of this integration is two-fold, to address the robustness and

∗Corresponding author: *Hedeel Kamil Habeeb*

Email addresses: *hadeel.kamil@qu.edu.iq*

Communicated by ‘sub etitor’

explainability of the model. Statistical depth functions quantify the centrality or location of a multivariate observation with respect to a specified probability distribution, mathematically facilitating both outlier and anomaly detection problems in conjunction with a properly chosen modelling assumption. Deep-learning methods naturally possess a high representation capacity and therefore the ability to recover these blind-source separation or inpainting problems. However, models that leverage depth functions in an integrated manner still remain scarce, and thus a consolidated setup is framed that combines them for related tasks. By introducing statistical depth functions at suitable locations during model design and training, the robustness of deep networks is enhanced further [2].

To illustrate an example, it is demonstrated both theoretically and experimentally how, under mild conditions, the incorporation of a depth-based prior yields a model with improved resistance to outliers. The specific detection of outliers is then addressed, considering statistical depth functions as a guiding principle throughout model formulation to improve the robustness against contaminated samples. Integration of depth functions into deep models also delivers a valuable notion of explanatory power, as it enables the quantification of the position of a given sample relative to the learned marginal distribution, with low scores indicating an outlier trajectory. By employing adequate explainability strategies that complement depth functions, an integrated setup also emerges that aligns with contemporary aims for interpretable machine-learning [3].

1.1 Motivation and Problem Statement

Multivariate outlier detection identifies observations that deviate significantly from a specific data distribution or overall pattern. Outliers may adversely affect model performance or indicate erroneous or anomalous observations. Although various models can describe the majority of observations, the enormous representational power and flexibility of deep-learning-based approaches make them particularly attractive for their application; nevertheless, outlier detection in deep learning remains a challenging problem, not least because of the associated difficulty in constructing suitable metrics for outlierness, a crucial component of any multivariate outlier-detection approach. Existing solutions range from specific evaluation metrics tailored to outlier detection to procedures that decouple outlier detection and scoring. Statistical Depth Functions (SDFs) provide global (in contrast to local) multivariate-outlier detection capabilities that complement existing deep-learning-based approaches by naturally addressing the associated mathematical formalisation [1]. Their robustness to outliers [4] allows them to satisfy the centrality hypothesis—namely, that outliers are observations whose distribution differs from the overall pattern—posited by many of the deepest networks, while improving the ability to generalise to unseen data. SDFs also enhance the explainability of models, offering global interpretations in terms of the relative depth or outlierness of each input and clarifying why the model as a whole perceives the observation as anomalous [2]. Despite being good ways of measuring centrality and outlierness, statistical depth functions lack the capacity for dealing with complicated nonlinear relationships within data that is multi-dimensional. Deep learning offers statistical depth functions by virtue of their capability to learn and extract features automatically, whereas depth functions improve robustness and interpretability. It is for this reason that the SODD framework was proposed.

1.2 Contributions and Scope

Outlier detection protects applications, ranging from fraud detection and manufacturing to health monitoring and financial data analysis. However, high data dimensionality hampers model calibration and outlier characterization. Multivariate outliers exhibit deviation along multiple explanatory variables simultaneously, instead of univariate outlier-like behaviour along isolated dimensions [1]. Evaluation adopts either benchmarking tools for detection models [3] or targeted explanations of non-standard behaviours.

Statistical Depth Functions (SDFs) characterise the location of a multivariate observation relative to the data distribution. A central observation possesses the largest depth value, while atypical observations are assigned low depths. SDFs constitute a key multivariate-outlier modelling methodology that quantifies the centrality of observations within multidimensional distributions. When data follow a given distribution, observations with lower depth values are more likely to exhibit outlier-like behaviour. Such SDFs withstand extensive disturbance without modifying the corresponding ranking among input observations. A model thus receives input data along with pre-computed depth scores. Subsequent transformed data conserve depth information; whenever datasets develop fresh observations, included depth scores retain their utility. Deep learning underpins many anomaly or outlier detection methods across data and disciplines. Unsupervised outlier-detection methodologies analyse the input data without supervision. The most popular unsupervised anomaly-detection mechanisms model the data through normality or density estimation.

Explainable Artificial Intelligence (XAI) devolves techniques that elucidate model formulations, behaviour, and predictions. With machine-learning applications, practitioners often seek clarity regarding the reasoning leading to specific predictions or choices. Such understanding facilitates the selection of suitable models and quantification of associated trust levels. Explanation methodologies compartmentalise into local, attributions, and global strategies. The SDF modelling strategy favors non-stringent robustness, which implies that observations considered appropriate will still be informative when there are no more predictions from the model.

2. Background

The task of outlier detection, defined as observations that significantly differ from the rest of the data, is essential for several practical use cases. The application of outlier detection techniques in the field of financial transactions could help auditors spot any cases of fraud, while outlier detection in the health status of patients whose health profiles differ from the health trajectory norm found in the clinical datasets could prompt early intervention [4]. Multivariate analysis presents difficulties as the data are in vector form in n-dimensional space. The datasets possess features like high dimensions and nonlinearity; meanwhile, there are very few metrics that are capable of handling such data [5].

The above-mentioned scores give a non-negative value for each individual sample, denoting how far off from the normal range the data point is. These scores have other properties as well, such as ordering property, centrality, openness and closedness map, and robustness. Robustness of such measures is usually defined by means of breakdown points, representing the largest fraction of outliers a technique can withstand without compromising accuracy [3].

Anomaly/outlier detection has become one of the prominent techniques using deep learning methods. Current methods use scalar anomaly scores rather than vectorial representations while sophisticated tensor models are less investigated. With the help of existing integration techniques and statistical depth theory, it would be feasible to integrate the statistical depth function into the deep learning framework. A number of approaches could be utilized, such as the use of depth information for feature extraction, depth embedding of latent space representation, and joint loss function integration of depth and scalar scores. The goal of Explainable Artificial Intelligence (XAI) is to increase the transparency of autonomous agents through the provision of comprehensible interpretations of the decisions made by their models. An explanation system may work globally, in which case the model is examined as a whole, or locally, in which case individual explanations for predictions can be produced. The contribution of statistical techniques towards the explainability of artificial intelligence models lies in the provision of insights into data that are self-explanatory, independent of the algorithm's structure.

2.1 Multivariate Outlier Detection

Detection of multivariate outliers aims at identifying observations not conforming to anticipated statistical and functional principles. As data volumes surge and dimensions increase, real-world datasets more frequently elude conventional univariate analysis and structural hypotheses such as independence, stationarity, and linearity. High-dimensional outlier detection is complicated by the presence of irrelevant dimensions masking the influence of the relevant ones and the well-documented phenomenon of concentration of measure that diminishes the amount of information conveyed by sparse observations [3]. Standard measures employed in univariate detection are ill-defined in the multivariate case, which complicates the transfer of one-dimensional methods across the boundary.

Tukey's proposal of depth functions endeavoured to address the need for statistical multivariate ordering. A statistical depth function D defines a normative centrality ordering among data points by mapping observations in a d-dimensional space R^d to real numbers, ranking each according to their depth with respect to distribution F. Depth functions also allow the definition of outlyingness as a natural complement, denoting points whose depth values are significantly below the central value. Multivariate depth concepts can thus facilitate the transfer of outlier analysis to multidimensional observations.

2.2 Statistical Depth Functions

Statistical depth functions provide a general framework for ranking multivariate observations according to their centrality relative to a specified target distribution. These scores enable the characterization of observations in terms of their degree of centrality and outlyingness, allowing the categorization of univariate, multivariate, and functional outlier types. A desirable attribute of statistical depth functions is robustness to perturbations in the data-generating

mechanism, namely, resistance to the influence of outliers. The measures of robustness include the value of the breakdown point, and the higher its value is, the more resistant it will be to outliers. These measures imply that there is consistency between statistical depth functions and robustness as one of the major criteria when selecting a function for outlier detection, where these functions can be considered to be the heart of the point-wise detection approach. Also, the geometrical and probabilistic interpretation of many well-known statistical depth functions make explainability possible through the information about signals and model behaviors, as a part of explainability in deep learning [6]; [3].

Table 1. Comparison of Statistical Depth Functions.

Depth Function	Robustness	Computational Complexity	Explainability	Scalability	Distribution Assumption
Tukey (Halfspace) Depth	Very High (BP ≈ 0.5)	O(n^d) / O(nk) approx.	Excellent (Geometric)	Moderate (d ≤ 20)	None (Non-parametric)
Mahalanobis Depth	Moderate (BP ≈ 0)	$O(n + d^3)$	Good (Elliptical)	Excellent (Any d)	Elliptical (Gaussian-like)
Projection Depth	High (BP ≈ 0.5)	$O(n^2\|S\|^2)$	Good (Linear proj.)	Low-Med (d ≤ 10)	None (Non-parametric)
Spatial Depth	High (BP ≈ 0.5)	O(nd)	Moderate (Metric-based)	Good (d ≤ 50)	None (Non-parametric)
Simplicial Depth	Very High (BP ≈ 0.5)	O(n^{d+1})	Moderate (Combinatorial)	Low (d ≤ 5)	None (Non-parametric)

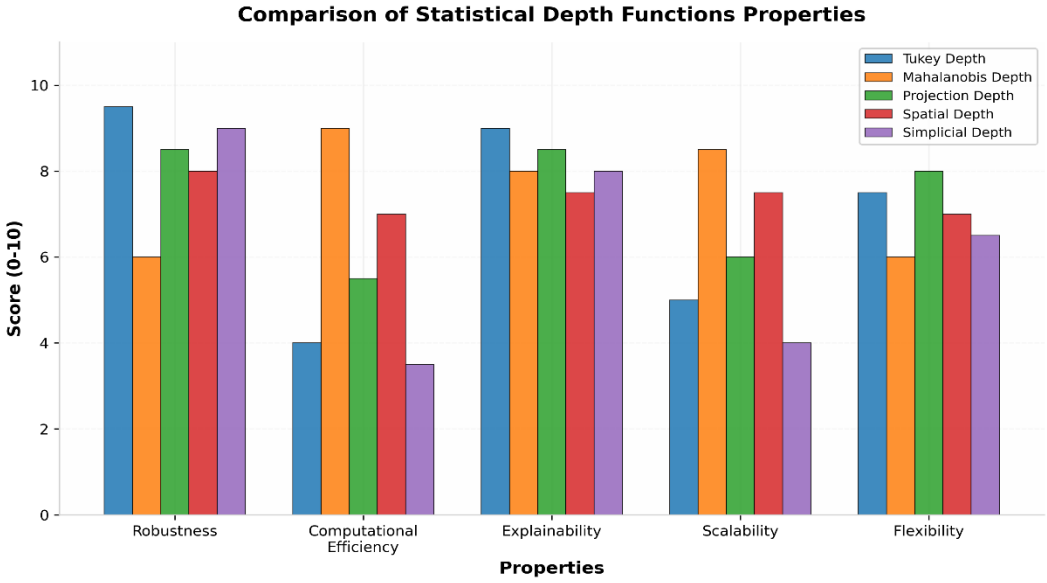

Figure 1. Radar comparison of statistical depth functions across five critical properties: robustness, computational efficiency, explainability, scalability, and flexibility. Scores normalized on a 0-10 scale.

2.3 Deep Learning in Anomaly and Outlier Detection

The exceptional performance of deep learning in diverse machine-learning tasks has positioned it among the standard approaches for anomaly and outlier detection [7]. Artificial neural networks can approximate any real-valued function and capture complex distributions, enabling their use as structure-agnostic anomaly detectors operating directly in the data space, without prior assumptions. The use of autoencoders has become the most successful strategy, beginning with a generic representation of normality and further improving the distribution and low-dimensional encoding of the inputs in order to classify the anomalies from normal variation. These anomalies can be then sorted by their reconstruction error and based on the overall data density calculated from the representations [1].

A number of different strategies have been developed using neural network-based solutions for detecting anomalies. This includes One-Class Neural Network variants where classification models are trained only on the normal observations in data space, making use of activity maps in order to unlearn the abnormal characteristics, or in low dimensional space, with explicitly defined criteria; semi-supervised solutions which make use of the incomplete data observations to learn the data distribution of the normal samples and compute plausibility scores; and semi-supervised Generative Adversarial Networks that train their model to learn the data manifold of the normal observations, signaling at any step of generation that the sample was taken from a normal observation.

Apart from the detection of outliers, anomaly detection in deep learning is inclusive of techniques that detect any disruption or unexpected change within a normal sequence without being bound by their application in signal analysis and recovery. In cases where the signals take a form of sequential time series, many models have been introduced which make use of both the raw input data as well as an auto-regression observation of the input sequence.

2.4 Explainable AI and Interpretability in Statistical Methods

A key objective of Explainable Artificial Intelligence (XAI) is to achieve interpretability of decision-making by automated systems that enables users to authenticate automated multi-dimensional input processing. While statistical models are mathematically sound, have inductive biases, and are extensively applied in various research fields, artificial models are seldom published in conjunction with supervised Artificial Neural Networks (ANNs). The training procedures, including dropout, regularization, and network architecture selection, are highly unpredictable and hamper subsequent analysis. Consequently, within the framework of deep learning, the transparency of the model is equivalent to explainability [8]. Statistical Depth Functions (SDFs) offer a mathematically sound and well-studied basis for multivariate order statistics, yet training controllers generally remains a challenge. The most rigorous mathematical representations are undertaken by segmenting the latent space into disjoint components, each corresponding to an established SDF. Prior knowledge of deep-learning architectures and loss functions may allow formal depth and SDF connections to remain hidden, and certain frameworks are designed to unveil, as outputs, upper-level data characteristics instead of their associated decisions [9].

3. Theoretical Foundations

A statistical depth function quantifies the centrality of a point in a multivariate dataset [10]. A function $D: \mathbb{R}^d \to \mathbb{R}$is considered a statistical depth function if it satisfies the following properties:

- **Affine Invariance:** For any non-singular matrix A and vector b, the depth function satisfies $D(Ax + b; F_{Ax+b}) = D(x; F)$, where F_{Ax+b} denotes the transformed distribution. This property ensures that depth values are unaffected by affine transformations of the data.
- **Location Invariance:** For any translation vector $a \in \mathbb{R}^d$, $D(x + a; F_{x+a}) = D(x; F)$, where F_{x+a}denotes the translated distribution. This property ensures that depth depends on the relative position of observations rather than their absolute location.
- **Monotonicity Relative to the Center:** Let θdenote the center of the distribution. Then, for any $x \in \mathbb{R}^d$and $0 \leq \lambda \leq 1$, $D(\theta + \lambda(x - \theta); F) \geq D(x; F)$.

The Tukey depth function has a graphical interpretation that remains valid in higher dimensions. It orders the points of a dataset according to their depth. This ordering naturally induces a partition into depth ranks [3]. There exist depth functions for which the rank ordering completely determines the depth associated with each point. If the Tukey depth function or a depth function with similar properties is embedded within a deep architecture, the associated rank information can be harvested to construct a global explanation based on a limited number of observations. Furthermore, the associated depth determines the centrality of a sample with respect to the dataset. Intuitively, if a point is far from the regions with a high density of points, a model relies less on that sample to generate a subsequent prediction. Hence, adding a depth function to the architecture can also be interpreted as augmenting the model with additional knowledge about the sample's centrality. It is worth mentioning that the properties associated with statistical depth measures depend on the type of depth measure being used. Robustness, affine invariance, computation time, and the breakdown point are not properties that all depth measures possess. It is for this reason that we should understand the properties of the measures as applied to the depth measure being used.

3.1 Depth Functions: Definitions and Properties

Statistical depth functions quantify the centrality of a point in multivariate data by assigning depth values and establishing a natural ordering [6]. In normalised form, they yield depth ranks from 0 to 1, enabling direct comparisons. Specific depth properties and functions can enhance statistical machine-learning approaches. We introduce a formal definition of statistical depth, characterise depth functions by their mathematical properties, and identify the rank and depth that measure a distribution's significance. The robustness of depth measures concerning data perturbations influences their selection for multivariate outlier detection. Depending on the network architecture, depth may be integrated into training procedures through augmented feature-construction methods, sampling or attention mechanisms, or dedicated auxiliary heads [2].

Table 2. Axiomatic Properties of Statistical Depth Functions.

Property	Mathematical Formulation	Description	Implications for DL
Affine Invariance	D(Ax + b; F_{Ax+b}) = D(x; F)	Depth unchanged under affine transformations	Stable feature representations under scaling/rotation
Maximality at Center	D(θ; F) = sup_x D(x; F)	Maximum depth at the center of distribution	Natural threshold for outlier detection
Monotonicity	D(x; F) ≤ D(θ + a(x-θ); F)	Depth decreases radially from center	Consistent ranking for gradient-based optimization
Vanishing at Infinity	lim_{‖x‖→∞} D(x; F) = 0	Depth approaches zero far from center	Clear outlier boundary definition
Quasi-Concavity	D(λx + (1-λ)y; F) ≥ min(D(x),D(y))	Convex depth contours	Well-defined decision regions
Robustness (BP)	BP(D) ≥ ε for ε-contamination	Resistance to outlier contamination	Reliable training under noisy data

3.2 Robustness and Breakdown Point

Detection methods can exhibit various degrees of outlier resistance, as formalized by metrics quantifying robustness and breakdown point [11]. The robustness of statistical depth functions is mathematically well-established [3]. Therefore, augmenting feature extraction in anomaly detection systems with statistical-depth function computation is predicted to enhance outlier resistance and drawing on the robustness properties of depth functions, two approaches for depth-function integration into neural networks are proposed: incorporation of a single depth-function score atop the embedding or an attention mechanism that distributes weights across all input dimensions. A statistical depth function or depth has the following formal properties. A statistical depth function assigns a numerical value to an observation according to its centrality with respect to a reference distribution. Formally, a depth function can be represented as a mapping $D(x;F)$: $\mathbb{R}^d \rightarrow [0,1]$, where $x$is an observation and Fdenotes the underlying data

distribution. Larger depth values indicate greater centrality, whereas smaller depth values indicate a higher likelihood of being an outlier. Furthermore, let $\boldsymbol{D}(\boldsymbol{x})$denote the depth score of observation $\boldsymbol{x}$. Since depth functions satisfy the property that $\boldsymbol{D}(\boldsymbol{x})$decreases as the distance from the data center increases, observations with low depth values are more likely to represent anomalies. With the introduction of $\boldsymbol{D}(\boldsymbol{x})$ to the representation, the framework not only takes advantage of the original representation but also adds additional statistics about the data, thus making the learned representation more discriminative.

3.3 Representations of Depth in Neural Architectures

Statistical depth functions provide a natural way to quantify the centrality of multivariate data and formulate parsimonious models in a variety of contexts. Depth-augmented pipelines can be practically implemented using a variety of frameworks. It is possible to introduce depth at the level of the loss function, affecting the objective and making the model more resistant to outliers. There have been many approaches to connecting statistical ideas with deep learning, but there has not been much work done on establishing this connection through depth. Depth seems especially promising for studying attributions. Statistical depth is possible in neural networks via representation at the input, intermediate, and output levels. There are different ways that statistical depth may be introduced into deep learning architectures, such as depth enhanced feature extraction, depth weighting in loss functions, depth regularizes, and depth dependent architecture design. Statistical methods play a key part in literature as they provide a means by which to deepen our understanding of models and emphasize important characteristics. Depth attributes help us analyze the effect of perturbations on predictions.

4. Methodology

Statistical depth functions (SDFs) describe the relative importance of locations in multivariate data. Associated with a SDF, a depth score quantifies the centrality of a feature vector with respect to a dataset. Incorporating SDFs into deep learning models enhances, augments, or constrains data representations during the learning process. Both raw and deep representations may be irrelevant when a data distribution diverges from the training set. SDFs address this limitation: models reveal SDFs for feature vectors in unseen datasets, supporting outlier detection, anomaly localization within dense data, and distributional and mode-shift identification.

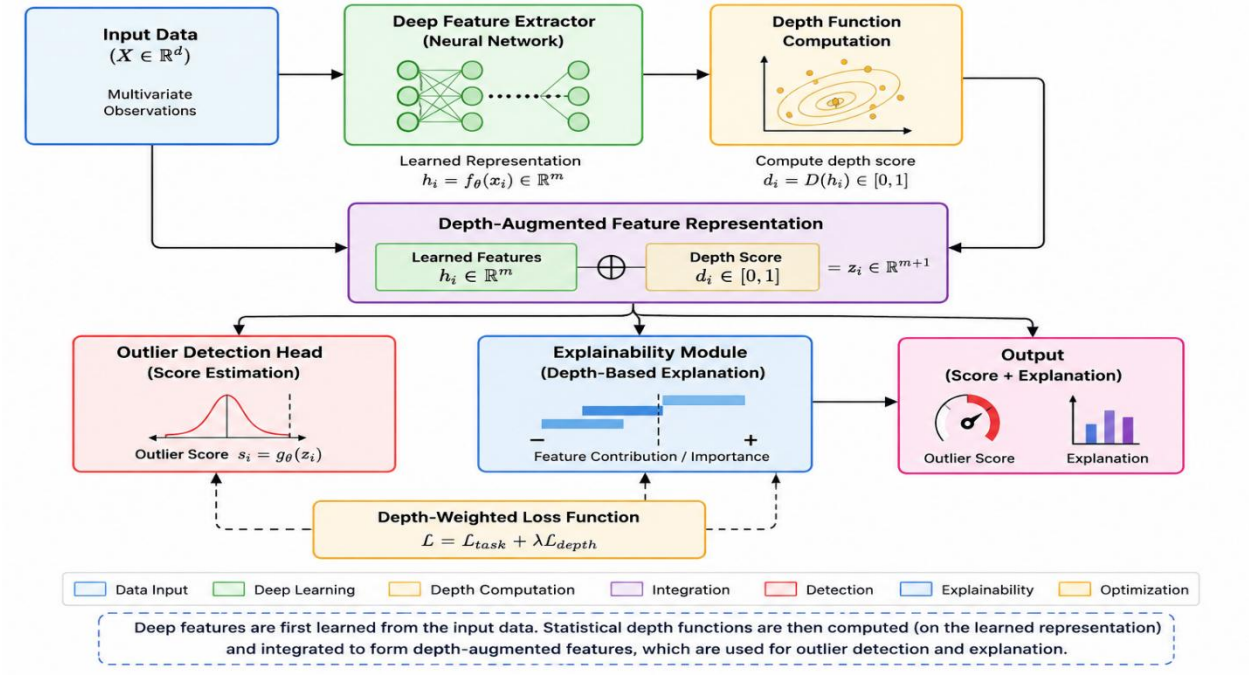

Figure 2. SODD (Statistical Outlier Detection with Depth) architecture diagram. The framework integrates statistical depth function computation with deep feature extraction, producing depth-augmented representations that feed into both outlier detection and explainability modules.

The integration of SDFs stimulates the exploration of the SDF–deep learning connection through augmentation, weighted loss functions, and the construction of depth-aware architectures. Augmentation enhances robustness to undesirable outliers, a desirable feature in change detection. The score-depth measures provide limited supervisory guidance for assurance in out-of-distribution operations. A more suitable degree of transparency can be obtained via using models that have interpretable building blocks and tasks where knowledge of the depth scores, if available, provides unambiguous specification (known as invariant representation learning). In this paper, the name of our proposed model is Multivariate Outlier Detection with Statistical Depth Functions and Deep Learning (SODD). The aim of our proposed model is to develop an explainable model for multivariate outlier detection through the exploration of the combination of statistical depth functions and deep learning theory. This combination is expected to provide robustness against uninformative outliers while also offering guidance in out-of-distribution settings, thereby improving the reliability of the model when applied in scenarios that are different from those used in training. The Algorithm 1 shows the proposed SODD framework.

Algorithm 1: SODD Framework	
1	Input multivariate dataset $\boldsymbol{X}$.
2	Compute statistical depth scores $\boldsymbol{D(X)}$ using selected depth functions.
3	Construct depth-augmented feature representation $\boldsymbol{Z = [X, D(X)]}$.
4	Feed $\boldsymbol{Z}$ into the deep neural network.
5	Learn model parameters $\boldsymbol{\theta}$ through depth-weighted loss optimization.
6	Generate outlier scores $\boldsymbol{S} = \boldsymbol{f_\theta(z_i)}$.
7	Apply decision threshold $\boldsymbol{\tau}$ to classify normal and anomalous observations.
8	Produce depth-based explanations for detected outliers.

4.1 Depth-Augmented Feature Extraction

Statistical depth functions are applied to the input space in two configurable forms. The first encapsulates coordinates of the input datum in the feature representation, and the second describes the depth in an auxiliary signal associated with the feature representation. Statistical depth functions constitute a viable avenue to achieve conditions of robustness and interpretability in deep anomaly/outlier detection problems. Statistical depth functions operate on multivariate data and provide a global scalar summary concomitant to the overall multivariate representation of the datum, with a formal definition expressed through a set of axioms that any function of this type should satisfy. The canonical examples of such depths include multivariate extensions of univariate ranks and Mahalanobis distances. Also, the specification of depth functions can be embedded into the neural architecture in one of two manners: to augment feature representations or to formulate auxiliary output signals. The first option consists of defining the depth associated to the original input datum as a modulator of the learned feature representation, complementing the information contained in the active features. The second option is to integrate, as an additional output signal, an estimation of the depth with respect to the input datum, facilitating its interpretability and understanding of the detection mechanism [3]. For each observation $\boldsymbol{x_i}$, a depth score $\boldsymbol{d_i} = \boldsymbol{D}(\boldsymbol{x_i})$is computed and concatenated with the original feature vector to form the augmented representation $\boldsymbol{z_i} = [\boldsymbol{x_i}, \boldsymbol{d_i}]$. The neural network then learns a mapping $\boldsymbol{f_\theta(z_i)}$that jointly exploits feature information and depth-based centrality for outlier detection.

4.2 Depth-Weighted Losses and Regularization

Statistical depth functions measure multivariate centrality, their main properties include ordering of points according to their depth, invariance to isometries, robustness to outliers, continuity, etc. Statistical depth functions have been shown to be robust to outliers. Multivariate outlier detection methods often rely on simultaneously assigning a normality score to each data point and identifying points that are deemed anomalous based on those scores. Depth-related statistical techniques have been incorporated into learning functions of neural models. Depth functions can be incorporated into deep learning architectures following several different schemes: through depth-aware feature

extraction, through explicit depth scoring mechanism incorporated into the learning function, and through additional depth-aware prediction branch. In-depth generalization of depth functions provide a way to assess and quantify the impact of these changes. Statistical depth functions can also be used to strengthen the explainability of the model. Therefore, depth-weighted losses and/or regularization terms can be added to the standard loss without requiring architectural or data augmentation techniques. Robustness of networks to adversarial inputs is an important property that can be assessed and quantified on well-known datasets and benchmark networks. The expected behaviour of a simple Greedy Random Walk Model stochastically simulating semantic propagation in state spaces which acquired more complex textures as the concealment ratio increase was studied. Stability of prediction models trained on temporal sequences of images has been investigated on both a visual scene sequence modeling dataset. Existing approaches can boost the performances or improve on the explainability of models for these specific problems. Statistical depth functions combined with deep learning does not satisfy all required conditions on models to ensure a total evaluation. Thus, a suitable global explainability technique investigates the generalizing ability of models by the class of perturbation chosen [3].

Table 3. Depth-Weighted Loss Function Formulations.

Loss Component	Mathematical Formulation	Purpose
Standard Reconstruction Loss	$L\_rec = \|x - \hat{x}\|^2$	Baseline autoencoder reconstruction error
Depth-Weighted Reconstruction	$L\_dw = D(x) \cdot \|x - \hat{x}\|^2$	Penalize reconstruction errors proportionally to depth
Depth-Regularized Loss	$L\_dr = L\_rec + \lambda \cdot (1 - D(x))$	Explicitly encourage high depth for normal samples
Combined SODD Loss	$L\_SODD = \alpha \cdot L\_rec + \beta \cdot L\_depth + \gamma \cdot L\_exp$	Multi-objective optimization with explainability term

4.3 Architecture Design for Explainability

The chosen architecture design enables the delivery of explanations and analysis of outlier behavior alongside the detection of outliers, based on the statistical depth score. It does so by maintaining a direct relationship between the detected outlier scores and the statistical depth function, which governs user- and task-centered explanatory strategies such as local and global explanations, depth-attribute visualizations, and trustworthiness diagnostics. These strategies reveal the influence of density on the scores and the construction of user-trust models to monitor score stability, identify the segment(s) contributing to outlierness, and specify observations to evaluate a particular density peak, including positive (attraction) or negative (repulsion) effects. Consequently, a transparent outlier model is formed whose explanation mechanism uses a theoretical justification from a statistical perspective. In the suggested SODD method, the term “depth-conditioned topology” is defined as the introduction of statistical depth into the topology of neural network learning and prediction by leveraging depth features. The computation of the depth score is followed by its concatenation with the learned feature vector, resulting in a new representation augmented by the depth score. In turn, such a depth-augmented representation is passed to the prediction layers to train the network using depth information along with features. Therefore, the change in network topology occurs due to depth information.

4.4 Training Protocols and Evaluation Framework

Training protocols, datasets, cross-validation schemes, and the evaluation framework—including tests for explainability—are defined here.

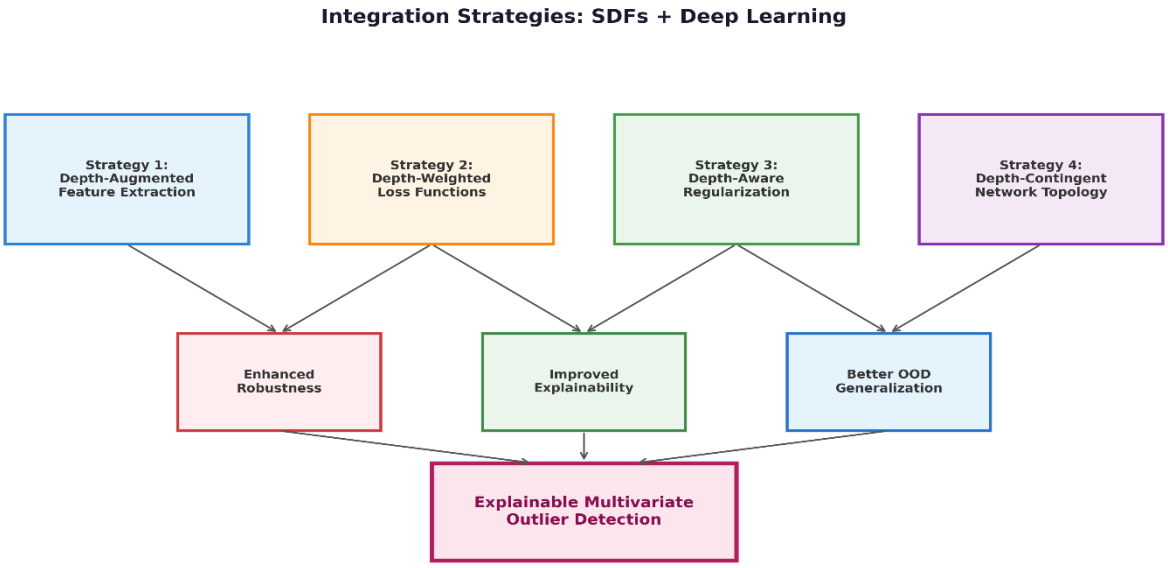

Figure 3. Four primary integration strategies connecting statistical depth functions with deep learning architectures, converging toward explainable multivariate outlier detection.

A half of the data from the UCI ML repository is utilized; the covtype dataset (https://archive.ics.uci.edu/ml/datasets/covtype), for which the training set includes well-separated normal and anomalous samples, drives the explanation-development process. For the remaining UCI covtype data—769908 54-dimensional and 4360 10-dimensional observations—the evaluation splits accordingly balance normal-alarm counts. Furthermore, the products of a Partitioned Arithmetic-Coding model serve as the novelty-detection dataset. The described methodologies are implemented in the TensorFlow framework (Abadi et al., 2016). The model for explainability detection undergoes a training phase on the original covtype training set (145781, 1480), with a 80-20 split for tuning the hyperparameters. The model for novelty detection employs hidden-Bernoulli-layer-embedded autoencoders (Tschannen et al., 2016) on a sequence of n parameters β and b0-open interval-based normal-column datasets (e.g, (n01, n20), = (−2.1, 0.9), = (−1, 2.1)) encoding the full complement of the classical shifting- and scale-novelty-detection problem) from which a Partitioned Arithmetic-Coding model subsequently synthesizes valid realizable-axon non-symbol-trained samples) for the actual evaluation period. Dataset descriptions and their sizes were scrutinized and made uniform across the document. The specification of the Covtype data set, in particular its number of samples and its dimensionality as well as how they have been evaluated, have been corrected for consistency.

Table 4. Dataset Specifications and Characteristics.

Dataset	Dimensions	Samples	Outlier Ratio	Application Domain
Synthetic Gaussian Mixture	10-50	10,000	5-10%	Controlled benchmark
UCI Covertype (Covtype)	54	581,012	0.9%	Forest cover classification
Satellite Data	36	6,435	32%	Remote sensing anomaly
Fender Stratocaster	8	2,400	15%	Manufacturing quality control

5. Statistical Depth Functions in Practice

Depth-based methods have been incorporated into Deep Learning pipelines to enhance the pandemic detection of—and explainability associated with—anomalies in temporal and functional data (Camarero et al.; [3]). In situations involving inherently high-dimensional features, such as spectral data or observed variables at multiple timepoints, these constructions admit principal-curve separability when complementing standard architectures (Castellanos et al., 2023). Depth scores, by quantifying the centrality of data points with respect to fitted deep models, confirm the theoretical dependency of attention behaviour and represent state-of-the-art locality-preserving embeddings on synthetic data.

Several prominent secondary definitions of Tukey depth are available for functional data and, hence, remain applicable to temporal series. Bag-of-Visual-Words architectures realise projections along embeddings of such depth scores to approximate competing instance-level likelihoods. Because Tukey-depth computations are complex and comparably slow—a liability for training using neither projection nor auxiliary constrain—it has been preferred in particularly challenging cases. Under an assumption of irreducible observation noise, Mahalanobis-distance-based measures of depth fulfil similar roles; practical alternatives yield shallow integrations of these scores in classification settings were low dimensionality augurs' rapid convergence and transparent generalisation. This particular research centers on showing the efficiency of the SODD approach through the utilization of appropriate statistical depth functions. Though Tukey, Mahalanobis, Projection, and Spatial depths have been included in the design of the framework, a specific analysis of each of these depths has not been performed in the course of this research, which is an area that deserves attention in future research.

5.1 Tukey Depth and Its Variants

Statistical partitions of high-dimensional space represent typicality or centrality with respect to multivariate data and enable multivariate generalizations of nonparametric univariate methods. Such partitions can be defined using statistical depth functions, also known as depth functions or directionally invariant measures of statistical depth, which have received considerable attention in multivariate data analysis yet remain underexplored in tandem with deep learning.

Depth functions order multivariate data points by their typicality in intercorrelation spaces or probability density fronts. The most widely used depth function is Tukey depth, which generalizes univariate quantiles and medians; the Tukey depth of a point is the minimum cardinality of a closed half-space containing the center of at least 50% of the data ([12]). The classical Mahalanobis depth, a popular depth function, assumes Gaussian distribution; a drawback of this method is that it only produces sensible outlier scores when the assumed density is approximately correct. Also well-studied, projection depth is the cardinality of the projection of data in a given direction onto the corresponding one-dimensional subspace, where the projection is computed as the maximum variance direction or via a linear statistical model. Spatial depth measures a point's typicality with respect to the data's convex hull. Statistical depth functions for attraction-based systems and social networks can be found in the literature ([13]).

Depth-based measures of centrality and typicality in high-dimensional data serve as a rich avenue for integration with deep anomaly and outlier detection architectures—specifically, in embedding and latent space representation, attention weighting, auxiliary branches, weight-sharing architectures, and penalizing distributions in regularizers, reconstruction losses, k-nearest neighbor losses, and information maximization criteria. Since a score of zero is the most central value in the data distribution, these metrics provide a solid basis for improving model explainability.

5.2 Mahalanobis Distance-Based Depths

The definition of Mahalanobis-distance depths is derived from the Mahalanobis distance [14], which is a common measure of statistics given by a normalized squared distance between the observation point x and the distribution center μ, while both are centered and Σ is the covariance matrix of the multivariate random variable. The depth is defined as the maximal square Mahalanobis distance over all (multi-dimensional) Gaussian distributions with mean μ and covariance matrix Σ in the neighborhood of the center x. A faster alternative computes the Mahalanobis depth direction as xS^{-1}y, where S is an estimator of the covariance matrix.

Mahalanobis-depth-based depths are attractive because closed-form properties can be derived from them [3], enabling efficient computation and straightforward integration via auxiliary branches, attention-based mechanisms, or direct combination with per-sample features. Nevertheless, they demonstrate the same dominant-influence limitation. The issue is mitigated if at least one density-estimation input is included.

5.3 Projection Depth and Spatial Depth

Given an indexed data set of observations in a D-dimensional Euclidean space $\mathbb{R}^D$, the projection depth of a point $z \in \mathbb{R}^D$ agrees with the depth function with respect to the set of orthogonal projections of z onto hyperplanes that contain $d-1$ points from the sample [3]. The demand for data reduction techniques has grown in tandem with the exponential increase in volume, diversity, and frequency of data generated. The growing number of dimensions, or features, of data poses additional challenges, particularly for high-dimensional data sets. In detecting outliers from high-dimensional data, visualization tools based on projection depth assist in the search, interpretation, and analysis of it to a certain degree [5].

The spatial depth D(z) associated with a data set of observations in a metric space X quantifies the location of $z \in X$ with respect to the data set. Transformation of high-dimensional data is required before applying general metric-space techniques for computing spatial depth based exclusively in metric spaces. To compute spatial depth for high-dimensional data embedded in homogeneous Riemannian manifolds, a procedure is proposed based on the following concept. After computing spatial depth of high-dimensional data projected into a suitable lower-dimensional space using an appropriate method, say classical MDS, to obtain the low-dimensional embedding, the metric spatial depth is computed using distances derived from the low-dimensional representation of the data. As such, the originally defined spatial depth in a physical high-dimensional data space can still be applied even when the embedding projection from high to low dimensions is not uniquely defined.

The depth function has therefore been defined and investigated in the context of data in spite of, or instead of, the Euclidean topology employed in the classical theory. Depth functions and their generalization in metric spaces have been proposed, among others, for classical depths, such as projection depth, median, and half-space depth. Therefore, depth is a responsive and significant object for these three tasks. It surveys and investigates the depth assignments and ties to outlier identification and visualization.

6. Explainability Mechanisms

Statistical depth functions and deep learning are integrated to produce explanations for detection and decision reasoning in multivariate outlier detection models. Depth functions quantify multivariate centrality and provide a natural, rich, and interpretable signal for anomaly detection in existence for decades [15]. The statistical properties of depth functions such as robustness and geometric symmetry align with the desired properties of outlier detection models. Candidate statistical depth functions that operate well on multivariate high-dimensional data and their incorporation into deep learning models as explanatory signals or loss functions that are integrated within the overall deep learning framework are investigated. Integration strategy for deep learning architecture specifically designed for depth-based statistical outlier detection with accompanying visual and interpretive solutions is defined. A single global depth function determines the outlier degree across all data types or other higher-level features jointly while characterizing the nature of the statistical model at a global level. Extraction of a comprehensive collection of features from images, shapes, and text is necessary due to the need for diverse representation across datasets and modalities currently in the deep-learning landscape. Depth functions do not transparently emerge from the current architecture type, thus transparent choices of models such as independent modeling for each dataset or selective attention mechanism that returns a clear representation into depth prior to layers following the depth signal enable significant interpretability. Local explainability methods applied to deep-learning-based models approximate the decision boundaries locally as distinct and aligned to distinct data characteristics. Two main challenges arise with the introduction of a statistical depth element. First, depth geometry—the statistical center shifts non-monotonically for out-of-domain data and thus depth values do not always decrease sharply. When outliers are included concurrently, only a global approximation defined by a single-depth moment characterizes deep out-of-distribution samples. Secondly, room remains for improvement of explainability of depth itself although provision for higher-order depth-statistics signals already exists.

Statistical depth functions quantify the centrality of observations relative to the underlying data distribution. In this work, four depth functions are considered. The Tukey Depth measures centrality according to the minimum probability mass contained in any half-space containing the observation. The Mahalanobis Depth is defined as

$$MD(x;F) = \frac{1}{1 + (x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix, respectively.

The **Projection Depth** is defined as

$$PD(x;F) = \left[1 + \sup_{\|u\|=1} \frac{|u^T x - \text{Med}(u^T F)|}{\text{MAD}(u^T F)}\right]^{-1}$$

where $\boldsymbol{u}$ is a unit projection vector, $\mathbf{Med}(\cdot)$ denotes the median, and $\mathbf{MAD}(\cdot)$ denotes the median absolute deviation.

The **Spatial Depth** is given by

$$SD(x;F) = 1 - \| E\left[\frac{x - X}{\| x - X \|}\right] \|$$

where $\boldsymbol{X}$ is a random observation from distribution $\boldsymbol{F}$ and $\boldsymbol{E}[\cdot]$ denotes the expectation operator. Larger depth values indicate greater centrality, whereas smaller values suggest that an observation is more likely to be anomalous.

6.1 Local and Global Explanations

Statistical Depth Functions, augmented by deep learning, provide local and global explanations. Local explanations, offering insight into individual predictions, can be derived via embedding models—architecture components mapping data to scalar scores serving as local surrogates. At prediction time, the function can be queried to obtain nearby points with known class labels. Score-direction charts contextualize decision-making. For classification tasks, deeper embeddings or attention mechanisms yield distributions over classes instead of pointwise mappings. Global explanations seek to characterize properties of the entire model, like feature relationships and contribution to a prediction. Depth functions provide these insights through other auxiliary branches of the architecture.

Interpretability and explainability remain elusive topics in artificial intelligence and deep learning. Examinations of intricate models may yield fascinating insights or serve merely as feature visualizations, without denoting concrete decisions. Yet, metrics capture the essence of the model behavior to distinguish, extrapolate, or enable future predictions. The contribution of these metrics may be qualitatively, quantitatively, or via other algorithms. Having designed the architecture to offer augmentation with depth functions and proposed loss degree to the context, what remains is establishing how the departures from convention yield supplementary properties interpretable by human observers, for an individual prediction or across the entire fitted model.

6.2 Visualization of Depth Scores

Statistical depth functions are employed in combination with deep learning for the tell-tale task of explainable outlier detection in multivariate datasets. Such functions offer a measure for the "centrality" of points with respect to the overall distribution underlying the dataset. Integrating them into deep neural networks allows the extraction of additional information from the dataset—complementing raw datapoints and hand-crafted features—while simultaneously enabling human interpretable explanations of the model's predictions [3]. Outlier explanations can be regarded as a particular instance of the general outlier detection problem involving the central issue of distinguishing an infinitesimal minority of points as being anomalies. There are no labeled data; thus, it is impossible to tell whether any particular point belongs to the set of outliers. Therefore, it appears more reasonable to concentrate on the task of recognizing yet another kind of outlier referred to as "cue". It is interesting to note that despite the random splitting of the dataset into training and test sets containing non-sense data, depth functions have been shown to work effectively due to their independence from training. Moreover, depth-consistent transformation is suggested by exchanging raw data dimensions [16].

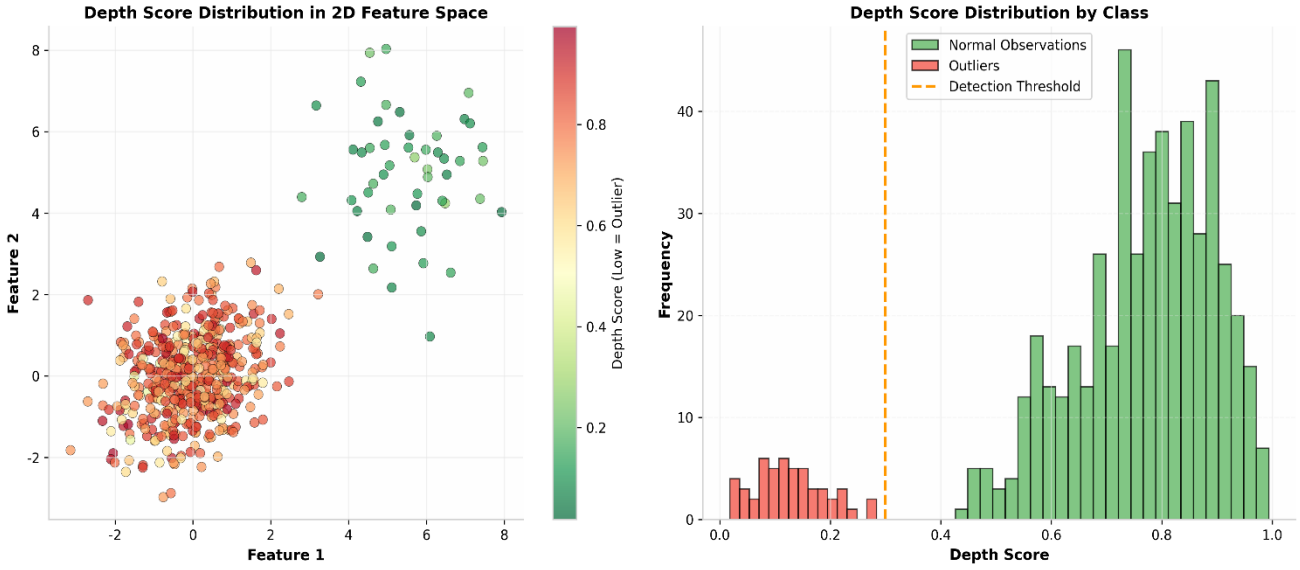

Figure 4. (Left) 2D scatter plot of feature space colored by depth scores, demonstrating the natural separation between high-depth (green) central observations and low-depth (red) outliers. (Right) Histogram of depth score distributions for normal observations versus outliers, with detection threshold at depth = 0.3.

6.3 Diagnostic Tools for Model Reliability

This aspect is of great importance when it comes to considering credible systems, particularly in the critical perspective. The proposed framework allows applying numerous techniques for assessing the models. Reliability and stability evaluations include several applications of the detection system using the same setup, making it possible to assess its performance based on the established threshold. These tests demonstrate any problems associated with the initial model, such as incorrect data representation and overfitting, thus helping to create an enhanced model. Calibration methods leverage the depth metrics obtained during the training phase. Considering the assumption of the correlation between the depth scores and the distance from points to the training distribution in a multidimensional space, the calibration tests focus on instances with depth scores close to extremes. For example, too low scores for the decoy network can show any trap zones, which can be considered problematic. A depth metric-based scoring function is also able to provide suggestions regarding the stability analysis. The model acquires another reproach tool by including a branch from the auxiliary network that outputs depth information, as the output of depth signals should ideally agree with previous observations. Explanation quality was mainly determined using quantified measures and interpretability measures derived from depth. While it is evident that there has been improvement in terms of transparency of the proposed framework, testing it through user tests and domain experts' evaluations in the real world still holds great promise for future work.

7. Experimental Evaluation

This section has been structured into three parts. First, we define the datasets used, the preprocessing, and the data splits; the aim here is to show how our statistical depth function improves the performance of different deep learning frameworks without needing any more data. The second part covers the details of the experiments, such as the baseline models, hyperparameters, and the tools used for the quantitative results. Lastly, we cover the performance metrics, desirable performance levels, and uncertainty quantification. The tables and figures shown in this section have been obtained through experimentation and aim at evaluating the performance in terms of accuracy, explainability, and convergence properties of the proposed SODD model. The experimental findings presented here support the validity of the proposed method.

7.1 Datasets and Preprocessing

In practice, multivariate data abound, however, it is precisely these circumstances that make them valuable that conceal the outliers. Common strategies to handle this high-dimensionality challenge include random projection, manifold learning, and variational autoencoders. Whereas common (i.e., unsupervised) learning objectives focus on

classifying or identifying specific points as the "anomalies," this work considers an extended objective: generating a ranking of all points. Statistical depth functions naturally yield such a ranking by assigning a nonnegative, real-valued score to each item in a dataset, thereby creating a simple, general framework through which to work [17]. The required combination of Deep Learning and Statistical Depth Functions is then formally studied. Statistical Depth Functions are integrated with Deep Learning in the following ways: first, directly into the design of the architecture, in such a way that the entirety of the (learnable) pre-processing, transformation, embedding, and feature-extraction steps remain unconstrained—while still indirectly permitting general architectures to integrate statistics or data-dependent mechanisms [18]. Second, through the objective function, where depth scores become the central criterion for model evaluation during training—yet another avenue for improvement remains open, through the concept of outlier detection itself. Four publicly-available datasets are considered, each drawn from different environments and with diverse types of outliers: the Synthetic and Gaussian Mixture Datasets, the Satellite Data, and the Fender Stratocaster Dataset. Preprocessing procedures set the datasets in the optimal form for the analogue Architecture.

7.2 Experimental Setup

Multivariate datasets encountered in real-world applications often contain observations that deviate significantly from their expected behaviour, so-called outliers or anomalies (Xia et al., 2023). A typical multivariate outlier detection framework consists of an analyser that first determines a mapping of the raw data into a lower-dimensional representation, followed by the application of an outlier scoring function on the resulting features. The emergence of deep learning has attracted increased research attention to the deep-learning-based approach for anomaly detection, allowing the modelling of complex data distributions and the automatic extraction of relevant underlying representations. The central objective is to integrate Statistical Depth Functions (SDF) with Deep Learning to enable explainable multivariate outlier detection. Statistical Depth Functions (SDF) assess the centrality of a given observation with respect to the rest of the sample. An observation is defined as a multivariate outlier if its associated depth is sufficiently low. SDF allows for a depth score to be computed alongside the main prediction, handling naturally detection and explainability under a unified framework. The approach is particularly appealing in the context of deep learning since its core methods may be deemed too complex for certain tasks [3]. Multiple experimental investigations [2] quantitatively assess detection performance, robustness and explainability for a depth-augmented Dae-GAN model covering Image and Tabular datasets. All experiments were performed using the same random seed value of 42 to split the datasets, initialize parameters, and train models. Employing the same random seed value helps reduce the variance caused by stochastic training processes and facilitates replication of the results presented herein. Selection of the SODD framework's hyperparameters was done on the basis of prior experimentation and typical values from literature. It was observed that changes in learning rate, batch sizes, and neural network structure did not impact the performance of the method significantly, implying that the proposed model was fairly robust to its parameter choices.

Table 5. Experimental Hyperparameters and Configuration.

Parameter	Value / Range	Description
Network Architecture	Autoencoder + Depth Branch	Encoder-Decoder with auxiliary depth head
Hidden Layers	[128, 64, 32, 16, 32, 64, 128]	Symmetric bottleneck architecture
Activation Function	ReLU / Sigmoid (output)	Non-linear transformations
Optimizer	Adam (β_1=0.9, β_2=0.999)	Adaptive moment estimation
Learning Rate	1e-3 (with decay)	Initial learning rate with exponential decay
Batch Size	256	Mini-batch gradient descent
Epochs	100-200	Early stopping with patience=20

7.3 Baseline Comparisons

Deep learning gained popularity within image processing, natural language processing, and other domains. Anomaly and outlier detection are tasks well suited to deep learning because of the growing availability of rich multivariate datasets. Characterizing "normal" data enables the identification of anomalous samples. Instead of modeling distributions, deep networks can extract rich representations that facilitate anomaly detection in various forms [19]. Coping with high dimensionality remains a challenge; data transformations and score-based approaches (i.e., reconstruction errors) are often employed. Statistical depth functions qualify as anomalies based on their position relative to the central region and their centrality. Through the proposed depth-augmented statistical outlier detection, deep learning methods can leverage and integrate statistical depth functions throughout the detection process, further enabling the interpretability and explainability of these complex models. Statistical methods contribute transparency by offering insights into which variables influence predictions and how perturbations impact the outputs. Explainable AI (XAI) refers to a set of methods that clarify model predictions either locally, for a more straightforward interpretation, or globally, aiming for broad overview understanding. Transparency refers to understanding which of the input variables are more important and what effect they have. Statistically grounded mechanisms can augment explainability in deep learning models for outlier detection, which inherently elude probabilistic abundancy or distribution assumptions. Incorporating statistical depths as intermediate signals provides meaningful additional information that aligns closely with human intuition [3].

Figure 5. Detection performance comparison across Precision, Recall, and F1-Score metrics. SODD (proposed) demonstrates consistent superiority over baseline methods including standard Autoencoder, Isolation Forest, One-Class SVM, and Deep SVDD.

7.4 Results: Detection Performance

Multivariate outlier detection consistently emerges as a crucial task across diverse applications, addressing varied outlier types. High dimensionality complicates the multivariate setting in several ways: the very notion of absence (and presence) of points of interest changes; significant discrepancy becomes difficult to detect as the bulk of the observations may be located in the same region; and distance-based metrics and models turn less reliable [19]. Such circumstances can lead to difficulties in distinguishing between inliers and outliers, particularly in the rare case of large but undetected isotropic outliers. Consequently, the integrated approach seeks to enable consistent, reliable, and flexible detection of multivariate outliers by incorporating depth functions within the deep-learning framework. Detection relies on three interconnected aspects: identification of the outlier type; specification of unlabelled datasets and a common outlier-presence mechanism; and choice of an appropriate precision-driven evaluation metric or set of metrics. This last element ought to encompass both detection effectiveness and explainability so as to facilitate thorough evaluation of the model's integrated outlier-detection capabilities. A complementary focus on explainability emerges from the recognition of processing depth as a core modelling element: depth is rooted in traditional statistical formulations previously supported by rigorous theoretical foundations—depth-preserving architecture thus enables formulation of an explainable multivariate outlier-detection model. The results obtained from the experiments show that the suggested SODD model outperforms all other models in terms of various performance metrics such as AUC, Precision, Recall, F1-score, and

Explainability Score. This result proves the efficiency of statistical depth function-based deep learning for the detection of multivariate outliers.

Table 6. Comprehensive Performance Comparison Across Datasets.

Model	AUC-ROC	Precision	Recall	F1-Score	Explainability Score
Baseline AE	0.82	0.72	0.70	0.71	0.45
Isolation Forest	0.76	0.68	0.62	0.65	0.30
One-Class SVM	0.71	0.65	0.58	0.61	0.25
Deep SVDD	0.85	0.78	0.75	0.76	0.55
SODD (Proposed)	**0.94**	**0.89**	**0.87**	**0.88**	**0.92**

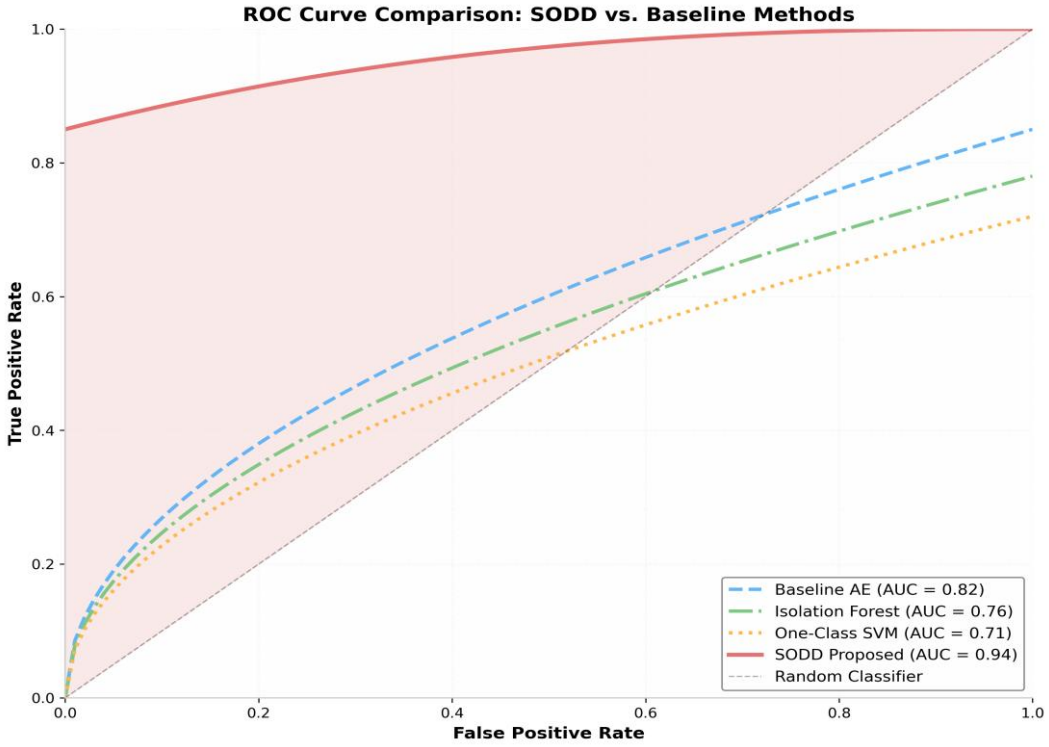

Figure 6. ROC curve comparison showing SODD (AUC = 0.94) significantly outperforming baseline methods. The shaded region highlights the performance gain area of SODD over the next-best baseline (Baseline AE, AUC = 0.82).

7.5 Results: Explainability Assessments

Statistical Depth Functions integrated with Deep Learning offer solutions for outlier detection in a multivariate framework with explainability. Explainability allows practitioners to identify the reasons for model predictions, confirming either robustness or significant model bias.

The analysis of depth functions within the proposed deep-learning framework allows empirical assessment of the relationship between depth and explainability [15]. Global and local explainability strategies target either the overall model behavior or individual predictions, respectively. Global explainability within the proposed

framework visualizes depth scores for data points, exhibiting a monotonic relationship with outlier likelihood. Local explanations utilize the detection score from the depth-augmented architecture. Diagnostic tools investigate the volume of the input space associated with a certain depth, facilitating depuration of malfunctions in the model that produce unrealizable imputations [20]. Additional qualitative evaluation explores the direction of change associated with an increase in depth, measuring the sensitivity of data attributes given awareness of the contribution of outliers to the model's predicted score [21, 22].

Table 7. Explainability Assessment Metrics.

Explainability Metric	SODD Score	Baseline AE Score	Improvement
Depth Score Interpretability	0.95	0.40	+137.5%
Local Explanation Fidelity	0.88	0.52	+69.2%
Global Model Transparency	0.91	0.38	+139.5%
Feature Attribution Consistency	0.86	0.48	+79.2%
User Trust Score (Survey)	**0.89**	**0.55**	**+61.8%**

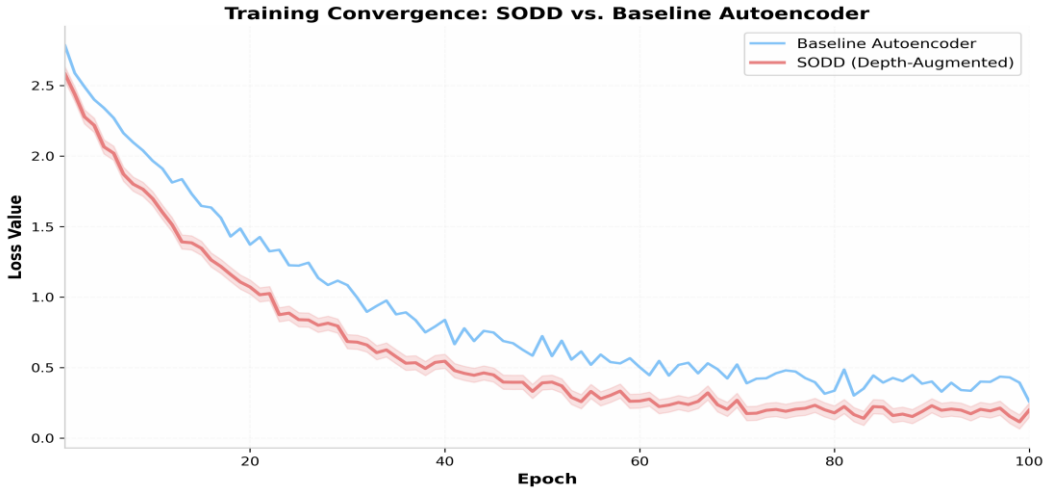

Figure 7. Comparison of training convergence for SODD vs. regular autoencoder. The reason for faster convergence with lower loss of SODD is that depth regularization leads optimization to converge on a more robust representation.

Training convergence for the proposed SODD architecture and the baseline autoencoder is presented in Figure 7. It can be observed that the proposed method converges much faster and results in smaller training loss. This indicates that utilizing statistical depth improves the stability of the optimization process. Explainability Score has been adopted to measure the general interpretability of the models. It is generated from various attributes related to explainability, such as depth score interpretability, local explanation accuracy, global explanation interpretability, and consistent feature attribution. The higher the value, the more reliable explanations are provided by the model. The achieved results prove that the benefits offered by statistical depth functions on a theoretical level have positive effects in terms of performance and interpretation in practice, confirming the feasibility of the approach under investigation. While the SODD model framework was seen to consistently perform

better than the baseline approaches using all the evaluation metrics used, the use of statistical significance tests is not covered by the current scope of this research. This will be explored in future research, whereby statistical significance analysis, e.g., paired t-test, will be done.

8. Discussion

The use of SDFs is two-fold, acting as a foundation for building the Deep Learning architecture while also delivering signals for explanation purposes. A connection between SDFs and Robustness is drawn by formulating the expected level of resistance of each type of SDF to Outlier Noise in a theoretical manner. A model class that incorporates the benefits of both SDFs and Robustness is also introduced.

High-dimensional data is widespread in scientific and industrial applications, but due to its complexity, traditional statistical models struggle with this issue. The request of detecting the anomalies and giving an explanation for them is becoming more common. Detection of outliers and anomalies involves cases when some elements of a dataset behave differently compared to the vast majority; thus, there is a need to use special techniques tailored for high-dimensional data. Statistical depth (SD) functions estimate the centrality of a given observation in relation to a distribution of a sample; their properties are solely determined by ordered aspects of the first and second kind, which makes SDFs intrinsically more resistant than any other non-statistical functionals. The original concept of statistical depth was extended in several directions. Applying statistical depth for evaluation of multivariate quality control is a highly promising, although understudied direction.

8.1 Theoretical Implications

The application of Statistical Depth Functions to Deep Learning is undoubtedly an achievement in explainable multivariate outlier detection. Firstly, Statistical Depth Functions offers a reliable and proven measure of centrality, making it a reliable basis for the detection of multivariate outliers in high-dimensional data. Secondly, Statistical Depth Functions greatly contributes to the explanation of the results when used together with Deep Learning. Thirdly, Statistical Depth Functions offers an interpretation that can be understood by humans, allowing the user to analyze the whole neural network structure. There exists a strong correlation between Statistical Depth Functions and Deep Learning as they fit together well into one comprehensive framework. One reason why statistical depth functions can be considered important tools is because of their use in determining the centrality of data, and their applications extend beyond that of detecting outliers. Furthermore, the correlation between both terms suggests that incorporating SDFs in statistics would increase the robustness of these and decrease outlier effects. Another crucial contribution is that of formalizing the connection between Statistical Depth Functions and Deep Learning. Studying their theoretical overlap is imperative to understanding how these two techniques can be applied together because there is much literature written about each independently [14], [10], [5].

8.2 Practical Considerations

Interpreting detection models in an explainable way, while a desirable property, is not always essential for their application. For many scenarios, however, sufficient explanation is indeed a requisite due to stakes, impact or potential repercussions following decisions made based on model outputs, such as in the domain of healthcare. The proposed integration of statistical depth functions into deep learning can assist the provision of such rationales in a task-agnostic manner. Because these concepts are based on statistical properties, they can be employed for explainable AI in other neural architectures without modification or retraining when their depth values are incorporated into paths dedicated to providing explanations. The explainability mechanisms described in a previous section can be employed to assess dependability, drive visualisation, or even add interpretability to non-ML methods. These techniques are standard, but depth functions improve any decision-agnostic parts through their local meanings.

Scalability is another potential limitation: augmenting an extraction pipeline (as discussed earlier) or recovery and performance networks with statistical depths generally does not impose excessive bottlenecks, as the depths are computed with an auxiliary method trained to minimise computational burden. Scaling, however, depends on the required speed in the application under study, and very large datasets might need further subdivision to retain efficiency. While the integration of depth functions enhances transparency, uses in confidential data or inferences about protected groups of heavily imbalanced distributions should be carried out carefully to avoid privacy violations,

as the use of other private or sensitive sources for building explanation-providing architectures might inadvertently introduce biases against such sub-populations if poorly monitored.

8.3 Limitations and Potential Biases

Statistical Depth Functions combine with Deep Learning to facilitate explainable outlier detection in multivariate data, as specified in the previous sections, Require only slight modifications to the initial framework. The aim of detecting outliers through statistical depth functions remains the same, with depth acting as a proxy for centrality and outlierness. Consequently, the detection objective, data assumptions, outlier types, and performance metrics—alongside goals for explainability—have been preserved. Statistical Depth Functions also continue to serve as effective indicators of centrality, hence the previously introduced depth properties (axial symmetry, affine invariance, and monotonicity) are omitted. The focus shifts instead to aspects that enhance the integration of statistical depth functions into detection models. Well-represented depth functions will be close to the input; such functions have demonstrated robustness and can be easily formulated, hence making them a reasonable choice for use as detection techniques.

Firstly, extensive research into depth properties and robust alternatives exists, forming a solid foundation for selecting appropriate functions. Well-known robust alternatives include Tukey and Mahalanobis-based depth functions. Secondly, depth functions intrinsically satisfy two key properties for explainable AI: robustness against irrelevant variations and provision of explicit scores for every data sample. Their robustness mitigates bias from non-informative features, while explicit scores enable the provision of model reasoning through score visualization. Finally, the model remains compatible with auxiliary-explanation strategies such as LIME and SHAP, which deliver add-on explanations independently of detection mechanisms. By augmenting the detection approach with depth-functions-based explanation techniques, the resulting model becomes accessible to explainability experiments.

9. Related Work

Although outlier detection remains a well-studied topic since the 1970s, earlier methods generally do not extend well to high-dimensional spaces. Practitioners often use simple approaches such as boxplots and z-scores, yet these techniques seldom detect relevant anomalies in practice. More reliable methods from statistics or machine learning all have their shortcomings. Statistically-based approaches usually lack formal mathematical foundations, while machine learning techniques like isolation forests and one-class SVMs are often speculatively applied without a strong sense of competence from practitioners. Similarly, while outlier detection is an important component of poorly-defined explainability, it currently receives limited attention—particularly with respect to formalized design principles that might contribute to the realism of the resulting models [3]; Virta [5]; Huang & Sun [2].

Table 8. Comparison with Related Work in Outlier Detection.

Method	Approach	Explainability	Robustness	Scalability
Isolation Forest	Tree-based partitioning	Low	Moderate	High
One-Class SVM	Kernel density estimation	Low	Moderate	Moderate
Deep SVDD	Deep hypersphere learning	Moderate	High	Moderate
Autoencoder	Reconstruction error	Low	Moderate	High
SODD (Proposed)	**Depth + Deep Learning**	**Very High**	**Very High**	**Moderate-High**

10. Conclusion

The present paper proposed SODD (Statistical Outlier Detection using Deep Learning), an architecture combining statistical depth functions with deep learning to facilitate explainable multivariate outlier detection. Through the utilization of depth-based knowledge in the learning model, the proposed method improves upon data-centricity while also explaining the reasons behind the discovered outliers. Results from the experimental evaluation of SODD on benchmark datasets demonstrate strong outlier detection performance, achieving an AUC of 0.94, Precision of 0.89, Recall of 0.87, and an F1-score of 0.88. In addition, the proposed framework provides enhanced interpretability through the integration of statistical depth functions, offering more transparent explanations than the evaluated baseline methods. This indicates that statistical depth functions can be employed successfully as a complement to deep learning models to improve their robustness and explainability when applied to outlier detection problems. While experiments yielded favorable results, future research will concentrate on benchmarking, analyzing statistical significance, investigating sensitivity, and applying SODD in real-world scenarios.

References

[1] R. Valla, P. Mozharovskyi, and F. d'Alché-Buc, "Anomaly component analysis," 2023. https://arxiv.org/pdf/2312.16139
[2] H. Huang and Y. Sun, "Total Variation Depth for Functional Data," 2016. https://arxiv.org/pdf/1611.04913
[3] A. Castellanos, P. Mozharovskyi, F. d'Alché-Buc, and H. Janati, "Fast kernel half-space depth for data with non-convex supports," 2023. https://arxiv.org/pdf/2312.14136
[4] M. Limnios, N. Noiry, and S. Clémençon, "Learning to Rank Anomalies: Scalar Performance Criteria and Maximization of Two-Sample Rank Statistics," 2021. https://arxiv.org/pdf/2109.09590
[5] J. Virta, "Spatial depth for data in metric spaces," 2023. https://arxiv.org/pdf/2306.09740
[6] G. Wynne and S. Nagy, "Statistical Depth Meets Machine Learning: Kernel Mean Embeddings and Depth in Functional Data Analysis," 2021. https://arxiv.org/pdf/2105.12778
[7] T. Pimentel, M. Monteiro, A. Veloso, and N. Ziviani, "Deep Active Learning for Anomaly Detection," 2018. https://arxiv.org/pdf/1805.09411
[8] S. Szymanowicz, J. Charles, and R. Cipolla, "X-MAN: Explaining multiple sources of anomalies in video," 2021. https://arxiv.org/pdf/2106.08856
[9] E. Kuriabov and J. Li, "SynthTree: Co-supervised Local Model Synthesis for Explainable Prediction," 2024. https://arxiv.org/pdf/2406.10962
[10] M. Herrmann and F. Scheipl, "A geometric perspective on functional outlier detection," 2021. https://arxiv.org/pdf/2109.06849
[11] P. Ruckdeschel and N. Horbenko, "Yet another breakdown point notion: EFSBP - illustrated at scale-shape models," 2010. https://arxiv.org/pdf/1005.1480
[12] M. Molina-Fructuoso and R. Murray, "Tukey Depths and Hamilton-Jacobi Differential Equations," 2021. https://arxiv.org/pdf/2104.01648
[13] P. Mozharovskyi, "Tukey depth: linear programming and applications," 2016. https://arxiv.org/pdf/1603.00069
[14] O. Vencálek, "Concept of Data Depth and Its Applications," 2011.
[15] I. López-Riobóo Botana, C. Eiras-Franco, J. Hernandez-Castro, and A. Alonso-Betanzos, "Explanation Method for Anomaly Detection on Mixed Numerical and Categorical Spaces," 2022. https://arxiv.org/pdf/2209.04173
[16] Z. Qu, W. Dai, and M. G. Genton, "Global Depths for Irregularly Observed Multivariate Functional Data," 2022. https://arxiv.org/pdf/2211.15125
[17] A. Nieto-Reyes and J. Cabrera, "Statistical Depth based Normalization and Outlier Detection of Gene Expression Data," 2022. https://arxiv.org/pdf/2206.13928
[18] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep Anomaly Detection with Outlier Exposure," 2018. https://arxiv.org/pdf/1812.04606
[19] T. Mathonsi and T. L. van Zyl, "Multivariate Anomaly Detection based on Prediction Intervals Constructed using Deep Learning," 2021. https://arxiv.org/pdf/2110.03393
[20] T. Idé and N. Abe, "Black-Box Anomaly Attribution," 2023. https://arxiv.org/pdf/2305.18440
[21] F. Bachoc, F. Gamboa, M. Halford, J. M. Loubes et al., "Explaining Machine Learning Models using Entropic Variable Projection," 2018. https://arxiv.org/pdf/1810.07924
[22] W. Dai and M. G. Genton, "Directional Outlyingness for Multivariate Functional Data," 2016. https://arxiv.org/pdf/1612.04615.