

Unsupervised Anomaly Detection in Deep Learning: From Autoencoders to Transformers

Virendra Tank¹, Dr. Swati Agarwal², Shivangi Sharma³

¹Assistant Professor, Computer Science, Shri Mahaveer College, Jaipur (Rajasthan) India

²Associate Professor, Computer Science, Poornima University, Jaipur (Rajasthan) India

³Assistant Professor, Computer Science, Shri Mahaveer College, Jaipur (Rajasthan) India

vtank87@gmail.com¹  0009-0003-9126-0982

swati.agarwal@poornima.edu.in²

sharmashivangi2010@gmail.com³

Abstract:

Anomaly detection is a significant problem in various applications ranging from cyber security to manufacturing quality control, which necessitates the ability to detect rare and unusual patterns. With the development of deep learning from traditional methods, it is possible to learn finer-grained data representation and pattern structure in complex high-dimensional information. We thoroughly review unsupervised deep learning methods for anomaly detection in this work, ranging from classical ones to the latest transformer-based. The contributions of our paper are as follows: 1) We conduct a comprehensive survey on state-of-the-art unsupervised deep models with attention mechanisms; 2) We investigate and summarize the applicability and possible variants of them, covering their pros and cons. We study the subfamilies of autoencoders, such as variational and adversarial autoencoders, discuss GAN-based detection techniques and examine recently introduced transformer architectures targeted for outlier detection. By analyzing applications in cyber security, manufacturing defect detection and fraud detection, we explain how these techniques face practical issues while we provide academics with research directions on this fast-evolving domain.

Keywords: Anomaly detection, Unsupervised learning, Autoencoders, GANs, Transformers, Deep learning, Outlier detection

1. Introduction

Anomaly detection (or, outlier or novelty detection) lies in detecting patterns in data that do not conform to expected behavior [1]. Such anomalies often correspond to important information, e.g., network intrusions, defective products in manufacturing, fraudulent transactions in online business and abnormal samples in medicine. The problem is that the anomalies can be very sparse, usually less than 1% of all points and there are few labeled anomalies [2]. This key characteristic makes unsupervised learning techniques especially promising as they can learn normal patterns without dependence on large amount of labeled data.

Machine learning has always been the traditional way to find anomalies, such as Isolation Forest [3] and One-Class SVM (OCSVM) [4]. For their simplicity, explainability widely accepted however, such techniques are unsuitable for high-dimensional data, complex interactions between things if those

interactions are not linear. What will you do about the apparently random patterns in data now appearing everywhere--do you believe it is still appropriate to use these old methods or would you rather switch to new ones offering flexibility in terms of model structure and network configuration while still providing minimal overhead but with comparable performance? Deep learning has totally changed how we do anomaly detection. By making feature learning automatic, very fine data distribution structure is captured, and the algorithm scales to handle gigantic datasets smoothly [6].

Unsupervised methods need not be capable to learn how normal data are compressed in Neural Networks, but are these themselves. As a result, if anomalies violated this logic then it was discovered and they formed. [7] This review examines how from classical methods to modern deep learning twists and turns, with a focus on autoencoders, generative adversarial networks (GANs) and transformers. We will seek the theoretical foundations and innovative architectures,

as well as practical applications in different sectors such as cyber security, manufacturing or detecting financial frauds.

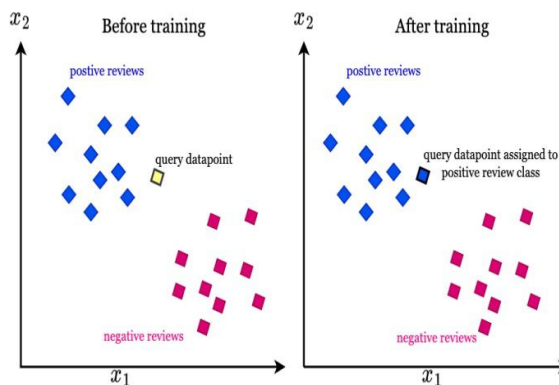
The organization of this review is as follows: Section 2 compares classical methods to those using deep learning techniques; Section 3 is about autoencoder based methods or theories or systems; Section 4 is about GAN-based methods derived from neural networks; Section 5 researches transformer architectures; and Section 6 offers some case studies from real-world application. As seven leads off--the last 1 is also a result of our discussions on Chapter 7.

2. Classical Methods versus Deep Learning Approaches

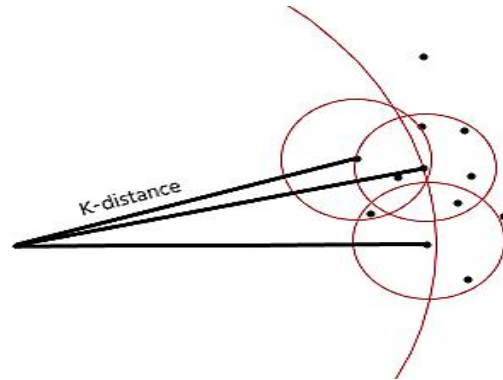
2.1 Traditional Anomaly Detection Techniques

Classical anomaly detection methods provide the foundation for today's techniques. The Isolation Forest's basic premise is that the anomalies are few and different, which makes it easier to isolate them than other instances.[3] The algorithm creates a mean combination that is obtained by averaging all these 'isolation trees' that grow along independent axes in n-dimensions (isolation forests) of real instances. Anomalies require few splits to become isolated in this way. It's as simple as that! This simple split feature can handle high-dimensional data efficiently, but it has trouble with local density variations and complex manifold structures.

Learning a decision boundary around normal instances in high- dimensional feature space regarding [4]. It maps the data into a so-called kernel -induced space and uses this hyperplane to separate normal instances from the origin maximally. In areas where the normal regions are well-delineated moderate- dimensional data Information is limited but effective little challenges come from One-Class SVM. This is not the case with large datasets-normal instances can be anything! Careful selection of suitable kernels and parameters must be made [8].



Other classical techniques include **k-Nearest Neighbors (kNN)**, which identifies outliers by examining the distances from each point to its closest neighbors [9], and



Local Outlier Factor (LOF), a method for measuring local density deviations [10]. Statistical methods such as Gaussian Mixture Models assume the distribution form is known and fixed on a priori basis, thereby greatly reducing their applicability to complex living data [11].

2.2 Advantages of Deep Learning Approaches

Deep learning methodology tackles a number of basic limitations of classical techniques by taking advantage of several key characteristics. Automatic feature learning eliminates the need for manual engineering. This allows neural networks to deduce hierarchical representations of raw data(at least in theory) because its output simply responds to previous values in some fashions wide conceptually similar to more flexible forms like reticular fashion feedback networks via non-linear filters) [6].Convolutional networks extract spatial features from images; recurrent nets capture temporal patterns; and attention mechanisms for locating long-range dependencies can be used alongside smaller networks in creating very simple multi-head structures that are capable of finding interesting hooks in very complex data sets completely automatically without any human intervention.

It is the second advantage giving deep networks non-linear modeling capacity. Neither a simple geometric model nor a statistical model can cope with these complex data distributions [12]. With multiple levels of non-linear transformations, rich representational space can, model even fine normal patterns. Thirdly, scaling up to high-dimensional data facilitates the processing of images, videos, and all kinds of multivariate time series that outclass traditional methods [13].

However, deep learning methods bring problems such as long training times, needing many hyperparameters customization, caught the normal data, and compared with the traditional model, lower interpretability [14]. Whether to choose a traditional or deep learning model depends on the characteristics of the dataset, computing power, interpretability requirements, and normal pattern complexity.

3. Autoencoder-Based Anomaly Detection

3.1 Standard Autoencoders

The essence of deep learning-based anomaly detection is the capability of autoencoders to learn a compressed representation of input data [15]. Autoencoders thus consists of an encoder network, which maps input x to the latent representation $z = f_{enc}(x)$, and a decoder network, which reconstructs the input $\hat{x} = f_{dec}(z)$. In the training process, we minimize reconstruction error $L = ||x - \hat{x}||^2$ for normal data. This forces the network to absorb important patterns and exclude noise and irrelevant changes [16].

Reconstruction error counts as anomaly score in anomaly detection. Ordinary examples that resemble training data will ever credited to their data. Actually in this way, these points can produce low discrimination between anyone of them and any other. Unorthodox examples, however, possess high reconstruction errors. This arises because they diverge from the learned patterns this method uses as a reference [17]. It is just an attempt to guarantee that the brain will not learn these anomalies. What's more, this code makes sure our model will not be able to exist in we assume the autoencoder's limited capacity rules out memorizing all possible patterns, and concentrates instead on normal patterns of higher prevalence.

There are a large number of architectural choices that influence performance. One choice is deep autoencoders with multiple layers of hidden units in them: these have the capacity to capture hierarchical features. Another kind is convolutional autoencoders: such an arrangement can make use of the spatial structure apparent within pictures [18]. Denoising autoencoders take clean inputs and reconstruct corrupted versions of them. This means they learn more stable and error-tolerant ways to represent what needs to be learned by the system [19]. The bottleneck dimension critically balances compression and representation capacity – too small limits expressiveness, while too large enables memorization of anomalies.

3.2 Variational Autoencoders (VAE)

Variational Autoencoders bring probabilistic modeling to the autoencoding process; the sampling probability for latent space is learned rather than fixed encodings [20]. First, the encoder outputs parameters (μ, σ) for Gaussian distribution $q(z|x)$, from which the latent codes are randomly sampled. Reconstructed inputs are then sent through the decoder. At the same time, some KL divergence term helps keep latent space 'nice'--moving it closer to an assigned 'prior' $p(z)$ (typically standard Gaussian).

The VAE aims to amalgamate the loss of reconstruction with regularization so that $L = E_q [||x - \hat{x}||^2] + KL (q (z | x) || P (z) [20]$. With the approach of this form, latent spaces are smooth and continuous--and you get something meaningful when interpolating from point A to B [21]. It provides a variety of anomaly scores for abnormality detection: the probability of reconstruction, the density in latent space and a combination simultaneous multiple values--such as these different anomalies resulted in a regularized log-loss.

But with the ability to model uncertainty and produce varied reconstructions, a VAE can detect anomalies that might go unnoticed by a standard autoencoder. They are subject to collapse of the posterior where the decoder ignores latent codes, and struck a balance must constantly be struck between reconstruction and regularization terms [22]. Such weaknesses have recently been dealt with in modern versions, e.g., the β -VAE [23] and Wasserstein autoencoders [24], for improved regularization schemes.

3.3 Adversarial Autoencoders

Using generative adversarial training Adversarial autoencoders to match the posterior distribution of latent space into any prior distribution [25]. Adversarial autoencoders use a discriminator network, unlike VAEs that use KL divergence as regularization, to tell if samples are from the encoder's distribution or its prior.

With adversarial training, the encoder is forced to generate latent representations that look just like those from prior at the same time as the decoder learns how to turn codes into inputs [25]. With this method we can pick our prior distribution to be any complex or composite one, and we get sharper reconstructions than if using a VAE. To assess anomalies, Anomaly scores of adversarial autoencoders include both the reconstruction error, and the score from the discriminator thus providing fairly robust anomalousness measurements [26].

4. GAN-Based Anomaly Detection

4.1 Fundamentals of GAN-Based Detection

Generative Adversarial Networks consist of a generator G that creates synthetic samples and a discriminator D that distinguishes between real and generated samples [27]. Through adversarial training, the generator learns to produce samples indistinguishable from real data, implicitly modeling the data distribution. This generative capacity enables novel approaches to anomaly detection.

AnoGAN pioneered GAN-based anomaly detection by training a GAN on normal data, then detecting anomalies through their inability to be accurately generated [28]. For a test instance x , AnoGAN searches for a latent code z such that $G(z)$ closely resembles x . The search involves optimizing z to minimize $L(z) = \lambda \cdot \|x - G(z)\|^2 + (1-\lambda) \cdot \|f(x) - f(G(z))\|^2$, where f represents intermediate discriminator features. High optimization losses indicate anomalies poorly represented by the learned normal distribution.

4.2 Advanced GAN Architectures

GANomaly is an improvement over **AnoGAN** because it makes use of an encoder network mapping inputs directly to latent codes [29]. Thus avoiding costly iterative optimization. The design comprises of an E based encoder network, a G based generator and a D based discriminator. In addition there is one extra encoder E' to re-encode the latent codes. In the normal data training phase, by means of latent space reconstruction error and traditional reconstruction error, the efficient anomaly scoring process can now perform normal information detection at speeds which exceed those of yesterday's transmissions-tangibly beneficial to people with heartbeats.

f-AnoGAN successfully enhanced **AnoGAN**, strong following encoder network trained concurrently with GAN is introduced so as to maintain benefits such as the dog is going rightwards while suffering from left front-standing problems in terms of efficiency when implementing this scheme[30]. Leveraging (isolation in the cross talk) combines reconstruction errors in image space, and incline differential analysis discriminator features to catch high-level semantic anomalies.

ALAD (Adversarially Learned Anomaly Detection) should be realized by bidirectional GANs acting together on both forward and backward paths tersely [31]. The architecture consisted of encoder and decoder networks trained adversely, with discriminators operating on both data and latent spaces. This consists of a bidirectional approach for

robust anomaly detection using consistency measures between forward and backward mappings.

4.3 Challenges in GAN-Based Detection

However, GAN-based methods run into numerous problems. The training instability means the generator may end up producing only a limited variety of images [32]. In particular, hyperparameter sensitivity-meaning that the balance between discriminator and generator training must be tuned carefully-is an issue [33]. Finally, concrete examples support this view: immense computational costs To train GANs and, in some methods, optimize latent codes for each test example can also hamper large-scale applications [34].

5. Transformer-Based Anomaly Detection

5.1 Transformer Architectures for Anomaly Detection

Transformers, while originally intended for natural language processing [35], have turned the focus of anomaly detection back through its attention mechanisms that capture both long-range dependencies and intricate relationships. The self-attention mechanism captures every position's relationship to each other in a sequence, lifting the system's ability to detect context based anomalies linked with wider pattern recognition [36].

This presents **Anomaly Transformer**, an architecture for time series anomaly detection that incorporates explicit and direct memory about "is the attention pattern traditional, or abnormal?". The model learns a "Divine Inequality?" that represents anticipated attention pattern changes. This method can be extremely effective in temporal data; point, contextual and collective anomalies are readily detected using it.

In its attentuary training scheme, **TranAD (Anomaly Transformer)** utilizes transformer models for multivariate time series [38]. The architecture has attention mechanisms revealing the relationships between sensors and taking place in time, trained through adversarial objectives that improve sensitivity to anomalies in the data while not impairing its normal part retrievability.

5.2 Vision Transformers for Visual Anomaly Detection

Vision Transformers (ViT) use transformer architecture to process images. The model treats images as sequences of patches and performs self-

attention on them. Returning to application, self-attention from ViT-based anomaly detectors will capture not only spatial relationships but also contextual information throughout an image thus enabling detection for example when there's a small fault that convolutional techniques at best can't yet recognize (or may miss altogether) [40].

PatchCore adopts precursors of vision transformers to Deal information of disable features and patchy images establish a Memory with normal examples [41]. With this feature-space nearest-neighbor technique, Anomaly Detection also contracts the benefits of transformers to essential memory-based techniques. With this in mind it is unsurprising that our method leads when employed in industry defect detection benchmarks.

5.3 Advantages and Limitations

Another advantage of transformers is that their attention mechanisms mean it is possible to interpret an AI's results: whatever area of data the neural net finds important for prediction stands out, and this is called 'saliency' in the literature of deep learning. Furthermore, its architecture naturally handles sequences of variable length. Pretrained models and transfer learning both somewhat remove these deficiencies, a style that has been labeled 'induction bias' [42]. But transformers consume significant computational resources. For example attention mechanisms require large amounts of memory, and without proper regularization can overfit to training data [43].

6. Real-World Applications

6.1 Cybersecurity and Intrusion Detection

The detection of network intrusion is an important cybersecurity application, with anomaly detection being used to recognize malicious activities in network traffic [44]. Deep learning methods can analyze packet-level data, flow statistics, and patterns of behavior for detecting zero-day attacks, advanced persistent threats, and insider threats.

By learning normal traffic patterns, Autoencoders have been deployed in detecting network anomalies [45]. LSTM-based autoencoders capture the temporal dependencies that lie behind order of occurrence in network events, linking attack patterns with temporal artifacts [46]. For GAN-based approaches, simulated scenarios of attacks can be produced at scale and novelty attack types discovered on the basis of distribution differences [47].

Transformer-based methods excel at capturing long-range dependencies in network sessions and user behaviors [48]. Attentional mechanisms help to find odd correlations between distant events that are easy for traditional methods to overlook. Problems need to be addressed include the handling of the high-dimensional feature space, adaptation to new attack strategies, and maintaining a low false positive rate in production environments [49].

6.2 Manufacturing Defect Detection

Product manufacturing quality, detects product surface defects, dimensional anomalies and functional problems which especially need to be controlled [50]. In visual inspection, performed manually in the traditional way, automated anomaly detection can yield significant dividends--especially on high-throughput production lines.

With a CNN autoencoders that has been trained on images of defect-free products "learn" how to turn what they inspect into normal appearances. And where the error rate is high, this means a defect [51]. Diverse defect types are handled by this method, from scratches of various shapes and sizes; dents and surface contamination; or even misalignments without needing examples for training--just common sense and a slightly educated eye. VAEs also offer uncertainty measures that distinguish natural product variations from real defects without false alarms [52]. GAN-based methods, such as GANomaly and f-AnoGAN, have achieved stellar performance on industry-standard datasets like MVTEC AD [53]. They capture subtle anomalies such as missing components, miss-assemblies, texture changes. Using features that have been finely-tuned from training data, Transformer-based approaches like PatchCore can almost detect all defects at a rate of around 100%. They manage to do it with computational efficiency which allows them to be used for real-time inspection [41].

To bring these systems into practical use requires solving a number of problems. Particularly illumination variations; transformation changes in perspective; the rare occurrence of defects which has little data on them; and getting information that can be comprehended readily by human operators [54].

6.3 Financial Fraud Detection

Credit card fraud, insurance fraud, money laundering, and identity theft [55] all fall within the scope of financial scams. The dynamical nature of fraud patterns requires anomaly detection methods to go beyond simply knowing what types of fraud are out there.

Autoencoders analyze series of transactions, learning normal spending patterns and flagging inconsistency [56]. LSTM-autoencoders model the time-based relationships in transaction records. They can detect abnormal episodes where one user's behavior has changed drastically from her accustomed mode, listening to Bob Marley instead of B B King all evening [57]. But there's a trade-off involved between fraud detection sensitivity and false positives that block good transactions.

Different GAN-based approaches are used to generate synthetic fraud cases, which are then intravenously fed into models of classification. This means that the model now does much better than traditional statistical models using decision trees. Supervised classifiers improve as a result [58]. Similarly, GANs trained on legitimate transactions can act as models for the distribution of decision trees, with fraud defined by areas that have low-density within this learned distribution [59]. Methods based on transformers are capable of capturing intricate relationships between attributes of transactions, merchant categories, patterns in geo-location features, and the rhythmic of time [60]

It's important to consider practical issues such as low latency requirements for real-time processing--less than 100 milliseconds. Companies should also think about the number of operations to be performed on what is then sent back out to customers and regulators. On top of these, there are additionally prioritizable factors that come up when tracking how fraud evolves: adaptive learning techniques or methods for privacy protection in forensic finance that protect sensitive information from spilling out into the public sphere [61]

6.4 Medical Anomaly Detection

Medical imaging can be employed to diagnose tumors, lesions and such in X-rays, CT scans and MRI's [62]. When an autoencoder, trained on healthy tissue images, works with another as yet unseen image, it generates a reconstruction error. This error measures the abnormality present in this new image [63]. VAEs provide an anomaly score based on probability, which expresses uncertainty. For medical decisions such as diagnosis it is crucial that patients be made aware.

Recent work has tried applying transformers for medical image analysis, using attention mechanisms so that they concentrate on those parts of an image which are clinically relevant [64]. These methods are able to discover minute pathologies only if one grasps the general context of anatomy and how parts are arranged. Major problems include lack of training data due to confidentiality constraints, the rarity of

each type of aberrant disease and need for clinical workflows that can explain themselves [65].

7. Challenges and Future Directions

7.1 Current Limitations

However, in unsupervised anomaly detection, there remain difficult problems. Because anomalies are less than 0.1% of the data, imbalanced learning makes it hard to train models properly as they tend to ignore rare patterns [66]. For metrics like AUC-ROC, which measure how well a model separates rare anomalies from regular data in practice, we need viable alternatives such as precision-recall curves or anomaly detection accuracy [67].

Accountability for high-stake applications remains an issue, whereas deep Learning models are often unable to explain why such-a place has been identified as anomalous [68]. Furthermore, coping with distribution shift, Alternatives seek to make use of information in new data over time - is something that most existing methods cannot attain [69]. Structural efficiency for real-time applications demands good performance with transformers and complex models that does not compromise detection performance [70].

7.2 Emerging Directions

There are several hopeful prospects in anomaly detection research. Self-supervised learning approaches using pretext tasks exploit the information in unlabeled data to learn representations, which will improve future anomaly detection [71]. Few-shot anomaly detection is about adapting models after observing very small numbers of new instances of anomaly-types [72].

Interpretable anomaly detection combines attention mechanisms, saliency maps, and counterfactual explanations of detected anomalies into human-understandable explanations [73]. Federated anomaly detection Encourages distributed data collaboration that safeguards personal privacy; this especially suitable for medical and financial contexts [74].

Hybrid approaches — for example, combining transformers with traditional methods for feature extraction to achieve both efficiency and interpretability. Continual learning frameworks alter model structures as new data distributions emerge, without spilling over a previously learned pattern [76].

7.3 Future Research Opportunities

Next research needs to solve these problems by working out unified frameworks that smoothly combine many architectures and loss functions,

choosing automatically based on data characteristics [77]. But better theoretical understanding of why certain types of architecture do well on particular anomaly categories would help for rational design choices [78].

Standardizing the benchmarks between domains is even more important. What is more, developing efficient neural architecture search methods especially suited for anomaly detection would mean that the discovery of optimal architectures for diverse applications could be automated [80].

8. Conclusion

From classical statistical and computer science learning techniques to complex deep learning structures, the style of unsupervised anomaly detection has undergone dramatic changes. Autoencoders, including VAEs and adversarial forms, offer amazing reconstruction-based detection by learning normality itself. Methods using GAN-based approaches use generative modeling to learn data distribution complexity and check for distribution anomalies. Attention mechanisms and the modeling of long-range dependence allowed by the transformer let us discover abnormality in context and on a large scale.

Real-world application scenarios in cybersecurity, manufacturing e-commerce apps, finance services and healthcare services highlight the effectiveness of these methods and their practical use-cases. Meanwhile, they also present sticking points include (class imbalance, interpretability); also (how can we make these methods adaptable to different domains), in particular healthcare. efficiency of computation The field continues to advance through innovations in self-supervised learning, few-shot detection, explainability, and continual learning.

Anomaly detection is becoming more and more important in self-driving cars, protecting critical infrastructure and safety-critical applications. This means that we must produce methods that are robustly efficient and interpretable. The convergence of learning paradigms brought by multiple deep neural networks, along with combining it with domain-specific expertise and rigorous evaluation guidelines assures that this is still a critical area of advancement in machine learning.

References

[1] Chandola, V., et al. (2009). "Anomaly Detection: A Survey", ACM Computing Surveys 41 (3):1-58.

[2] Pang, G., Cao, L., van der Hengel, A. (2021) "Deep Learning-Based Anomaly Detection: A Review" ACM Computing Surveys 54 (2):1-38.

[3] Liu, FT, et al. (2008) "Isolation Forest," Proc. ICDM 2008:413-22.

[4] Scholkoph B, Smola A, et al. (1999) "Support Vector Method for Novelty Detection," in NIPS 12(1).

[5] Zimek,A, et al. (2012). "A Survey of Unsupervised Outlier Detection in High-Dimensional Numerical Data" Data Mining and Statistical Analysis 5 (5):363-87

[6] LeCun, Y., Yoshua, V., Geoffrey, H. (2015) "Deep Learning." Nature 521 (7553) 436-44.

[7] Through this unpaid labour, the result is to be obtained without using any Algorithm of loss of interest.

[8] Tax, D., & Duin, R. (2004). Support vector data description. Machine Learning.

[9] Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. SIGMOD.

[10] Christian Even, M., Kedem A.j, Tibshirani R. (1979). LOF: Identifying density-based local outliers. SIGMOD.

[11] Reynolds, D. A. (2009). Gaussian mixture models. In Encyclopedia of Biometrics (741-659).

[12] Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep learning." MIT Press.

[13] Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition.

[14] Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., et al. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE.

[15] Hinton G., & Salakhutdinov R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786):504-507.

- [16] Sakurada, M. and Yairi, T. (2014), "Anomaly detection using autoencoders with nonlinear dimensionality reduction", MLSDA Workshop, 4-11
- [17] R Chalapathy, A K Menon, S Chawla, Anomaly detection using one-class neural networks, arXiv:1802.06360 (2018).
- [18] J Masci, U Meier, D Cireşan, J Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, ICANN, 52-59 (2011).
- [19] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A., "Stacked denoising autoencoders.", JMLR Vol. 11, (2010)., No. 12 (Wang et al 2016).
- [20] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes", ICLR (2014).
- [21] An & Cho, Variational autoencoder based anomaly detection using reconstruction probability, 1-18 (2015)
- [22] J Lucas, G Tucker, R Grosse & M Norouzi (2019) Understanding posterior collapse in generative latent variable models, ICLR Workshop.
- [23] I Higgins, L Matthey, A Pal, et al., "β-VAE: Learning basic visual concepts with a constrained variational framework", ICLR (2017).
- [24] I Tolstikhin, O Bousquet, S Gelly, B Schoelkopf. Wasserstein auto-encoders, ICLR (2018)
- [25] A Makhzani, J Shlens, N Jaitly, I Goodfellow & B Frey (2016). Adversarial autoencoders, ICLR
- [26] H Zenati, C.S Foo, B Lecouat, G Manek & V.R Chandrasekhar, Efficient GAN-based anomaly detection. ICLR Workshop (2018).
- [27] I J Goodfellow, J Pouget-Abadie, M Mirza, et al., "Generative adversarial nets", NIPS (2014).
- [28] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised Anomaly Detection with Generative Adversarial Networks. In: IPMI, pp. 146–157.
- [29] Akcay, S., Atapour-Abarghouei, A., & Breckon, T. P. (2018). Semi-Anomaly GAN: Reducing the Anomaly Discovery Gap using Adversarial Training. In: ACCV, pp. 622–637.
- [30] Schlegl, T., Seeböck, W., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). f-AnomaliesGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis. 54:30-44.
- [31] Zenati, H., Romain, M., Foo, C.S., Lecouat, B., and Chandrasekhar, V. (2018). Adversarially Learned Anomaly Detection. In: ICDM 2018, pp. 727–736.
- [32] Salimans, T., Goodfellow, I., Zaremba, W., et al. (2016). Techniques for training generative adversarial networks continue to improve. In: NeurIPS, 2234–2242.
- [33] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In: ICML, pp. 214–223.
- [34] Di Mattia, F., Galeone, P., De Simoni, M., and Ghelfi, E. (2019). A survey on GANs for anomaly detection. arXiv:1906.11632.
- [35] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. NeurIPS, 5998-6008.
- [36] Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL, 4171–4186.
- [37] Xu, J., Wu, H., Wang, J., & Long, M. (2022). Distilling Anomalies with Association Discrepancy: A New Transformer Model for Time Series Anomaly Detection. In: ICLR.
- [38] Tuli, S., Casale, G., Jennings, N.R., TranAD: Deep transformer networks for anomaly detection in multivariate linear stochastic simulation models. In: VLDB, 15(6):1201-1214.
- [39] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.
- [40] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, & G. L. Foresti., VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization, ISIE, pp. 01-06, 2021.

- [41] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, & P. Gehler.,Towards Total Recall in Industrial Anomaly Detection CVPR, pp. 14318-14328, 2022.
- [42] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, & M. Shah.,Transformers in Vision: A Survey, ACM Computing Surveys, vol. 54, no. 10s, pp. 1-41, 2022.
- [43] Y. Tay, M. Dehghani, D. Bahri, & D. Metzler.,Efficient Transformers: A Survey, ACM Computing Surveys, vol. 55, no. 6, pp. 1-28, 2022.
- [44] and [45] A. L. Buczak, E. Guven A survey of data mining and machine learning methods for cyber security intrusion detection., IEEE Communications Surveys Tutorials, vol. 18, no. 2, pp. 1153-1176, 2016, : Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection, NDSS. Y. Mirsky, T. Doitshman, Y. Elovici, & A. Shabtai,2018
- [46] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, & G. Shroff.,LSTM-based Encoder-Decoder for Multi-Sensor Anomaly Detection, ICML Workshop, 2016.
- [47] M. Ring, D. Schlör, D. Landes,& A. Hotho,Flow-based Network Traffic Generation using Generative Adversarial Networks,Computers Security, vol. 82, pp. 156-172, 2019.
- [48] L. S. Lin, R. Clark, R. Birke, et al.,LSTM-based Network Intrusion Detection using attention mechanism, IEEE Access, vol. 8, pp. 30747-30758, 2020.?
- [49] and [50] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido,& M. Marchetti,On the Effectiveness of Machine and Deep Learning for Cyber Security,CYBER, pp. 371-373, 2018. : P. Bergmann, M. Fauser, D. Sattlegger, & C. Steger.,MVTEC AD: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection, CVPR, pp. 9592-9600, 2019.
- [51] S.,Loewe, P. Bergmann, M. Fauser, D. Sattlegger, & C. Steger,"Improving Unsupervised Defect Segmentation By Applying Structural Similarity To Autoencoders," in VISIGRAPP, pp. 372-380,2019.
- [52] Baur, C., Wiestler, B., Albarqouni, S., & Navab, N. (2019). Delving deep into convolutions will allow us to autoencode data. In Proceedings of the European Conference on Computer Vision (ECCV), pages 161–169.
- [53] Zavrtnik, V., Kristan, M., & Skočaj, D. (2021). DRAEM: a discriminatively trained reconstruction embedding for surface anomaly localisation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8330–8339.
- [54] Czimmermann, T., Ciuti, G., Milazzo, M., et al. (2020). Review of visual defect detection and classification methods in the field industrial production. Sensors, 20(5), 1459.
- [55] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: State of development and challenges. Computers & Security, 57, 47-66.
- [56] Zheng, L., Liu, G., Yan, C., et al. (2018). Researches into the improved Adaboost and its application in transaction fraud detection. The IEEE Transactions on Computational Social Systems, 5(4), 1304-1316.
- [57] Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection Base on deep learning: An auto-encoder and Boltzmann machines in combination. Future Mathematics Journal, 9(1), 18-25.
- [58] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks to improve classification effectiveness in credit card fraud detection. Information Sciences, 479, 448-455.
- [59] Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. (2014). Multi-channel deep CNNs for time series classification. WAIM, 298-310.
- [60] Wang, C., Han, D., Liu, Q., & Luo, S. (2021). P2P Loan credit scoring using deep learning with attention mechanism and LSTM. IEEE Access, 9:211–222.
- [61] Van Vlasselaer, V., Bravo, C., Caelen, O., et al. (2015). APATE: A novel approach to Automated Credit Card transaction fraud detection using network information. Decision Support Systems, 75, 38-48.
- [62] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey of deep learning in medical image analysis Ban de l'e m y. Medical Image Analysis, 42, 60-88.

- [63] Schlegl, T., Seeböck, P., Waldstein, S. M., et al. (2017). Unsupervised anomaly detection by generative adversarial networks guided marker discovery. In Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI), pages 146–157.
- [64] Hatamizadeh, A., Tang, Y., Nath, V., et al. (2022). UNETR: 3D Medical Image Segmentation Using Transformers. WACV, 574-584.
- [65] Shen, D., Wu, G., & Suk, H. I. (2017). Machine Learning in Medical Image Analysis. Annual Review of Biomedical Engineering, 19, 221-248.
- [66] He, H., & Garcia, E. A. (2009). Overcoming the problem of imbalanced data. IEEE TKDE, 21(9), 1263-1284.
- [67] Saito, T., & Rehmsmeier, M. (2015). Precision-recall plots are preferable to ROC plots when evaluating the performance of binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432.
- [68] Guidotti, R., Monreale, A., Ruggieri, S., et al. (2018). Overview of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42
- [69] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). Concept drift in relation to model adaptation. ACM Computing Surveys, 46(4), 1-37.
- [70] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018). Detecting anomalies in spacecraft missions with LSTMs and dynamic thresholding. KDD, 387-395
- [71] Golan, I., & El-Yaniv, R. (2018). Anomaly detection using geometric transformations. arXiv preprint arXiv:1808.08028. NeurIPS, 9781-9791.
- [72] Lu, Y., Xu, P., Tan, J., & Xing, K. (2021). An Introduction to Few-shot Learning for Anomaly Detection. arXiv:2107.08028
- [73] Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Explaining the decisions of deep neural networks, using Layer-wise relevance propagation in MRI-based Alzheimer's disease classification. Frontiers in Aging-cmgt-yüçşh
- [74] Zhao, Y., Zhao, J., Jiang, L., et al. (2020). A Block-Chain-Based Federated Learning Method for Privacy-Preserving IoT Devices. IEEE Internet of Things Journal, 8(3), 1817-1829
- [75] Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., & Rabiee, H. R. (2021). Multiresolution Knowledge Distillation for Anomaly Detection. CVPR, 14902-14912
- [76] De Lange, M., Aljundi, R., Masana, M., et al. (2021). A Survey on Continual Learning Methods: Defeating Forgetting in Classification Tasks. TPAMI, 44(7), 3366-3385
- References [77] C. Zhou and R. C Paffenroth, "Anomaly detection with robust deep autoencoders," proc. KDD 2017, pp. 665–674.
- [78] L. Ruff, R. Vandermeulen, N. Goernitz, et al., "Deep one-class classification," in ICML 2018, pp. 4393–4402.
- [79] G. O. Sequeira Campos, A. Zimek, J. Sander, et al., "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," Data Mining and Knowledge Discovery, vol. 30, no. 4, pp. 891–927, 2016.
- [80] L. Bergman, N. Cohen and Y. Hoshen, "Deep nearest neighbour anomaly detection," 2020, ArXiv preprint ArXiv:2002.10445.