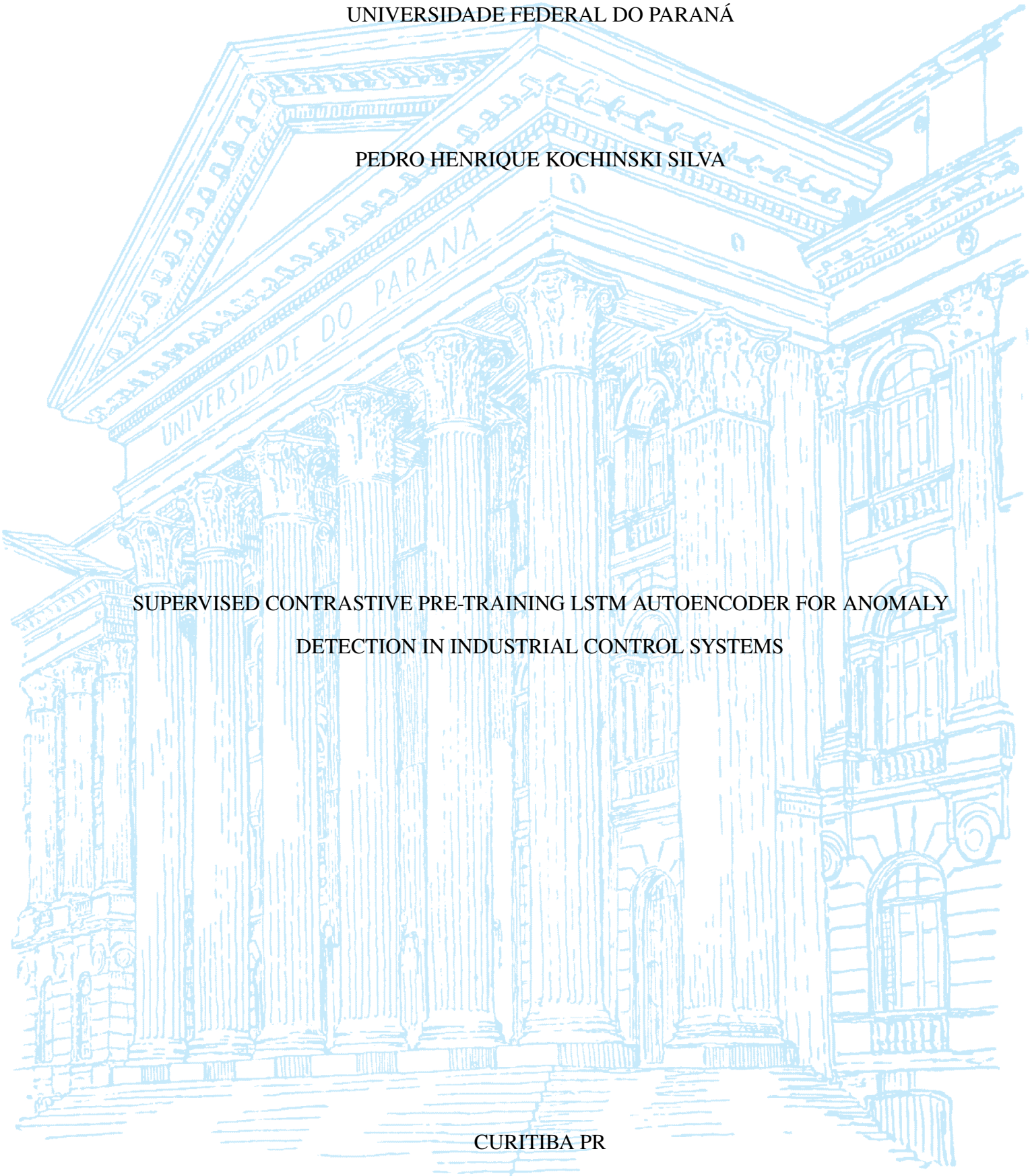
UNIVERSIDADE FEDERAL DO PARANÁ

PEDRO HENRIQUE KOCHINSKI SILVA



SUPERVISED CONTRASTIVE PRE-TRAINING LSTM AUTOENCODER FOR ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEMS

CURITIBA PR

2025

PEDRO HENRIQUE KOCHINSKI SILVA

SUPERVISED CONTRASTIVE PRE-TRAINING LSTM AUTOENCODER FOR ANOMALY DETECTION IN INDUSTRIAL CONTROL SYSTEMS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Fabiano Silva.

CURITIBA PR

2025

RESUMO

A detecção de anomalias em Sistemas de Controle Industrial é crucial para manter a segurança e a integridade operacional de Sistemas Ciberfísicos. Abordagens tradicionais de Aprendizado Profundo, como Autoencoders e Transformers, geralmente dependem do erro de reconstrução como um indicador para a pontuação de anomalias. No entanto, esses métodos baseados em reconstrução frequentemente sofrem de generalização excessiva, onde os modelos reconstroem bem as anomalias, levando a detecções perdidas. Este artigo propõe uma alternativa discriminativa usando um codificador de Memória de Longo Prazo (LSTM) treinado com a Função de Perda Triplet. Em vez de focar na reconstrução, nossa abordagem impõe explicitamente uma margem de separação entre padrões normais e anômalos no espaço latente. Avaliamos essa metodologia no conjunto de dados Secure Water Treatment. Nossos resultados demonstram que o LSTM baseado na Função de Perda Triplet atinge uma pontuação F1 de 0,92, comparável aos modelos Transformer de última geração, mas com sobrecarga computacional significativamente menor e um tamanho de janela temporal reduzido. Além disso, mostramos que inicializar um Autoencoder LSTM com pesos aprendidos por meio de treinamento contrastivo melhora suas capacidades de reconstrução, validando a transferibilidade do espaço métrico aprendido.

Palavras-chave: Triplet-Loss, LSTM, Autoencoders, Detecção de Anomalias, Séries Temporais

ABSTRACT

Anomaly detection in Industrial Control Systems is critical for maintaining the security and operational integrity of Cyber-Physical Systems. Traditional Deep Learning approaches, such as Autoencoders and Transformers, typically rely on reconstruction error as a proxy for anomaly scoring. However, these reconstruction-based methods often suffer from overgeneralization, where models reconstruct anomalies well, leading to missed detections. This paper proposes a discriminative alternative using a Long Short-Term Memory (LSTM) encoder trained with Triplet Loss. Instead of focusing on reconstruction, our approach explicitly enforces a margin of separation between normal and anomalous patterns in the latent space. We evaluate this methodology on the Secure Water Treatment dataset. Our results demonstrate that the Triplet Loss-based LSTM achieves an F1-score of 0.92, comparable to state-of-the-art transformer models, but with significantly lower computational overhead and a reduced temporal window size. Furthermore, we show that initializing an LSTM Autoencoder with weights learned via contrastive training improves its reconstruction capabilities, validating the transferability of the learned metric space.

Keywords: Triplet-Loss, LSTM, Autoencoders, Anomaly Detection, Time Series

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AE	Autoencoder
AUROC	Area Under the Receiver Operating Characteristic
CPS	Cyber-Physical Systems
CPU	Central Processing Unit
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
ICS	Industrial Control Systems
k-NN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MSE	Mean Squared Error
NDT	Non-parametric Dynamic Thresholding
PLC	Programmable Logic Controller
RAM	Random Access Memory
SWaT	Secure Water Treatment
TL	Triplet Loss
TN	True Negative
TP	True Positive
TSAD	Time Series Anomaly Detection
USAD	UnSupervised Anomaly Detection
VAE	Variational Autoencoder

CONTENTS

1 INTRODUCTION

Cyber-Physical Systems (CPS) and Industrial Control Systems (ICS) form the backbone of critical infrastructure, managing processes ranging from water treatment to power distribution. The increasing connectivity of these systems exposes them to cyber-physical attacks, making robust Time Series Anomaly Detection (TSAD) a priority for operational security (Mathur and Tippenhauer, 2016).

The current state-of-the-art in unsupervised TSAD is dominated by reconstruction-based Deep Learning architectures, such as Autoencoders, Variational Autoencoders, and recent Transformer-based models like USAD (Audibert et al., 2020) and TranAD (Tuli et al., 2022). These models operate on the assumption that a network trained exclusively on normal data will fail to reconstruct anomalous sequences, yielding a high reconstruction error that serves as an anomaly score (Chalapathy and Chawla, 2019).

However, reconstruction-based methods face significant limitations. As noted in recent surveys, these models often suffer from overgeneralization, where the network learns to reconstruct anomalous inputs with low error, masking the presence of attacks (Chalapathy and Chawla, 2019). Additionally, in high-dimensional industrial datasets, the reconstruction objective does not necessarily guarantee that normal and anomalous data are linearly separable in the latent space.

To address these shortcomings, this paper investigates a supervised metric learning approach using Triplet Loss (Yang et al., 2019). Unlike reconstruction methods that implicitly model normality, Triplet Loss explicitly optimizes the embedding space to cluster normal samples while pushing anomalous samples away by a defined margin.

We propose an architecture consisting of a Long Short-Term Memory (LSTM) encoder trained with Triplet Loss to generate discriminative embeddings. We validate our approach on the Secure Water Treatment (SWaT) dataset, a realistic testbed for water treatment processes. Our contributions are as follows:

1. We propose a lightweight LSTM-based architecture trained with Triplet Loss (LSTM-TL) that achieves an F1-score of 0.92, competitive with complex Transformer architectures.

2. We demonstrate that our metric learning approach generates superior latent space separability compared to standard LSTM Autoencoders, as verified by k-Nearest Neighbor probes.

3. We explore Transfer Learning, showing that encoders pre-trained with Triplet Loss can improve the performance of reconstruction-based Autoencoders (LSTM-AE-TL) from an F1-score of 0.85 to 0.89.

4. We provide an empirical analysis of the trade-offs between temporal window size and detection accuracy, showing that metric learning enables effective detection even with minimal temporal context (window size of 1).

2 THEORETICAL BACKGROUND

2.1 AUTOENCODERS

Autoencoders are neural architectures designed to learn compact representations of data through an encoder-decoder structure. Given an input vector $x \in \mathbb{R}^d$, the encoder $E(\cdot)$ maps the input to a latent vector $z = E(x)$, and the decoder $D(\cdot)$ reconstructs the input as $\hat{x} = D(z)$ (Bank et al., 2023). Training minimizes the reconstruction loss, typically the mean squared error (MSE):

$$\mathcal{L}_{rec} = \|x - \hat{x}\|_2^2.$$

In Time Series Anomaly Detection (TSAD), autoencoders rely on the assumption that normal patterns are well modeled during training and therefore produce low reconstruction errors, whereas anomalous inputs yield significantly higher errors (Chalapathy and Chawla, 2019). Variants such as Variational Autoencoders (Park et al., 2018) and adversarially trained autoencoders (Audibert et al., 2020) aim to improve robustness against noise, distributional shifts, and overlapping series patterns.

2.2 LONG SHORT-TERM MEMORY NETWORKS

Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are a type of recurrent neural network capable of learning long-range temporal dependencies through gated memory mechanisms. Their structure mitigates vanishing and exploding gradients, enabling them to capture temporal trends in multivariate industrial time series, where correlations among sensors evolve over time.

When used as encoders in LSTM-based autoencoders (LSTM-AE) (Provotar et al., 2019), they compress sequences into fixed-length embeddings while preserving temporal structure. This makes them particularly suitable for anomaly detection in ICS, where abnormal behavior often manifests as temporal deviations.

2.3 EMBEDDINGS FOR TIME SERIES

An embedding is a latent representation that captures the semantic and structural properties of input data (Chalapathy and Chawla, 2019). In TSAD, an effective embedding space should cluster normal samples while separating anomalous ones. Embeddings can be learned:

- implicitly, through reconstruction (AE, LSTM-AE) (Provotar et al., 2019; Srivastava et al., 2015);
- or explicitly, through contrastive learning techniques such as Triplet Loss (Yang et al., 2019).

High-quality embeddings enable downstream anomaly detection using simple classifiers such as k-Nearest Neighbors (kNN), making representation learning crucial when dealing with high-dimensional industrial datasets such as SWaT.

2.4 TRIPLET LOSS AND METRIC LEARNING

Triplet Loss is a metric learning objective that enforces a discriminative embedding space by comparing triplets of samples (Yang et al., 2019). Each triplet consists of an anchor (x_a), a positive sample of the same class of the anchor (x_p), and a negative sample of a different class (x_n). The goal is to ensure that the anchor is closer to the positive than to the negative by a margin m:

$$\mathcal{L}_{tri} = \max(0, d(x_a, x_p) - d(x_a, x_n) + m),$$

where $d(\cdot)$ is a distance metric, typically Euclidean distance.

2.5 QUALITY METRICS FOR ANOMALY DETECTION

Anomaly detection in ICS and CPS is highly imbalanced. Consequently, metrics such as accuracy or AUROC can be misleading (Chalapathy and Chawla, 2019). The F1-score Balances false positives and false negatives, widely used in TSAD benchmarks:

$$\text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}.$$

Matthews Correlation Coefficient (MCC) is a robust metric to heavy class imbalance and is considered one of the most reliable binary classification metrics (Boggia et al., 2025):

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}.$$

Where TP, TN, FP and FN are true and false positive and negative predictions.

2.6 1-NEAREST NEIGHBOR (1-NN) AS A BASELINE

The 1-Nearest Neighbor (1-NN) classifier is a classical and widely adopted baseline in time-series research due to its non-parametric nature and its theoretical consistency guarantees (Boggia et al., 2025; Cover and Hart, 1967). In the context of TSAD, 1-NN is commonly used as a reference benchmark for comparing encoder architectures and metric-learning objectives.

When an encoder produces representations that yield high 1-NN accuracy, it indicates that the latent space preserves discriminative structure, which is desirable for anomaly detection tasks involving metric-based decision rules. Although 1-NN does not capture all temporal dependencies relevant to TSAD, it remains a valuable diagnostic tool for assessing the quality of learned representations and for establishing a consistent comparison across models.

3 RELATED WORK

3.1 THE SWAT DATASET

The Secure Water Treatment dataset serves as the primary benchmark for evaluating the proposed TSAD methodology. SWaT is a scaled-down but fully operational industrial testbed designed to facilitate research in the security of Cyber-Physical Systems (CPS). Unlike simulation-based datasets, SWaT captures the physical properties, response delays, and sensor noise inherent in real-world industrial environments (Mathur and Tippenhauer, 2016).

The complete dataset contains 449,919 multivariate time series records, each representing synchronous readings from 51 sensors and actuators sampled at 1 Hz. The data are labeled into two classes: Normal and Attack. The distribution is highly imbalanced, with 395,298 normal samples and 54,621 attack samples, reflecting the natural predominance of normal operational conditions in industrial systems.

The testbed represents a modern water treatment plant capable of producing 5 US gallons per minute of filtered water. The physical process is divided into six autonomous stages (P1 through P6), each managed by a dedicated dual Programmable Logic Controller (PLC):

- **Stage P1 (Raw Water Intake):** Controls the inflow of water and manages the raw water tank;
- **Stage P2 (Chemical Dosing):** Performs water treatment, such as chlorination, based on water quality metrics;
- **Stage P3 (Ultrafiltration):** Filters feed water using micrometer-sized membranes; this stage includes differential pressure sensors to trigger cleaning cycles;
- **Stage P4 (De-chlorination):** Removes free chlorine using ultraviolet lamps and chemical dosing of sodium bisulphate prior to Reverse Osmosis;
- **Stage P5 (Reverse Osmosis):** The core filtration unit, separating water into permeate and reject streams;
- **Stage P6 (Backwash Cleaning):** Cleans the membranes from Stage P3 using a backwash pump.

The sensors (level, flow, conductivity, pH, etc.) and actuators (pumps, valves) communicate via an EtherNet/IP stack using the Common Industrial Protocol. PLCs share state information across a network, allowing for distributed control decisions.

The SWaT dataset is particularly valuable for supervised metric learning because it contains ground-truth labels derived from controlled cyber-physical attacks. Mathur and Tippenhauer (Mathur and Tippenhauer, 2016) describe the attack model used to generate these anomalies:

- **Attack Vector:** Attacks were executed on the communication network, assuming an attacker with access to the plant network;
- **Mechanism:** Man-in-the-Middle attacks were employed to intercept and manipulate sensor tags and actuator commands in real-time;

- **Impact:** The attacks aimed to drive the system into unsafe states or degrade performance. For instance, manipulating the differential pressure sensor value from 20kPa to 42kPa caused the PLC to initiate an unnecessary backwash cycle. In another scenario, falsifying the level sensor data caused pumps to shut down, significantly reducing water production.

3.2 UNSUPERVISED ANOMALY DETECTION

UnSupervised Anomaly Detection (USAD) (Audibert et al., 2020) introduced an adversarial autoencoder architecture for multivariate TSAD. The model consists of two autoencoders trained jointly: the first reconstructs the input time series, while the second attempts to discriminate between real inputs and reconstructions. This adversarial architecture encourages strong representations and mitigates reconstruction of anomalous patterns.

USAD demonstrated competitive performance on several public datasets, including SMAP, MSL, SMD, and SWaT, while maintaining low computational cost. Its adversarial training improved generalization in noisy industrial environments.

However, USAD still relies entirely on reconstruction error, making it vulnerable to anomalies present in training data, situations where reconstruction quality does not correlate with anomalous behavior, and high-dimensional series where reconstruction becomes excessively smooth.

This motivates approaches like ours, which avoid reconstruction and instead optimize latent discriminability.

3.3 TRANAD

TranAD (Tuli et al., 2022) is a transformer-based model for multivariate time series anomaly detection that integrates attention mechanisms, self-conditioning, and adversarial training. Its principal technical contributions are:

- Transformer-based temporal modeling: multi-head self-attention captures long- and short-term dependencies while enabling fast, parallelizable inference;
- Two-phase inference with self-conditioning: the first reconstruction pass produces a focus score that conditions a second reconstruction, amplifying deviations and enabling improved short-term trend extraction;
- Adversarial training with dual decoders: a minimax objective encourages one decoder to amplify errors and the other to reduce them, improving sensitivity to subtle anomalies;
- Meta-learning: enhances generalization under limited training data.

Despite its strong performance, TranAD has notable limitations:

- Sensitivity to window size: excessively small windows reduce contextual information, while very large windows dilute short anomalies; performance drops for both extremes;
- Modest performance margins on some datasets: Although TranAD achieves the highest average rank across benchmarks, its improvements over simpler baseline models are not uniformly large, and on datasets such as MSL and WADI the performance differences are small or even reversed.

Our work investigates whether simpler models trained with Triplet Loss can achieve comparable performance with significantly lower computational requirements and shorter window sizes.

3.4 BENCHMARKING UNSUPERVISED STRATEGIES FOR TSAD

The benchmark by Boggia et al. (2025) (Boggia et al., 2025) provides the most comprehensive evaluation of unsupervised TSAD algorithms to date. The authors evaluated four primary models: USAD, TranAD, a vanilla Transformer, and the iTransformer (in both reconstruction and forecasting modes) across 10 datasets such as SWaT, WADI, NASA (MSL/SMAP), SMD, GECCO, UCR, and IEEECIS. Key findings of the benchmark include: no single model performs best across all datasets; transformer architectures are highly sensitive to window size and sampling rate; unsupervised methods degrade significantly when anomalies contaminate training data; and embedding quality strongly influences anomaly detection performance.

The benchmark focuses exclusively on unsupervised and self-supervised approaches and does not evaluate supervised metric learning techniques such as Triplet Loss. This gap motivates our proposed method, which:

- Leverages labeled SWaT data to train a discriminative embedding space;
- Avoids reconstruction bottlenecks present in most unsupervised TSAD models;
- Evaluates the learned latent representations using simple classifiers (k-NN);
- Assesses whether smaller window models retain competitive performance.

3.5 DEEP ANOMALY DETECTION METHODS

In an extensive review of state-of-the-art techniques, Chalapathy and Chawla (Chalapathy and Chawla, 2019) categorized deep learning models into supervised, semi-supervised, hybrid, and one-class frameworks. Their research highlighted the key insights regarding model performance and applicability.

Supervised approaches are more accurate than semi-supervised and unsupervised models when labeled data is available. Their main advantages include rapid testing phases, as test instances are compared against a precomputed model. However, they exhibit critical limitations: (i) multi-class supervised techniques require accurate labels for normal and anomalous classes, which are often unavailable; and (ii) deep supervised techniques fail to separate normal from anomalous data if the feature space is highly complex and non-linear. Furthermore, the performance of deep supervised classifiers is often sub-optimal due to class imbalance, where positive class instances far outnumber negative ones.

Semi-supervised methods assume that training instances have only one class label. These approaches leverage labeled data (usually of one class) to produce performance improvements over unsupervised techniques. However, they suffer from specific disadvantages: the hierarchical features extracted within hidden layers may not be representative of the fewer anomalous instances, making these models prone to the over-fitting problem

Unsupervised methods, particularly autoencoders, are widely used as they do not require annotated data for training. They aim to learn inherent data characteristics to separate normal from anomalous points. Despite their cost-effectiveness, the survey emphasizes several critical limitations: (i) it is often challenging to learn commonalities within data in complex and high-dimensional spaces; (ii) the choice of the right degree of compression (dimensionality reduction) is a hyper-parameter that requires tuning for optimal results; and (iii) unsupervised techniques are very sensitive to noise and data corruptions, often resulting in lower accuracy compared to supervised or semi-supervised techniques.

4 PROPOSED APPROACH

The proposed model investigates the use of discriminative latent representations learned by an LSTM encoder trained with Triplet Loss as an alternative to traditional reconstruction-based anomaly detection methods. The central hypothesis is that enforcing explicit separation between normal and anomalous samples in the embedding space can yield performance comparable to or superior to reconstruction-based LSTM autoencoders, while requiring lower computational cost and eliminating the need for long temporal windows. This idea is supported by recent work showing that reconstruction-centric approaches often suffer from generalization issues, causing anomalies to be reconstructed too well (**?**Audibert et al., 2020). Metric-learning strategies such as Triplet Loss offer a promising direction for inducing structured latent spaces with clear inter-class boundaries.

When applied to LSTM encoders, Triplet Loss enables learning class-discriminative embeddings without relying on reconstruction. This is advantageous in ICS datasets, where reconstruction-based models can inadvertently learn to reproduce anomalous patterns, reducing detection reliability.

4.1 MOTIVATION FOR TRIPLET LOSS

Reconstruction-based methods, including LSTM autoencoders, operate under the assumption that the network learns to faithfully reconstruct normal patterns but fails to reconstruct anomalous sequences. While effective in many settings, this paradigm presents several limitations well documented in the literature (**??**Boggia et al., 2025). The weaknesses identified across all categories converge around common themes: (i) a lack of explicit anomaly-aware objectives in feature learning; (ii) overgeneralization in reconstruction-based approaches, causing anomalous inputs to receive low reconstruction error; (iii) difficulty modeling complex, noisy, high-dimensional distributions; and (iv) limited latent-space separability between normal and anomalous samples.

Triplet Loss directly addresses these limitations by optimizing the embedding space for *relative* similarity. Instead of reconstructing inputs, the encoder is trained to ensure that:

$$d(f(x_a), f(x_p)) + m < d(f(x_a), f(x_n)),$$

where x_a is an anchor sample, x_p is a positive (same class), x_n is a negative (different class), and m is a margin. This objective induces two desirable properties: (i) normal samples cluster tightly, despite noise or multi-modality, and (ii) anomalous samples are explicitly pushed away from normal ones. Classical deep anomaly detection methods lack such anomaly-aware representational constraints.

Moreover, Triplet Loss avoids decoder-based overgeneralization, reduces reliance on large temporal windows, and produces embeddings that remain discriminative even in high-dimensional sensor settings. This makes contrastive metric learning a principled and theoretically grounded alternative to reconstruction-based TSAD methods, particularly for industrial multivariate environments such as SWaT.

To address these issues, this work explores a supervised contrastive-learning strategy based on Triplet Loss. Instead of reconstructing input sequences, the LSTM encoder is trained directly to cluster normal samples and push anomalous samples away in latent space. This approach enables explicit control over the embedding and removes the reliance on reconstruction mechanisms. Moreover, unlike transformer-based frameworks such as TranAD, the proposed

approach avoids heavy temporal modeling and significantly reduces computational overhead, making it suitable for industrial environments.

4.2 HYPOTHESES

This work is organized around the following hypotheses:

- **Latent Separability**: An LSTM encoder trained with Triplet Loss produces a latent space with higher separation between normal and anomalous points than the latent space produced by a traditional LSTM autoencoder;
- **Dependence on Training Volume**: Increasing the number of samples per class in contrastive training ($1000 \rightarrow 2000 \rightarrow 5000$) improves the quality and robustness of the learned embeddings, consistent with findings in deep metric learning;
- **Transfer Learning Benefits Autoencoders**: Initializing the encoder of an LSTM autoencoder with weights pre-trained using Triplet Loss provides a more informative starting point, improving its reconstruction capacity and, consequently, anomaly detection performance.

4.3 METHODOLOGICAL STEPS

To validate our hypotheses, the research methodology is structured into four distinct phases: Training of the proposed model, Evaluation of latent space separability, Application of Transfer Learning and, finally, Comparative Evaluation using k-NN of the encoder from LSTM Autoencoder.

1. **Preprocess and Training**: The SWaT dataset contains approximately 13% anomalies. To ensure balanced contrastive training, we construct subsets with 1000, 2000, and 5000 samples per class. The test set remains fixed with 1000 samples of each class. We train the proposed LSTM encoder using Triplet Loss. Crucially, each data point is processed as a 51-dimensional vector with a window size of $w = 1$, $w = 5$ and $w = 10$. This design choice is intentional to investigate if instantaneous features are sufficient for anomaly detection, treating the problem as multivariate classification rather than sequence modeling as shown in Table 6.2;
2. **Proposed Model Evaluation**: After training, embeddings produced by the encoder are evaluated using a k-NN classifier. This step measures the quality of the latent space separability independent of any reconstruction mechanism;
3. **Model integration**: The weights from the contrastively trained encoder (Stage 2) are transferred to initialize the encoder of an LSTM Autoencoder. The encoder is frozen, and only the decoder is trained. This tests if the discriminative manifold benefits the reconstruction task;
4. **Comparative Evaluation**: Conversely, we extract the encoder from a standard reconstruction-based Autoencoder and evaluate it using the k-NN method from Stage 3. This serves to verify if reconstruction objectives naturally induce linear separability in the latent space.

5 EXPERIMENTAL SETUP

All experiments were conducted on a workstation equipped with an NVIDIA GTX 1660 GPU, an AMD Ryzen 5 3600 CPU, and 16 GB of RAM. The models were implemented in PyTorch and executed on the GPU.

To ensure statistical correctness and mitigate the variability inherent to GPU-trained models, each experiment was executed 10 times using distinct random seeds. For all runs, the training and test sets were kept fixed, varying only weight initialization and data shuffling order. The reported results correspond to the mean and standard deviation across the 10 runs, enabling an evaluation not only of the central tendency of each approach but also of its stability. This strategy is particularly relevant given the non-deterministic nature of LSTM operations on GPUs, and the use of aggregated metrics mitigates bias and increases the reliability of model comparisons.

The original dataset is, due to the nature of anomaly detection, highly imbalanced. Balanced subsets containing 1000, 2000, and 5000 samples per class were extracted for the training experiments, while the test set was kept fixed at 1000 samples per class.

For LSTM autoencoder with transfer learning from Triplet-Loss trained encoder, the samples selected for pretraining were removed from the corresponding subsets.

We designed four specific experimental configurations to compare the proposed method against the baseline and evaluate the impact of transfer learning:

1. **LSTM-AE (Baseline)**: The standard reconstruction-based autoencoder used as a reference;
2. **LSTM-TL (Proposed)**: The standalone metric-learning encoder trained with Triplet Loss;
3. **LSTM-AE-TL (Hybrid)**: The autoencoder initialized with weights from the LSTM-TL model;
4. **AE-Encoder (Analysis)**: The encoder extracted from the baseline LSTM-AE evaluated as a feature extractor.

The specific architecture and training details for each configuration are described below.

5.1 LSTM AUTOENCODER (LSTM-AE)

The first approach consists of a faithful reproduction of the LSTM autoencoder proposed by (Boggia et al., 2025), used here as a reference baseline. The original pipeline applies a downsampling factor of 10 to reduce temporal redundancy and decrease training cost. After downsampling, the time series are segmented into windows of size $w = 10$, each containing measurements from 51 variables.

The model follows a classical autoencoder structure: the LSTM encoder receives input sequences of shape (10, 51) and outputs a 64-dimensional embedding. The LSTM decoder reconstructs the original sequence from this embedding. A final Fully Connected layer adjusts the output shape to reproduce the (10×51) matrix.

Training is performed exclusively on normal samples, under the assumption that abnormal patterns will yield substantially higher reconstruction errors. The loss function used is

Mean Squared Error (MSE), which is standard in reconstruction-based approaches for anomaly detection.

In this setting, the classical LSTM autoencoder outperformed both the F1-score of 0.74 reported in (Wang et al., 2023) and other LSTM-based variants, including the LSTM-VAE proposed in (Audibert et al., 2020) (F1-score of 0.80) and the forecasting-based LSTM-NDT introduced in (Tuli et al., 2022) (F1-score of 0.61). Under the same evaluation protocol, the classical LSTM autoencoder achieved an F1-score of 0.85, as shown in Table 6.3.

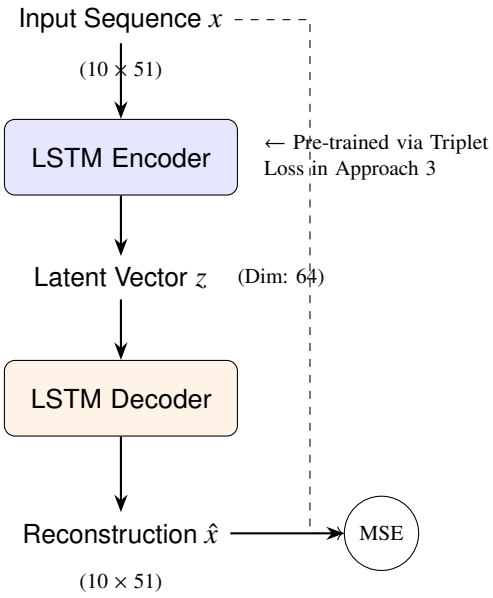


Figure 5.1: **LSTM-AE Architecture (Approach 5.1 & 5.3).** In the Transfer Learning approach (Approach 3), the 'LSTM Encoder' weights are initialized from the Triplet Loss model and frozen, while the Decoder is trained to minimize Reconstruction Error (MSE).

5.2 LSTM ENCODER WITH TRIPLET LOSS (LSTM-TL)

The second approach investigates the use of contrastive learning to capture separability between normal and anomalous samples. Unlike Approach 5.1, which relies on temporal reconstruction, this approach focuses solely on learning discriminative embeddings using the LSTM as an encoder.

Each record is treated as a 51-dimensional vector. Although this removes temporal structure, the underlying hypothesis is that SWaT anomalies often manifest as instantaneous or short-term deviations, enabling meaningful separation even without temporal windows.

During training, balanced triplets (anchor, positive, negative) are constructed. The LSTM encoder maps each sample to a 64-dimensional embedding, and Triplet Loss enforces the formation of two distinct clusters: normal and anomalous. Regarding hyperparameters, a margin of 0.1 was used for the 1000 and 2000 sample subsets with a batch size of 200 and a learning rate of 0.0001. For the larger 5000 sample subset, we increased the batch size to 700 and the learning rate to 0.002 to maintain convergence stability, while reducing the margin to 0.001.

After training, embeddings are evaluated using k-NN classification to assess cluster separability in the latent space.

This setting achieved a maximum F1-score of 0.93 when using a window size of $w = 1$, supported by the results in Table 6.2

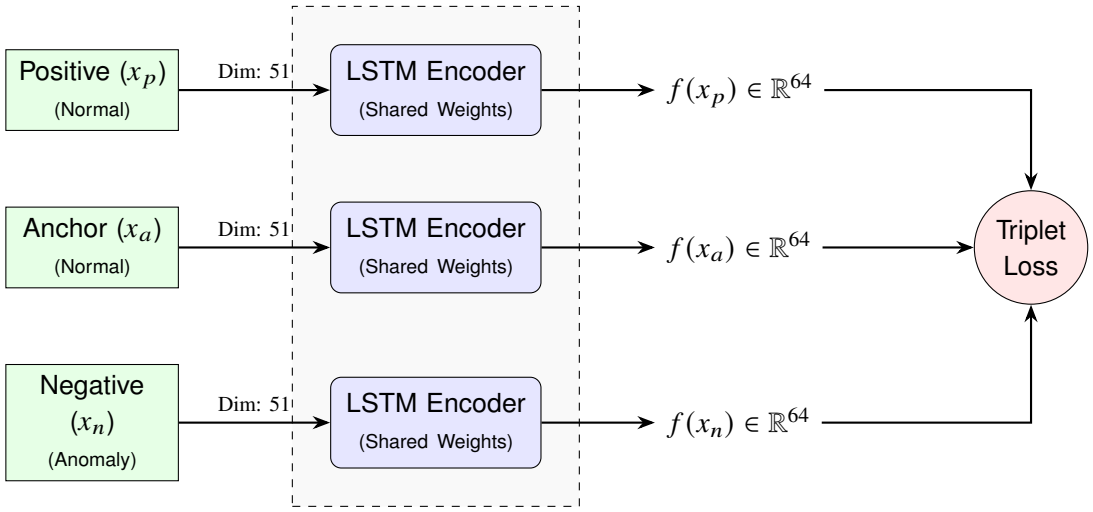


Figure 5.2: **LSTM-TL Architecture (Approach 5.2).** The model uses a triplet configuration (Anchor, Positive, Negative) to learn a discriminative embedding space. The LSTM Encoder maps 51-dimensional inputs ($w = 1$) to a 64-dimensional latent space optimized by Triplet Loss.

5.3 AUTOENCODER WITH PRETRAINED ENCODER (LSTM-AE-TL)

This approach integrates contrastive learning with reconstruction. First, an encoder is trained following Approach 5.2 to learn explicitly discriminative representations. Its weights are then transferred to the encoder of the autoencoder in Approach 5.1. At this point, the encoder is frozen and only the decoder is trained as shown in figure 5.2.

The objective is to evaluate whether embeddings obtained through contrastive learning provide a more informative initialization for reconstruction, potentially reducing reconstruction error in regions that are critical for class separation. The results show an improvement in F1-score from 0.85 to 0.89, together with an increase in MCC from 0.84 to 0.88, indicating that autoencoders can benefit from discriminative latent representations. These results were achieved when using 5000 samples from each class to train the LSTM encoder with Triplet Loss (Approach 5.2).

5.4 EVALUATION OF THE AUTOENCODER AS AN EMBEDDING GENERATOR

The fourth approach investigates whether the encoder trained for reconstruction in Approach 5.1 indirectly captures features capable of distinguishing normal and anomalous patterns. The encoder weights are frozen and used as embedding extractors, followed by k-NN classification.

This analysis assesses whether sequential reconstruction drives the encoder to learn semantically meaningful representations for TSAD, even without explicit supervision. The results show that this configuration yields an F1-score of 0.61, indicating that the learned embeddings are considerably less discriminative than those produced by contrastive learning.

6 RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained from the four evaluated approaches. As an initial step, we define a baseline to enable consistent comparisons. We adopt the LSTM Autoencoder proposed in (Boggia et al., 2025), whose implementation serves as a performance reference.

6.1 ANALYSIS OF TEMPORAL DEPENDENCIES AND WINDOW SIZE

A notable distinction in our experimental setup is the difference in window sizes between the baseline LSTM-AE ($w = 10$) and the proposed LSTM-TL ($w = 1$). While the baseline relies on a temporal context of 10 steps to reconstruct the signal, our metric-learning approach operates on instantaneous vectors.

The fact that the LSTM-TL model achieves an F1-score of 0.92 with a window size of $w = 1$ and it drops to 0.80 with $w = 10$ provides critical insight into the nature of anomalies within the SWaT dataset. In time series literature, anomalies are typically categorized as point anomalies or collective anomalies (sequences that are anomalous only when viewed over time).

The high performance of our model suggests that the majority of attacks in the SWaT testbed manifest primarily as point anomalies or distinct operational state changes that do not require long-term historical context to detect. This implies that the temporal dependencies captured by complex Transformer models or large-window Autoencoders may be redundant for this specific dataset. By treating the problem as a multivariate classification task rather than a sequence modeling task, the LSTM-TL approach eliminates the computational overhead of processing temporal windows without sacrificing detection accuracy.

6.2 PROPOSED MODEL PERFORMANCE

The proposed model shows strong F1-Score, outperforming the approach currently used in the original system. With an score of 0.92, its performance is comparable to state-of-the-art methods in time-series anomaly detection. Furthermore, the model supports transfer learning, enabling its encoder to be reused in autoencoder-based and classifier-based methods as shown in table 6.2.

The results show that increasing k from 1 to 5 yields a small improvement in F1-score, indicating greater robustness to noise in the embeddings. The transfer learning variant trained with larger subsets (5000 samples per class) maintains strong performance (F1 = 0.9204), demonstrating that the model benefits from additional data even without architectural changes.

The results of loading the pretrained encoder in the LSTM Autoencoder shows an F1-Score increse from 0.85 to 0.89.

Conversely, the approach using the pretrained encoder from the LSTM Autoencoder seen in 5.4 exhibits a significant drop in F1-score (0.6812). This suggests a potential mismatch between the latent space learned through reconstruction and that required by triplet loss. As a result, the embeddings become insufficiently discriminative for k-NN classification.

It was also observed that, particularly in the SWaT dataset, the embeddings produced by the model exhibited substantial overlap between normal and anomalous samples. This intrinsic characteristic of the dataset necessitated the use of a very small margin during training in order to obtain any meaningful separation in the embedding space. Consequently, both the number of

Table 6.1: Hyperparameters and Performance of LSTM with Triplet-Loss

Hyperparameter	**LSTM-TL**	**LSTM-TL (Low Data)**	**LSTM-TL (High Data)**	**LSTM Encoder Only**
k	1	5	5	5
Batch Size	200	200	700	100
Window Size	1	1	1	1
Epochs	145	145	220	10
Learning Rate	0.0001	0.0001	0.002	0.0001
Optimizer	Adam	Adam	Adam	Adam
Margin	0.1	0.1	0.0001	N/A
Training	2000 (per class)	2000 (per class)	5000 (per class)	49500
Test Samples	1000 (per class)	1000 (per class)	1000 (per class)	1000 (per class)
F1 Score	0.91 ± 0.04	**0.92 ± 0.03**	0.92 ± 0.03	0.61 ± 0.11

training epochs and the learning rate had to be increased to achieve competitive F1-scores when dealing with a large number of samples.

Table 6.2: Hyperparameters and Performance of LSTM-TL with Varying Window Sizes

Model	**Window Size (w)**	**F1-Score**
LSTM-TL	**1**	**0.92 ± 0.03**
LSTM-TL	5	0.77 ± 0.10
LSTM-TL	10	0.80 ± 0.07

Table 6.3: Performance of Autoencoder-Based Models

Model	**F1 Score**	**MCC**
LSTM-AE (Baseline)	0.85 ± 0.03	0.84 ± 0.02
LSTM-AE-TL	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.88 \pm 0.01}$

7 CONCLUSION AND FUTURE WORK

This study investigated a discriminative alternative to reconstruction-based Time Series Anomaly Detection (TSAD) by training an LSTM encoder with Triplet Loss to produce separable latent representations of normal and anomalous samples in the SWaT dataset. Unlike conventional autoencoder architectures, which rely on reconstruction error as an anomaly score, the proposed metric-learning formulation explicitly enforces inter-class separation in the embedding space. The results demonstrate that this strategy is highly effective as a standalone method. Furthermore, it proves competitive with traditional LSTM autoencoders when used to initialize the encoder weights via transfer learning, achieving these results with a window size of one and substantially lower computational overhead compared to transformer architectures

The evaluation shows that models trained with Triplet Loss consistently yield latent representations with superior separability, as measured by a k-NN classifier. This confirms the first hypothesis regarding enhanced discriminative structure in the embedding space. Additionally, increasing the number of balanced training samples per class was shown to improve robustness and stability across runs, validating the hypothesis concerning sensitivity to training volume. Furthermore, initializing an LSTM autoencoder with weights obtained from contrastive training provided tangible benefits: although only the decoder was trained, the resulting model exhibited improved reconstruction behavior, supporting the transfer learning hypothesis. Together, these findings indicate that metric-learning representations can serve as more stable and informative priors than those obtained through random initialization.

Beyond performance, the proposed approach achieved these results with significantly reduced computational complexity. By eliminating long temporal windows, decoder networks, and iterative reconstruction stages common in state-of-the-art transformer-based methods such as iTransformer, the Triplet-Loss LSTM encoder offers a lightweight and scalable alternative suitable for industrial environments with stringent latency and resource constraints. However, some limitations remain. First, the method requires labeled anomalies, which are often scarce in real industrial systems. Second, the embedding dimension and margin hyperparameters exert strong influence on performance, requiring careful tuning. Third, the LSTM-TL model does not capture long-range temporal dependencies, which may be relevant in datasets where anomalies manifest gradually rather than instantaneously. Finally, although contrastive training reduces overgeneralization, it does not address potential domain shifts across processes or sensor configurations.

Future work should explore several directions to extend this research. A natural next step is to evaluate larger or more diverse ICS datasets, including WADI, SMD, and GECCO, to assess generalization beyond SWaT. Another promising direction involves hybridizing the proposed embedding space with lightweight temporal models, such as temporal convolutional layers or attention blocks, to reintroduce temporal context without incurring the computational burden of full transformer architectures. Further investigation into hard triplet mining strategies could improve training efficiency and reduce sensitivity to margin selection. Finally, deploying the model on real-time telemetry streams from industrial systems (such as agricultural machinery or distributed sensor networks) would provide important evidence regarding robustness under domain shift, sensor drift, and online operational constraints.

In summary, this work demonstrates that Triplet-Loss-based metric learning is an effective and computationally efficient alternative to reconstruction-based anomaly detection methods in multivariate industrial time series and regular autoencoder architectures can benefit

from discriminative embeddings. By explicitly shaping the latent space, the proposed approach addresses several core limitations of deep reconstruction models and provides a solid foundation for future research on anomaly detection for cyber-physical systems.

REFERENCES

Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. (2020). Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3395–3404.

Bank, D., Koenigstein, N., and Giryes, R. (2023). Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374.

Boggia, L., de Lima, R. T., and Malaescu, B. (2025). Benchmarking unsupervised strategies for anomaly detection in multivariate time series. *arXiv preprint arXiv:2506.20574*.

Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mathur, A. P. and Tippenhauer, N. O. (2016). Swat: A water treatment testbed for research and training on ics security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*, pages 31–36. IEEE.

Park, D., Hoshi, Y., and Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551.

Provotar, O. I., Linder, Y. M., and Veres, M. M. (2019). Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pages 513–517. IEEE.

Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR.

Tuli, S., Casale, G., and Jennings, N. R. (2022). Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*.

Wang, F., Wang, K., and Yao, B. (2023). Time series anomaly detection with reconstruction-based state-space models. In *International Conference on Artificial Neural Networks*, pages 74–86. Springer.

Yang, Y., Chen, H., and Shao, J. (2019). Triplet enhanced autoencoder: Model-free discriminative network embedding. In *IJCAI*, pages 5363–5369.