



Delft University of Technology

Unmasking the Unexpected Towards Reliable Time Series Anomaly Detection

Ghorbani, R.

DOI

[10.4233/uuid:7baca279-cd15-4141-89fe-54e3073e164e](https://doi.org/10.4233/uuid:7baca279-cd15-4141-89fe-54e3073e164e)

Publication date

2025

Document Version

Final published version

Citation (APA)

Ghorbani, R. (2025). *Unmasking the Unexpected: Towards Reliable Time Series Anomaly Detection*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:7baca279-cd15-4141-89fe-54e3073e164e>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

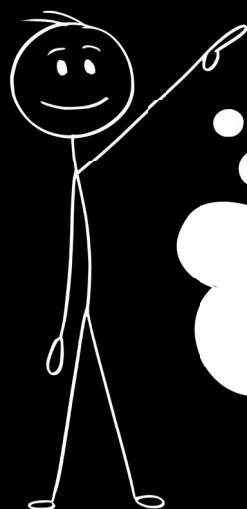
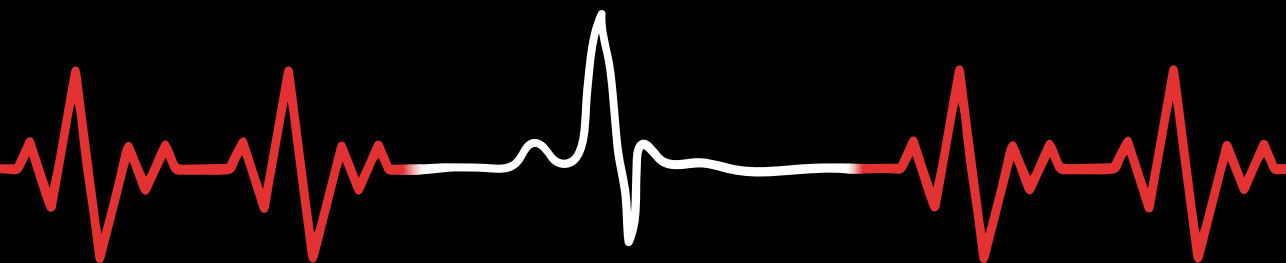
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

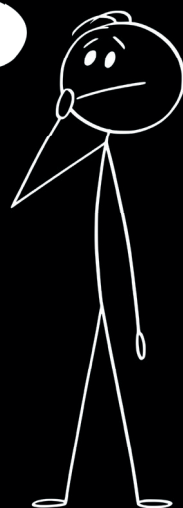
Unmasking the Unexpected

Towards Reliable Time Series Anomaly Detection



*Look, that is not
a normal Pattern!*

Are you Sure?!



UNMASKING THE UNEXPECTED

TOWARDS RELIABLE TIME SERIES ANOMALY DETECTION

UNMASKING THE UNEXPECTED
TOWARDS RELIABLE TIME SERIES ANOMALY DETECTION

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Tuesday, the 24th of June 2025 at 10:00 o'clock

by

Ramin GHORBANI

Master of Science in Industrial Engineering,
Iran University of Science and Technology, Iran
born in Esfahan, Iran

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T Reinders,	Delft University of Technology, <i>promotor</i>
Dr. D.M.J. Tax,	Delft University of Technology, <i>copromotor</i>

Independent members:

Prof.dr. J.J. van den Dobbelsteen,	Delft University of Technology
Prof.dr. K.J. Batenburg,	Leiden University
Dr.ir. R.C. Hendriks,	Delft University of Technology
Dr. X. Long,	Eindhoven University of Technology
Dr. I.A.C. van der Bilt,	University Medical Center Utrecht
Prof.dr. A. Hanjalic,	Delft University of Technology, <i>reserve member</i>



Keywords: Time Series, Anomaly Detection, Evaluation Metrics, Representation learning, Self-Supervised Learning, Photoplethysmogram (PPG), Inter-Subject Variability

Printed by: Proefschrift Specialist (www.proefschriftspecialist.nl)

Copyright © 2025 by R. Ghorbani

ISBN/EAN: 978-94-6518-074-8

An electronic copy of this dissertation is available at
<https://repository.tudelft.nl/>.

*To my parents, my brother, and Marieke — for your love, unwavering support, and
constant encouragement that carried me through this journey.*

CONTENTS

Summary	ix
Samenvatting	xi
1. Introduction	1
1.1. Time Series Data	2
1.2. Anomalies in Time Series Data	3
1.3. Anomaly Detection in Time Series Data	5
1.3.1. Supervised Anomaly Detection	5
1.3.2. Unsupervised Anomaly Detection	6
1.3.3. Challenges in Time Series Anomaly Detection	6
1.4. Foundations of the Thesis and Outline	13
References	14
2. PATE: Proximity-Aware Time Series Anomaly Evaluation metric	21
2.1. Introduction	22
2.2. Proposed Evaluation Metric - PATE	24
2.2.1. Categorizing the Events	25
2.2.2. Weighting Process	27
2.2.3. PATE Final Score	29
2.3. Experiments and Results	30
2.3.1. Synthetic Data Experiments	30
2.3.2. Real-World Data Experiments	31
2.3.3. Impact Analysis: SOTA Models	33
2.4. Ablation Analysis: Buffer Sizes	35
2.5. Discussion and Conclusion	36
References	36
Appendices	39
3. Self-Supervised PPG Representation Learning shows High Variability	45
3.1. Introduction	46
3.2. Proposed Framework	47
3.3. Experimental Setup	48
3.3.1. Dataset	48
3.3.2. Data Preprocessing	49
3.3.3. Implementation	49
3.4. Results	50
3.4.1. Pretext Task	50
3.4.2. Downstream Task	51

3.5. Discussion and Conclusion	52
References	53
4. Representation Learning and Personalization for PPG Anomaly Detection	59
4.1. Introduction	60
4.2. Proposed Framework	62
4.2.1. Definition of Anomalies	64
4.3. Experimental Setup	65
4.3.1. Datasets	65
4.3.2. Data Preprocessing	65
4.3.3. Implementation	65
4.3.4. Anomaly Detection Evaluation Scenarios	66
4.4. Results	69
4.4.1. Representation Learning	69
4.4.2. Anomaly Detection	69
4.4.3. Robustness of Representation Dimensionality	70
4.5. Discussion and Conclusion	71
References	74
5. RESTAD for Sensitive Time Series Anomaly Detection	79
5.1. Introduction	80
5.2. Methodology	81
5.2.1. RESTAD Framework	82
5.3. Experimental Setup	83
5.3.1. Datasets and Preprocessing	83
5.3.2. Implementation	83
5.4. Results	84
5.4.1. Ablation Analysis	85
5.5. Discussion and Conclusion	86
References	87
6. Discussion and Conclusion	93
6.1. Evaluation Metrics for Time Series Anomaly Detection	93
6.2. Representation Learning for Unsupervised Time Series Anomaly Detection	94
6.3. Anomaly Scoring Mechanism for Subtle Time Series Anomalies	95
6.4. Final Words	96
Acknowledgements	97
Curriculum Vitæ	101
List of Publications	103

SUMMARY

The integration of wearable technology into healthcare is revolutionizing health monitoring by enabling continuous tracking of vital metrics like heart rate and blood sugar. Devices such as smartwatches and glucose monitors empower proactive interventions, reducing hospital visits and personalizing care. For instance, wearables can detect irregular heart rhythms for early cardiovascular disease detection or assist individuals with diabetes in managing glucose levels. These advancements are enabled by technologies like photoplethysmography (PPG), a non-invasive method for real-time monitoring of physiological signals. Continuous monitoring generates time series data that captures dynamic health fluctuations over time. This data allows for identifying irregularities and deviations that isolated measurements might miss. Detecting anomalies, such as abrupt changes in heart rate or prolonged abnormal patterns, is essential for timely interventions in managing chronic conditions like hypertension and cardiovascular diseases.

However, the analysis of time series data introduces challenges. For instance, label scarcity arises because labeling health anomalies requires expert input, which is often infeasible for large datasets. Inter-subject variability becomes a concern as physiological patterns differ significantly across individuals, complicating model generalization. Furthermore, temporal dependencies in time series data add complexity, as observations are sequentially related and anomalies may not manifest as isolated points but as patterns or sequences deviating from normal behavior. Detecting subtle anomalies, minor deviations that accumulate over time but may signal early-stage conditions, becomes particularly challenging due to their resemblance to normal temporal variations and noise in the data. For example, a gradual change in heart rate variability might indicate the onset of an irregular rhythm but could easily blend into inherent variability if not carefully analyzed. Moreover, evaluation metrics for time series data are insufficient, failing to capture the temporal complexities of real-world applications. Consequently, conventional metrics can misrepresent model performance, leading to unreliable or misleading assessments.

This thesis addresses these challenges by advancing time series anomaly detection through innovative methodologies. A key focus is on addressing the limitations of existing evaluation metrics by introducing new evaluation metrics to better capture temporal complexities, ensuring reliable and meaningful performance assessments. Beyond evaluation, this thesis is guided by several core principles to address the challenges inherent in time series data analysis. Central to this is the use of unsupervised representation learning to tackle label scarcity and variability, enabling robust feature extraction from unlabeled data while maintaining generalizability. Finally, the thesis develops strategies for increasing sensitivity to subtle anomalies, providing effective solutions for identifying small yet significant deviations in complex

datasets. Together, these contributions present a comprehensive framework for improving anomaly detection systems across diverse applications, bridging theoretical advancements with practical, real-world needs.

SAMENVATTING

De integratie van draagbare technologie in de gezondheidszorg revolutioneert in de manier waarop gezondheid wordt gemonitord, doordat het continue tracking van vitale functies zoals hartslag en bloedsuiker mogelijk maakt. Apparaten zoals smartwatches en glucosemeters maken proactieve interventies mogelijk, verminderen ziekenhuisbezoeken en personaliseren de zorg. Zo kunnen wearables onregelmatige hartritmes detecteren voor vroege opsporing van hart- en vaatziekten, of mensen met diabetes helpen bij het beheren van hun glucosespiegels. Deze vooruitgang wordt mogelijk gemaakt door technologieën zoals fotoplethysmografie (PPG), een niet-invasieve methode voor het real-time monitoren van fysiologische signalen. Continue monitoring genereert tijdreeksdata die dynamische gezondheidsfluctuaties in de tijd vastlegt. Deze data maakt het mogelijk om onregelmatigheden en afwijkingen te identificeren die bij geïsoleerde metingen mogelijk onopgemerkt blijven. Het detecteren van anomalieën, zoals plotselinge veranderingen in hartslag of langdurig abnormale patronen, is essentieel voor tijdige interventies bij het beheer van chronische aandoeningen zoals hypertensie en hartziekten.

Het analyseren van tijdreeksdata brengt echter uitdagingen met zich mee. Een voorbeeld hiervan is het tekort aan gelabelde data, omdat het labelen van gezondheidsanomalieën deskundige input vereist, wat vaak onhaalbaar is bij grote datasets. Interpersoonlijke variabiliteit vormt ook een uitdaging, aangezien fysiologische patronen sterk verschillen tussen individuen, wat de generalisatie van modellen bemoeilijkt. Bovendien voegen temporele afhankelijkheden in tijdreeksdata extra complexiteit toe, omdat observaties sequentieel met elkaar verbonden zijn en anomalieën zich mogelijk niet manifesteren als geïsoleerde punten, maar als patronen of reeksen die afwijken van normaal gedrag. Het detecteren van subtiele anomalieën, kleine afwijkingen die zich over tijd opstapelen maar mogelijk wijzen op een vroege aandoening, is bijzonder uitdagend vanwege hun gelijkenis met normale temporele variaties en ruis in de data. Een geleidelijke verandering in hartslagvariabiliteit kan bijvoorbeeld wijzen op het begin van een onregelmatig hartritme, maar kan gemakkelijk opgaan in de natuurlijke variatie als dit niet zorgvuldig wordt geanalyseerd. Bovendien schieten evaluatiemetrieken voor tijdreeksdata tekort; ze houden geen rekening met de temporele complexiteit van realistische toepassingen. Hierdoor kunnen conventionele metrieken de modelprestaties verkeerd weergeven, wat leidt tot onbetrouwbare of misleidende evaluaties.

Dit proefschrift pakt deze uitdagingen aan door de detectie van anomalieën in tijdreeksdata te verbeteren met innovatieve methodologieën. Een belangrijk aandachtspunt is het aanpakken van de beperkingen van bestaande evaluatiemetrieken door het introduceren van nieuwe metrieken die de temporele complexiteit beter vastleggen, zodat prestaties op betrouwbare en betekenisvolle wijze kunnen worden

beoordeeld. Naast evaluatie wordt dit proefschrift geleid door verschillende kernprincipes om de uitdagingen in tijdreeksanalyse aan te pakken. Centraal hierin staat het gebruik van unsupervised learning om het tekort aan labels en variabiliteit te overwinnen, waardoor robuuste kenmerken kunnen worden geëxtraheerd uit ongelabelde data met behoud van generaliseerbaarheid. Tot slot ontwikkelt het proefschrift strategieën om de gevoeligheid voor subtiele anomalieën te vergroten, en biedt het effectieve oplossingen voor het identificeren van kleine maar significante afwijkingen in complexe datasets. Samen vormen deze bijdragen een allesomvattend raamwerk voor het verbeteren van systemen voor anomaliedetectie in uiteenlopende toepassingen, waarbij theoretische vooruitgang wordt verbonden met praktische, reële behoeften.

1

INTRODUCTION

THE integration of technology into healthcare is revolutionizing how we monitor and manage our health. One of the most promising advancements in recent years has been the rise of remote patient monitoring, which enables continuous, seamless health tracking, largely facilitated by wearable devices [1]. These devices, such as smartwatches and fitness trackers, allow individuals to stay connected to their health data and enable healthcare providers to monitor patient conditions in real time. For instance, a smartwatch can alert someone with a heart condition if their heart rate becomes irregular, prompting immediate care [2]. Similarly, wearable glucose monitors help patients with diabetes track blood sugar levels continuously, enabling more responsive adjustments in medication or diet [3].

Continuous monitoring is particularly valuable because health metrics can fluctuate widely throughout the day. A one-time measurement taken during a hospital visit might not capture a patient's typical health status or warning signs that appear at other times [4]. For cardiovascular health, for instance, subtle fluctuations in heart rate or irregularities can signal potential issues that may not be detected in a one-time hospital visit [5]. Wearable devices, by providing an always-on connection to vital health metrics, bridge the gap between patients and healthcare professionals, reducing the need for frequent hospital visits and enabling more personalized care [6]. For many conditions, such as hypertension or heart disease, critical changes in health metrics can occur between appointments, and timely interventions can mean the difference between a manageable situation and a health crisis. Continuous monitoring, therefore, not only empowers patients but also transforms the healthcare system, fostering a proactive approach that can significantly enhance patient outcomes [7].

One of the key technologies enabling this continuous, non-invasive tracking in wearable devices is photoplethysmography (PPG), which measures blood volume changes by emitting light onto the skin and detecting variations in light absorption [8]. PPG is especially useful for cardiovascular monitoring, allowing wearables to capture ongoing data about heart rate and detect irregularities that might signal early health risks, such as arrhythmias or hypertension. This steady stream of data allows healthcare providers a more accurate view of patient trends, supporting proactive interventions and enabling patients to better manage their health with fewer hospital visits [9]. Figure 1.1 illustrates the process of how wearable devices utilize PPG technology for continuous health monitoring. The diagram shows how light-based

PPG sensors in wearable devices capture variations in blood volume, generate physiological signals, and transmit this data to healthcare systems. This enables real-time monitoring not only by patients but also by doctors, who can use the information to detect potential health issues early and provide personalized care. Such integration between wearable devices and healthcare professionals ensures a proactive approach to health management, bridging the gap between individuals and clinical support systems. By enhancing individual health awareness and supporting broader public health efforts, PPG technology plays an important role in advancing remote healthcare. As wearable technology and PPG continue to evolve, the potential to predict and manage health conditions remotely is expanding, making personalized, proactive care more achievable than ever [10].

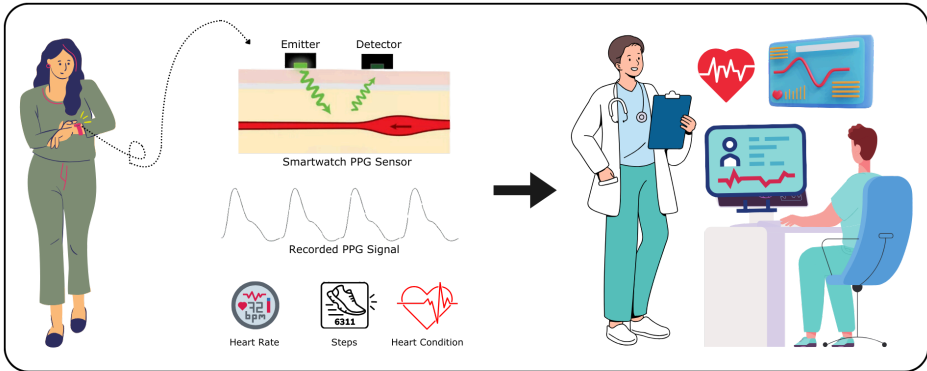


Figure 1.1: *Illustration of wearable health technology enabling continuous monitoring.* A smartwatch uses photoplethysmography (PPG) to measure blood volume changes, generating real-time physiological data that is processed and analyzed by healthcare providers for proactive, personalized care.

1.1. TIME SERIES DATA

As PPG data is recorded continuously over time, it forms what is known as a time series data—a sequence of measurements collected at successive intervals that are dependent on time, providing valuable insights when analyzed over time. Time series data is widely used across domains to capture changes in variables over time [11]. Common examples include daily temperature readings, stock market prices, and website traffic metrics [12–14].

One of the defining characteristics of time series data is its temporal nature—each observation is recorded in a specific order, and past observations often influence future ones [15]. This temporal dependency allows analysts to uncover patterns that evolve over time [16]. For example, weather forecasts rely on past and current temperature and pressure readings to predict future conditions, as these values are interdependent [17]. Similarly, in PPG data, the progression of heart rate during exercise and its subsequent recovery can reveal insights into cardiovascular fitness, as these measurements are influenced by preceding levels of exertion and recovery [18].

Building on this temporal nature, time series data often exhibits two key components: trends and seasonality. A *trend* reflects long-term changes in the data. For example, an upward trend in global average temperatures over decades indicates climate change, whereas a decline in resting heart rate in PPG data over months might suggest improved fitness [12, 19]. Figure 1.2b shows an example of a time series with a positive, increasing trend. *Seasonality* refers to recurring patterns that repeat at consistent intervals, such as daily peaks in electricity usage during the evening or hourly cycles in PPG data reflecting activity and rest [20, 21]. An example of seasonality in a time series, where a recurring pattern is apparent, is shown in Figure 1.2c.

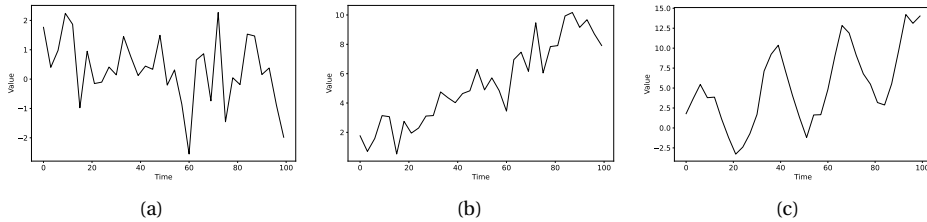


Figure 1.2: *Examples of time series data:* (a) a baseline signal without trend or seasonality, (b) the same baseline signal with a positive, increasing trend, and (c) a time series with a repetitive seasonal pattern combined with a slightly positive trend.

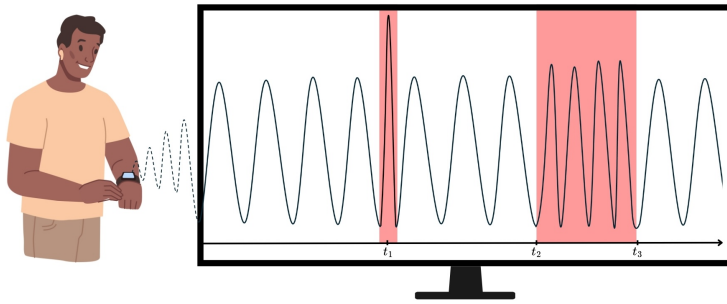
1.2. ANOMALIES IN TIME SERIES DATA

The specific characteristics of time series data enable the establishment of a baseline of expected behavior over time. This baseline represents the "normal" patterns observed within the system, providing a reference against which deviations can be identified [22]. By analyzing these typical patterns, we can determine what constitutes expected system behavior under varying conditions [23]. However, real-world systems are subject to changes and unexpected events that disrupt these patterns. Identifying these deviations from the established baseline of normal patterns is vital because they may indicate critical shifts in the system's state [22]. Such deviations, commonly referred to as anomalies, provide early warning signals that enable timely interventions and informed decision-making [24]. For example, Figure 1.3 illustrates two scenarios involving PPG signal monitoring: Figure 1.3a shows a person who is monitoring his heart rate while standing, and Figure 1.3b shows a person who is monitoring his heart rate while running. In both scenarios, the red-highlighted sections indicate deviations from the baseline pattern, representing anomalies. These deviations demonstrate how anomalies can manifest in different forms within time series data. Anomalies in time series data generally fall into three types [22]:

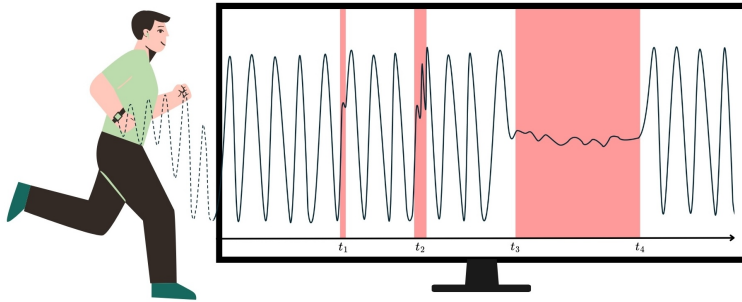
- *Point Anomalies:* These are single data points that stand out from the rest of the data. Such anomalies may indicate momentary sensor errors, environmental interference, or critical events such as an acute cardiac irregularity. For example,

in Figure 1.3a, where the person is standing, the PPG signal exhibits a sharp point anomaly at t_1 , highlighted by a pronounced deviation from the baseline pattern. This type of anomaly is typically easier to detect due to its clear and significant nature.

- *Contextual Anomalies:* These occur when a data point is anomalous in a particular context. For instance, an elevated heart rate during rest or sleep is unusual and could signal a health concern, even though such a heart rate might be normal during physical activity. In Figure 1.3a, where the person is standing, the PPG signal exhibits a contextual anomaly from t_2 to t_3 . While the pattern in this range resembles the typical PPG signal observed during running, shown in Figure 1.3b, it is considered anomalous in the context of the person being standing. The discrepancy arises because the observed PPG pattern does not



(a) PPG signal recorded while the person is standing: The signal shows a baseline pattern with a sharp "point" anomaly at t_1 and a "contextual" anomaly between t_2 and t_3 , reflecting a pattern inconsistent with the person's stationary state.



(b) PPG signal recorded while the person is running: The signal shows a "collective" anomaly with a subtle deviation at t_1 , a more significant anomaly at t_2 , and a sustained anomalous segment from t_3 to t_4 .

Figure 1.3: Examples of PPG signals collected during two scenarios: (a) while the person is standing and (b) while the person is running. The red-highlighted regions in both signals indicate various types of anomalies, including "point" anomalies, "contextual" anomalies, and "collective" anomalies, illustrating how deviations from the baseline pattern can manifest under different contexts and conditions.

align with the expected baseline for a standing individual. This highlights the importance of considering contextual information when identifying anomalies, as patterns must be evaluated relative to the conditions under which they occur.

- *Collective Anomalies*: These involve a sequence of data points that together represent an unusual pattern, even if individual points within the sequence may not appear anomalous on their own. Unlike point anomalies, where single data points can be identified as anomalous independently, collective anomalies require examining the entire sequence to recognize the deviation from the baseline. In Figure 1.3b, where the person is running, two types of collective anomalies can be identified. A short-duration collective anomaly is observed at t_1 , where a subtle deviation spans a few consecutive points, and at t_2 , where the deviation is more prominent but still occurs over a brief temporal window. A prolonged collective anomaly is observed from t_3 to t_4 . During this interval, the PPG signal deviates from the established baseline, reflecting a significant alteration in the overall pattern. This deviation may indicate a prolonged irregularity in data collection or another physiological abnormality that warrants attention. Unlike short-duration collective anomalies like those at t_1 and t_2 , prolonged anomalies are characterized by their persistence over a longer time frame.

1.3. ANOMALY DETECTION IN TIME SERIES DATA

Anomaly detection in time series data involves identifying these unusual events amidst the normal fluctuations and patterns. A variety of approaches can be employed, broadly categorized into *supervised* and *unsupervised* methods.

1.3.1. SUPERVISED ANOMALY DETECTION

In supervised anomaly detection, the model is trained on a labeled dataset, where each data point is explicitly marked as either "normal" or "anomalous." This approach treats anomaly detection as a classification problem, enabling the model to learn patterns associated with each class directly from the data [22]. For instance, Figure 1.3a illustrates a scenario where an expert has annotated a PPG signal recorded while the person is standing. In this figure, the regions around t_1 (sharp point anomaly) and between t_2 and t_3 (contextual anomaly) are labeled as anomalies. These labeled segments serve as ground truth for the model during training, allowing it to learn to distinguish anomalous patterns from normal behavior.

While supervised models can achieve high accuracy for detecting known anomalies, their effectiveness is heavily reliant on the availability and quality of labeled datasets. This dependency presents a significant challenge, as collecting labeled examples, particularly for rare anomalies, is often labor-intensive and costly [25]. Additionally, the scarcity of anomalies results in highly imbalanced datasets, making it difficult for models to learn effectively without biasing toward the majority class [26]. Supervised models are also prone to overfitting, especially when the training data fails to capture the full variability of both normal and anomalous patterns [27]. Moreover, these

models typically struggle to adapt to unknown or unexpected anomalies, failing to detect events that were not represented in the training data [28]. As a result, while supervised anomaly detection is powerful for well-defined problems with ample labeled data, its limitations underscore the need for alternative solution such as unsupervised anomaly detection.

1.3.2. UNSUPERVISED ANOMALY DETECTION

Unsupervised anomaly detection techniques do not rely on labeled datasets. Instead, they operate under the assumption that anomalies are rare and different from the majority of the data [22]. These methods focus on learning the inherent structure of normal behavior from the dataset. Once the model has established a baseline for what constitutes "normal," it identifies deviations from this baseline as anomalies. This flexibility makes unsupervised methods particularly suited for scenarios where labeled data is unavailable or where anomalies are not clearly defined.

However, while unsupervised methods alleviate the dependency on labeled data, they come with their own challenges. The effectiveness of these techniques heavily depends on the model's ability to learn a meaningful representation of the data, where normal and anomalous patterns are sufficiently distinct [29]. In this context, a meaningful representation refers to transforming the raw data into a reduced or structured form that highlights the essential patterns while discarding irrelevant variations or noise. If the model fails to capture the underlying structure of normal behavior accurately, it may misinterpret normal variations as anomalies, leading to false positives, or overlook subtle anomalies, resulting in false negatives [30].

1.3.3. CHALLENGES IN TIME SERIES ANOMALY DETECTION

Beyond the specific challenges mentioned for both supervised and unsupervised anomaly detection methods, both approaches are impacted by broader issues arising from the intrinsic properties of time series data.

HIGH DIMENSIONALITY

One primary challenge is the high dimensionality of time series data, particularly in multivariate datasets where multiple variables, or features, are recorded simultaneously at each time step. In this context, *features* refer to distinct measurements or variables, such as heart rate, blood oxygen levels, or motion sensor data in physiological monitoring.

In high-dimensional datasets, the relationships between features become increasingly complex. These interrelationships often play a critical role in detecting anomalies, as deviations from expected patterns may not occur within individual features but rather in their combined behavior [31]. For example, an anomaly might involve a subtle shift in the correlation between heart rate and oxygen saturation that would not be evident when examining either feature in isolation. Supervised methods, which rely on labeled datasets, face significant challenges in high-dimensional spaces because the model must learn these intricate relationships [22]. The scarcity of labeled data exacerbates this issue, making it difficult to adequately capture all

possible feature interdependencies. Without sufficient labeled examples, the model risks underperforming or failing to generalize. High dimensionality also introduces the risk of overfitting in supervised methods, where the model inadvertently learns noise or irrelevant correlations rather than meaningful patterns. This can lead to reduced accuracy when applied to unseen data. On the other hand, in unsupervised anomaly detection methods, which aim to establish a baseline of normal behavior, high dimensionality can obscure subtle anomalies or generate false alarms [29]. The complexity increases as the number of features grows, making it challenging for models to identify meaningful patterns. Additionally, the large number of possible interactions between features in high-dimensional spaces presents significant computational and modeling challenges for both supervised and unsupervised approaches [32].

Traditional approaches to address high dimensionality include dimensionality reduction techniques such as Principal Component Analysis (PCA), which project high-dimensional data onto a lower-dimensional space while preserving the most critical information [33]. More temporarily, Autoencoders, a type of neural network, are used to learn compact representations of data by compressing and reconstructing input features [34]. While these methods are effective at reducing the complexity of high-dimensional data, they have limitations. PCA assumes linear relationships between features, which may not hold in complex datasets [35], and autoencoders require careful tuning to avoid losing critical anomaly-related information [36]. Additionally, these techniques often struggle to preserve temporal dependencies in time series data, potentially diminishing their effectiveness for detecting anomalies [37].

TEMPORAL DEPENDENCIES

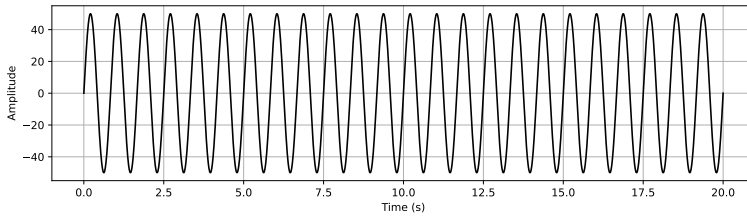
In addition to high dimensionality, temporal dependencies in time series data also affects anomaly detection [38]. These dependencies arise because observations, data points collected over time for each feature, in a time series are not independent but are sequentially related [39]. Anomalies, therefore, may not always occur as isolated point anomalies; they can also manifest as patterns or sequences that deviate from expected behavior over time (e.g., collective or contextual anomalies) [22]. For example, in Figure 1.3a, the anomaly at t_1 is an isolated point anomaly, while the anomaly section from t_2 to t_3 consists of a sequence of observations that are considered anomalous within the given context. Similarly, an irregular heart rhythm may develop gradually, with small changes in heart rate variability accumulating over time.

For supervised methods, temporal dependencies significantly complicate the task of labeling data. Gradual deviations or context-sensitive patterns require labeled datasets that not only capture isolated anomalies but also sequences of abnormal behavior [40]. However, constructing such datasets is often infeasible, especially when anomalies are rare or evolve unpredictably [22]. Consequently, supervised models trained predominantly on isolated point anomalies may fail to detect more complex temporal deviations embedded in sequences. Unsupervised methods also face considerable challenges when dealing with temporal dependencies [38]. Normal patterns in time

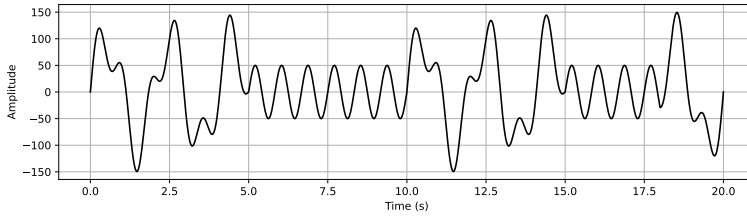
series often involve intricate, time-dependent relationships that vary across different time scales [39]. For example, PPG signals exhibit short-term fluctuations driven by breathing and long-term trends influenced by circadian rhythms. Distinguishing between these expected variations and genuine anomalies requires models capable of capturing both local and global temporal dependencies [29]. Failure to do so may result in false positives, where normal but unusual variations are flagged as anomalies, or missed detections for genuine sequential anomalies.

Moreover, temporal dependencies also complicate the detection of subtle anomalies—those that slightly deviate from normal temporal behavior but may still indicate significant underlying events—compared to more prominent anomalies. For example, in Figure 1.3b, the anomaly at t_1 represents a subtle deviation from the expected pattern. This type of anomaly may be harder to detect due to its minor deviation, which can easily blend into the inherent variability or noise present in time series data. By contrast, the anomaly at t_2 is a significant deviation that stands out from the baseline pattern, making it easier to detect. Supervised methods, which depend on labeled datasets, may fail to learn subtle anomalies like t_1 if they are underrepresented or absent in the training data. Additionally, supervised models tend to prioritize more distinct anomalies, such as t_2 , because these are easier to label and classify. This focus on significant anomalies can result in subtle deviations being disregarded, even when they are important. Unsupervised methods must accurately capture the intricate temporal dependencies in the data to distinguish between normal variability and subtle anomalies. Subtle deviations, like t_1 , may be misinterpreted as noise or expected temporal variations, particularly in the presence of complex patterns or noisy data. Furthermore, when significant anomalies like t_2 dominate the dataset, the model may be biased toward detecting these more obvious patterns, further reducing sensitivity to more subtle deviations.

To address temporal dependencies in time series anomaly detection, a variety of solutions have been developed. Traditional methods, such as autoregressive models (e.g., ARIMA and SARIMA), capture temporal relationships by modeling the current value of a time series as a function of its past values [41]. While effective for simpler, linear patterns or seasonal trends, these methods often struggle to handle non-linear dependencies or multi-scale temporal relationships. Another approach, the matrix profile, identifies anomalous subsequences by measuring similarities across the time series, providing a computationally efficient way to handle univariate series, though its applicability to multivariate or more complex anomalies is limited [42]. Neural networks, including LSTMs and Transformers, have become popular for their ability to model intricate, non-linear, and multi-scale temporal dependencies [39]. These models can capture both short-term fluctuations and long-term trends, making them powerful for detecting a wide range of temporal anomalies [43]. However, neural networks are prone to over-generalization, particularly when the training data predominantly reflects normal behavior. Subtle anomalies, which slightly deviate from normal patterns, may be overlooked if the model generalizes these deviations as part of the baseline [44, 45].



(a) Clean infrared PPG signal: The signal exhibits a smooth and consistent baseline pattern, representing ideal data collected without external noise or motion artifacts.



(b) Corrupted PPG signal with motion artifacts: The signal demonstrates significant distortion caused by motion artifacts, obscuring the original baseline pattern and making anomaly detection challenging.

Figure 1.4: Examples of PPG signals: (a) a clean signal without artifacts, and (b) a corrupted signal with motion artifacts. These examples illustrate the impact of noise and motion artifacts on data quality, highlighting the challenges in detecting anomalies in real-world scenarios.

NOISE AND DATA QUALITY

Another major challenge is the presence of noise and issues with data quality. In real-world applications, data may be collected from sensors or other devices in uncontrolled environments, where external factors like sensor placement, environmental conditions, or human activity can introduce noise [46]. Figure 1.4 provides an example of this issue in the context of physiological monitoring. Subfigure 1.4a illustrates a clean PPG signal, while Subfigure 1.4b shows the same signal corrupted by motion artifacts due to physical activity. These motion artifacts distort the underlying patterns, making it difficult to distinguish meaningful deviations from potential anomalies. For instance, anomalies embedded in the corrupted signal, such as subtle physiological changes, may be masked by the noise or misinterpreted as normal variations.

For supervised methods, noise and poor data quality can significantly degrade model performance by corrupting the labeled training data [47]. In the context of Figure 1.4, a corrupted PPG signal like the one in subfigure 1.4b could introduce inconsistencies in the labeled dataset, where noise may be mistaken for genuine anomalies or vice versa. When noisy labels or inputs are present, models may struggle to learn accurate distinctions between normal and anomalous patterns. In such cases, the model risks overfitting to noise, learning irrelevant patterns instead of generalizable features, and ultimately performing poorly on unseen data [47, 48]. Similarly, unsupervised methods struggle when noise obscures the true baseline of

normal behavior, as shown in subfigure 1.4b, where the motion artifacts make it challenging for the model to learn the inherent structure of the signal [49]. This issue is particularly pronounced in time series data, where noise can easily blend into complex temporal patterns, making it harder to distinguish true anomalies from random fluctuations.

Additionally, noise and data quality issues can compound other challenges in time series anomaly detection, such as high dimensionality and temporal dependencies. For example, when noise affects multiple features simultaneously or persists over extended periods, it can amplify the difficulty of detecting genuine anomalies across time and features. Models must disentangle true temporal patterns and feature correlations from spurious noise to effectively identify anomalies in such scenarios.

Traditional methods such as signal smoothing techniques, including moving averages and low-pass filters, aim to reduce noise by suppressing high-frequency variations that are unlikely to represent true anomalies [50, 51]. While these methods are computationally efficient and straightforward to implement, they risk removing subtle anomalies along with the noise, especially when anomalies themselves involve small, high-frequency deviations [52]. Noise-robust statistical models, such as Kalman filters, provide an alternative by estimating the underlying signal based on noisy observations, but their effectiveness diminishes when noise characteristics are complex or non-stationary [53]. Recent advances in machine learning have introduced more sophisticated solutions for handling noise. Neural networks, particularly autoencoders, are often used for denoising by learning compact representations of clean signals and reconstructing them from noisy inputs [36]. Variants like denoising autoencoders or convolutional neural networks (CNNs) are specifically designed to filter out noise while preserving meaningful patterns in the data [54]. However, these methods rely heavily on having representative training data that includes both clean and noisy signals, which may not always be available [55]. Moreover, neural networks can be prone to overfitting noise if the training process is not carefully controlled.

DOMAIN-SPECIFIC VARIABILITY

A further challenge in time series anomaly detection is domain-specific variability, which refers to differences that arise due to individual characteristics, device-specific factors, or variations in data collection sources. This variability complicates the development of robust models, as what constitutes "normal" behavior can vary significantly across domains, individuals, or environments [56]. For example, in physiological monitoring using wearables, PPG signals may differ greatly between individuals due to factors such as age, skin tone, body composition, or fitness levels [10]. This phenomenon, typically referred to as *inter-subject variability*, highlights how a pattern that appears normal for one individual may be considered anomalous for another, leading to difficulties in designing models that generalize across individuals [57].

In supervised methods, these variations can cause models to overfit to the specific characteristics of the training data, limiting their ability to generalize to new domains or contexts. Without diverse labeled datasets that capture the full range of variability, supervised models may misclassify domain-specific normal patterns as anomalies or

overlook genuine anomalies that deviate from these patterns. Unsupervised methods face similar challenges. Since these approaches rely on learning a baseline of normal behavior directly from the data, domain-specific variability in the training data can distort this baseline. For instance, an unsupervised model trained on PPG data from a single individual may struggle to accurately detect anomalies in data collected from another individual with significantly different physiological characteristics. This challenge is further exacerbated in multivariate time series, where models must simultaneously account for temporal dependencies and feature correlations while accommodating domain-specific differences across multiple variables. In such cases, variability across features or individuals can obscure true anomalies or lead to false positives, further complicating anomaly detection in diverse domains.

To address domain-specific variability in time series anomaly detection, traditional approaches often rely on feature standardization or normalization techniques to reduce variability by rescaling data to a common scale. While effective for minimizing inter-individual differences in straightforward cases, these techniques may not capture more complex variations arising from domain-specific factors [58]. Similarly, ensemble models that aggregate predictions from multiple models trained on different subsets of the data can enhance robustness, but they often require substantial computational resources and may still struggle with underrepresented variability in the training data [59, 60]. Transfer learning and domain adaptation techniques might handle domain-specific variability more effectively [61]. For instance, pretraining models on large, diverse datasets before fine-tuning them on smaller, domain-specific datasets can help mitigate overfitting to narrow contexts [62]. Transfer learning enables models to retain generalizable patterns learned during pretraining while adapting to specific domains during fine-tuning [63]. However, this approach relies on the availability of well-labeled source datasets, which may not always exist for niche applications [64]. Another promising method involves personalized modeling frameworks that adapt anomaly detection models to individual users or specific domains. However, personalized models face challenges such as data scarcity and the computational demands of managing multiple models [65–67].

EVALUATION OF TIME SERIES ANOMALY DETECTION METHODS

Once an anomaly detection model is built, evaluating its performance in time series data presents unique challenges, especially as we deal with rare, context-dependent events that unfold over intervals. Conventional evaluation methods often rely on metrics like Precision, Recall, or F1-score, which treat anomalies as isolated events and evaluate each data point independently [68]. While these metrics are effective for static data, they fall short in the context of time series, where anomalies frequently form part of larger temporal patterns and exhibit sequential dependencies. For example, an anomaly may span multiple time steps, making it difficult to assess performance using metrics designed for static, point-based evaluations.

These limitations underscore the need for evaluation criteria that address the specific complexities of time series data. A well-rounded assessment should consider several critical aspects of model performance. Among these, one criterion is *Early detection* which refers to identifying potential anomalies before they fully manifest,

often based on subtle changes in the data pattern over time. This capability is especially valuable for proactive interventions, as it enables timely responses to emerging issues. The other criterion is *Delayed detection* which occurs when an anomaly is identified after it has fully occurred. While not ideal, delayed detection reflects the model's capacity to eventually recognize anomalies, even if the response is not immediate. Figure 1.5 illustrates examples of both early and delayed detection. For instance, the anomaly event a_1 is detected early by prediction p_{11} , highlighting the value of early detection for timely intervention. Conversely, the same anomaly a_1 is detected with a delay by prediction p_{12} , demonstrating the importance of accounting for such cases in evaluation metrics.

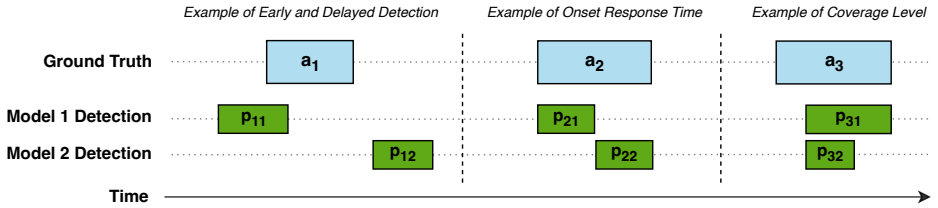


Figure 1.5: *Illustration of anomaly detection in time series data.* a_{1-3} represent the actual anomalies as ground truth. Predictions are denoted by p . The durations of both events are indicated by the length of the boxes. Overlapping areas indicated between p and a demonstrate where the model has correctly identified anomalies.

A concept closely related to *Early* and *Delayed detection* is *Onset response time*, which evaluates how accurately the detection of an anomaly aligns with the start of the event. *Onset response time* focuses on identifying the beginning of an anomaly or event as it starts to develop. In Figure 1.5, anomaly event a_2 is detected by p_{21} and p_{22} . However, p_{21} aligns more closely with the beginning of the anomaly event a_2 , indicating a faster response than p_{22} . Understanding and evaluating both early and *Onset response time* are critical for applications where timely responses can significantly mitigate potential risks or consequences.

Anomaly *Coverage level* is another essential criterion, assessing the model's ability to detect a wide range of anomalies present in the data. High coverage ensures the model is not overly selective, capturing not only clear outliers but also more complex anomalies, such as contextual or collective patterns that span multiple data points. In Figure 1.5, predictions p_{31} and p_{32} detect anomaly a_3 , but p_{31} offers greater coverage, aligning more extensively with the anomalous interval. Comprehensive coverage is particularly critical in applications where varied anomaly types carry significant implications.

Together, these criteria form a comprehensive framework for evaluating the performance of anomaly detection models in time series data. Various metrics have been proposed to evaluate time series anomaly detection, addressing its sequential nature and temporal dependencies. Metrics such as *R-based* [69], *TS-Aware* [70], and their enhanced versions like *ETS-Aware* [71] have advanced the field by incorporating

considerations for range coverage, delayed detection, and overlap scoring. Point-based metrics like *PA-FI* [72] simplify evaluation but often lead to optimistic scores, particularly when dealing with fragmented or subtle anomalies. Threshold-free metrics such as *VUS-ROC* and *VUS-PR* [73] offer an alternative by evaluating performance across a range of thresholds, yet they fall short in accounting for early and onset detection, critical in time-sensitive applications. While these metrics represent significant progress, gaps remain in fully capturing key evaluation criteria of time series anomalies.

1.4. FOUNDATIONS OF THE THESIS AND OUTLINE

The challenges outlined above underscore the complexity of time series anomaly detection, driven by issues such as label scarcity, high dimensionality, noise, intricate temporal dependencies, and evaluation metrics. Addressing these gaps is essential for developing more effective and reliable anomaly detection systems. A key focus of this thesis is addressing the limitations of existing evaluation metrics. By introducing a framework tailored to the unique characteristics of time series data, this work provides a more accurate and context-aware assessment of anomaly detection methods. Beyond evaluation, this thesis is guided by several core principles for addressing anomaly detection challenges in time series data. Central to this work is the use of unsupervised methods, which enable effective learning from unlabeled data, and representation learning, which provides a robust foundation for capturing meaningful patterns in complex datasets. Additionally, the thesis introduces novel model design and anomaly scoring mechanisms to improve the detection of subtle anomalies, emphasizing sensitivity and robustness in complex real-world datasets. Non-linear techniques are employed to capture the rich temporal structure of time series data, balancing generalization and sensitivity to ensure both prominent and subtle anomalies are effectively detected.

Building on these principles, this thesis makes several contributions to advance the state of time series anomaly detection, as detailed in the following chapters:

Chapter 2 focuses on the critical challenge of effectively evaluating anomaly detection models in time series data, which is often overlooked by traditional metrics. To overcome this, we introduce PATE (Proximity-Aware Time Series Anomaly Evaluation), a novel metric designed to account for the temporal complexity inherent in anomaly detection tasks. PATE considers the proximity of detected anomalies to the true anomaly onset, allowing for a more accurate evaluation. It captures aspects like *Early detection*, *Delayed detection*, *Onset response time*, and *Coverage level*, thereby addressing the shortcomings of existing evaluation metrics. This chapter not only explains the development of PATE but also demonstrates its effectiveness through experiments on both synthetic and real-world datasets, highlighting its superiority in capturing the intricacies of time series anomaly detection.

Chapter 3 addresses the challenges of label scarcity and high inter-subject variability in PPG data by adopting an approach that enables the model to learn

from unlabeled data. The method involves training the model on an auxiliary task, such as reconstructing the PPG signal, to capture essential patterns in the data. This auxiliary task, known as a *pretext task*, helps the model learn lower-dimensional, informative representations that filter out noise while retaining critical information. These learned representations are then fine-tuned for a specific application, such as human activity recognition, referred to as the *downstream task*. However, using reconstruction as the pretext task did not fully address the differences between individuals, and inter-subject variability remains a significant challenge, limiting the model's generalization capabilities. This chapter serves as a foundation for understanding the potential and limitations of representation learning in reducing data complexity and improving anomaly detection in subsequent chapters.

Chapter 4 builds on the limitations identified in Chapter 3 regarding the reconstruction pretext task and focuses on improving anomaly detection in PPG signals as downstream tasks. In this chapter, we propose a custom pretext task, where instead of reconstructing the original signal, we classify different transformations of the signal. This approach helps learn lower-dimensional representations that are more robust to variability and noise, effectively improving generalization across subjects for two different anomaly detection tasks. We demonstrate that this custom pretext task better captures essential features, tackling the challenges associated with representation learning in PPG signals. Additionally, we explore personalization, tailoring models to individual users to further reduce inter-subject variability, ultimately enhancing the reliability of the model.

Chapter 5 shifts focus to the challenge of detecting subtle anomalies in time series data, which are often overlooked due to the overgeneralization tendencies of traditional reconstruction-based methods. To address this, we introduce RESTAD (REconstruction and Similarity-based Transformer for Anomaly Detection), a novel framework that combines reconstruction errors with similarity-based scoring mechanisms. By integrating a Radial Basis Function (RBF) layer within a Transformer architecture, RESTAD improves detection sensitivity to subtle deviations by measuring the proximity of data points to learned reference centers in the latent space. This chapter demonstrates RESTAD's ability to detect both prominent and subtle anomalies across a variety of benchmark datasets, offering a practical solution to the challenges of missed anomalies and enhancing the field of anomaly scoring mechanisms.

REFERENCES

- [1] K. H. Bowles, P. Dykes, and G. Demiris. “The use of health information technology to improve care and outcomes for older adults”. In: *Research in gerontological nursing* 8.1 (2015), pp. 5–10.
- [2] J. P. Higgins. “Smartphone applications for patients’ health and fitness”. In: *The American journal of medicine* 129.1 (2016), pp. 11–19.
- [3] D. Rodbard. “Continuous glucose monitoring: a review of successes, challenges, and opportunities”. In: *Diabetes technology & therapeutics* 18.S2 (2016), S2–3.
- [4] M. Weenk, H. van Goor, B. Frietman, L. J. Engelen, C. J. van Laarhoven, J. Smit, S. J. Bredie, T. H. van de Belt, *et al.* “Continuous monitoring of vital signs using wearable devices on the general ward: pilot study”. In: *JMIR mHealth and uHealth* 5.7 (2017), e7208.
- [5] N. Aagaard, M. H. Olsen, O. W. Rasmussen, K. K. Grønbaek, J. Mølgaard, C. Haahr-Raunkjaer, M. Elvekjaer, E. K. Aasvang, and C. S. Meyhoff. “Prognostic value of heart rate variability for risk of serious adverse events in continuously monitored hospital patients”. In: *Journal of Clinical Monitoring and Computing* (2024), pp. 1–15.
- [6] Z. Lewczak, M. Mitchell, and M. G. Mitchell. “Wearable Technology and Chronic Illness: Balancing Justice and Care Ethics”. In: *Cureus* 16.11 (2024).
- [7] B. Noah, M. S. Keller, S. Mosadeghi, L. Stein, S. Johl, S. Delshad, V. C. Tashjian, D. Lew, J. T. Kwan, A. Jusufagic, *et al.* “Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials”. In: *NPJ digital medicine* 1.1 (2018), p. 20172.
- [8] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran. “A review on wearable photoplethysmography sensors and their potential future applications in health care”. In: *International journal of biosensors & bioelectronics* 4.4 (2018), p. 195.
- [9] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey. “Wearable photoplethysmography for cardiovascular monitoring”. In: *Proceedings of the IEEE* 110.3 (2022), pp. 355–381.
- [10] J. Allen. “Photoplethysmography and its application in clinical physiological measurement”. In: *Physiological measurement* 28.3 (2007), R1.
- [11] C. Chatfield and H. Xing. *The analysis of time series: an introduction with R*. Chapman and hall/CRC, 2019.
- [12] P. D. Jones and A. Moberg. “Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001”. In: *Journal of climate* 16.2 (2003), pp. 206–223.

- [13] E. F. Fama. “Efficient capital markets”. In: *Journal of finance* 25.2 (1970), pp. 383–417.
- [14] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu. “Machine learning-powered encrypted network traffic analysis: A comprehensive survey”. In: *IEEE Communications Surveys & Tutorials* 25.1 (2022), pp. 791–824.
- [15] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [16] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer. *Time series analysis and its applications*. Vol. 3. Springer, 2000.
- [17] D. S. Wilks. *Statistical methods in the atmospheric sciences*. Academic press, 2011.
- [18] A. Schäfer and J. Vagedes. “How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram”. In: *International journal of cardiology* 166.1 (2013), pp. 15–29.
- [19] D. Nunan, G. R. Sandercock, and D. A. Brodie. “A quantitative systematic review of normal values for short-term heart rate variability in healthy adults”. In: *Pacing and clinical electrophysiology* 33.11 (2010), pp. 1407–1417.
- [20] H. S. Hippert, C. E. Pedreira, and R. C. Souza. “Neural networks for short-term load forecasting: A review and evaluation”. In: *IEEE Transactions on power systems* 16.1 (2001), pp. 44–55.
- [21] P. H. Charlton, P. Celka, B. Farukh, P. Chowienczyk, and J. Alastruey. “Assessing mental stress from the photoplethysmogram: a numerical study”. In: *Physiological measurement* 39.5 (2018), p. 054001.
- [22] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [23] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano. “A review on outlier/anomaly detection in time series data”. In: *ACM computing surveys (CSUR)* 54.3 (2021), pp. 1–33.
- [24] V. Hodge and J. Austin. “A survey of outlier detection methodologies”. In: *Artificial intelligence review* 22 (2004), pp. 85–126.
- [25] B. Krawczyk. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in artificial intelligence* 5.4 (2016), pp. 221–232.
- [26] H. He and E. A. Garcia. “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.
- [27] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld. “Toward supervised anomaly detection”. In: *Journal of Artificial Intelligence Research* 46 (2013), pp. 235–262.
- [28] H. Xu, C. Caramanis, and S. Mannor. “Robustness and Regularization of Support Vector Machines.” In: *Journal of machine learning research* 10.7 (2009).

- [29] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. “Deep autoencoding gaussian mixture model for unsupervised anomaly detection”. In: *International conference on learning representations*. 2018.
- [30] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, *et al.* “Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 187–196.
- [31] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. “Outlier detection for temporal data: A survey”. In: *IEEE Transactions on Knowledge and data Engineering* 26.9 (2013), pp. 2250–2267.
- [32] M. Verleysen and D. François. “The curse of dimensionality in data mining and time series prediction”. In: *International work-conference on artificial neural networks*. Springer. 2005, pp. 758–770.
- [33] I. Jolliffe. “Principal component analysis”. In: *Encyclopedia of statistics in behavioral science* (2005).
- [34] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.
- [35] M. E. Tipping and C. M. Bishop. “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (1999), pp. 611–622.
- [36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. “Extracting and composing robust features with denoising autoencoders”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.
- [37] Z. Wang and T. Oates. “Encoding time series as images for visual inspection and classification using tiled convolutional neural networks”. In: *Workshops at the twenty-ninth AAAI conference on artificial intelligence*. 2015.
- [38] J. Gama, R. Sebastiao, and P. P. Rodrigues. “On evaluating stream learning algorithms”. In: *Machine learning* 90 (2013), pp. 317–346.
- [39] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, *et al.* “Long short term memory networks for anomaly detection in time series.” In: *Esann*. Vol. 2015. 2015, p. 89.
- [40] L. Akoglu, H. Tong, and D. Koutra. “Graph based anomaly detection and description: a survey”. In: *Data mining and knowledge discovery* 29 (2015), pp. 626–688.
- [41] R. Hyndman. *Forecasting: principles and practice*. OTexts, 2018.
- [42] Y. Wang, L. Wu, Z. Wu, E. Chen, and Q. Liu. “Selecting valuable customers for merchants in e-commerce platforms”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE. 2016, pp. 1281–1286.
- [43] A. Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).

- [44] T. Zhao, L. Jin, X. Zhou, S. Li, S. Liu, and J. Zhu. “Unsupervised Anomaly Detection Approach Based on Adversarial Memory Autoencoders for Multivariate Time Series.” In: *Computers, Materials & Continua* 76.1 (2023).
- [45] H. Zhong, Y. Zhao, and C. G. Lim. “Abnormal State Detection using Memory-augmented Autoencoder technique in Frequency-Time Domain”. In: *KSII Transactions on Internet and Information Systems (TIIS)* 18.2 (2024), pp. 348–369.
- [46] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers. “A review of wearable sensors and systems with application in rehabilitation”. In: *Journal of neuroengineering and rehabilitation* 9 (2012), pp. 1–17.
- [47] B. Frénay and M. Verleysen. “Classification in the presence of label noise: a survey”. In: *IEEE transactions on neural networks and learning systems* 25.5 (2013), pp. 845–869.
- [48] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115.
- [49] P. Perera and V. M. Patel. “Learning deep features for one-class classification”. In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5450–5463.
- [50] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, J. E. Dong, and R. C. Goodlin. “Adaptive noise cancelling: Principles and applications”. In: *Proceedings of the IEEE* 63.12 (1975), pp. 1692–1716.
- [51] G. Casiez, N. Roussel, and D. Vogel. “1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 2527–2530.
- [52] H. Magsi, A. H. Sodhro, F. A. Chachar, and S. A. K. Abro. “Analysis of signal noise reduction by using filters”. In: *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE. 2018, pp. 1–6.
- [53] R. E. Kalman. “A new approach to linear filtering and prediction problems”. In: (1960).
- [54] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155.
- [55] A. A. Hendriksen, D. M. Pelt, and K. J. Batenburg. “Noise2inverse: Self-supervised deep convolutional denoising for tomography”. In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 1320–1335.
- [56] Z. Z. Darban, G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi. “CARLA: Self-supervised contrastive representation learning for time series anomaly detection”. In: *Pattern Recognition* 157 (2025), p. 110874.
- [57] R. Ghorbani, M. J. Reinders, and D. M. Tax. “Self-supervised ppg representation learning shows high inter-subject variability”. In: *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*. 2023, pp. 127–132.

- [58] M. Kuhn. *Applied predictive modeling*. 2013.
- [59] A. Iqbal, R. Amin, F. S. Alsubaei, and A. Alzahrani. “Anomaly detection in multivariate time series data using deep ensemble models”. In: *Plos one* 19.6 (2024), e0303890.
- [60] A. Iliopoulos, J. Violos, C. Diou, and I. Varlamis. “Detection of Anomalies in Multivariate Time Series Using Ensemble Techniques”. In: *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE. 2023, pp. 1–8.
- [61] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.
- [62] K. Weiss, T. M. Khoshgoftaar, and D. Wang. “A survey of transfer learning”. In: *Journal of Big data* 3 (2016), pp. 1–40.
- [63] S. J. Pan and Q. Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [64] M. Wang and W. Deng. “Deep visual domain adaptation: A survey”. In: *Neurocomputing* 312 (2018), pp. 135–153.
- [65] M. Goldstein and S. Uchida. “A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data”. In: *PloS one* 11.4 (2016), e0152173.
- [66] R. Ghorbani, M. J. Reinders, and D. M. Tax. “Personalized anomaly detection in PPG data using representation learning and biometric identification”. In: *Biomedical Signal Processing and Control* 94 (2024), p. 106216.
- [67] B. Veeravalli, C. J. Deepu, and D. Ngo. “Real-time, personalized anomaly detection in streaming data for wearable healthcare devices”. In: *Handbook of large-scale distributed computing in smart healthcare* (2017), pp. 403–426.
- [68] C. C. Aggarwal and C. C. Aggarwal. *An introduction to outlier analysis*. Springer, 2017.
- [69] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich. “Precision and recall for time series”. In: *Advances in neural information processing systems* 31 (2018).
- [70] W.-S. Hwang, J.-H. Yun, J. Kim, and H. C. Kim. “Time-series aware precision and recall for anomaly detection: considering variety of detection result and addressing ambiguous labeling”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 2241–2244.
- [71] W.-S. Hwang, J.-H. Yun, J. Kim, and B. G. Min. “Do you know existing accuracy metrics overrate time-series anomaly detections?” In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 2022, pp. 403–412.
- [72] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, *et al.* “Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 187–196.

- [73] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. “Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection”. In: *Proceedings of the VLDB Endowment* 15.11 (2022), pp. 2774–2787.

2

PATE: PROXIMITY-AWARE TIME SERIES ANOMALY EVALUATION METRIC

Evaluating anomaly detection algorithms in time series data is critical as inaccuracies can lead to flawed decision-making in various domains where real-time analytics and data-driven strategies are essential. Traditional performance metrics assume iid data and fail to capture the complex temporal dynamics and specific characteristics of time series anomalies, such as early and delayed detections. We introduce Proximity-Aware Time series anomaly Evaluation (PATE), a novel evaluation metric that incorporates the temporal relationship between prediction and anomaly intervals. PATE uses proximity-based weighting considering buffer zones around anomaly intervals, enabling a more detailed and informed assessment of a detection. Using these weights, PATE computes a weighted version of the area under the Precision and Recall curve. Our experiments with synthetic and real-world datasets show the superiority of PATE in providing more sensible and accurate evaluations than other evaluation metrics. We also tested several state-of-the-art anomaly detectors across various benchmark datasets using the PATE evaluation scheme. The results show that a common metric like Point-Adjusted F1 Score fails to characterize the detection performances well, and that PATE is able to provide a more fair model comparison. By introducing PATE, we redefine the understanding of model efficacy that steers future studies toward developing more effective and accurate detection models.

This chapter has been published as:

Ghorbani R, Reinders MJ, Tax DM. PATE: Proximity-Aware Time Series Anomaly Evaluation. *In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2024 Aug 25 (pp. 872-883)*. [1].

Code available at:

<https://github.com/Raminghorbanii/PATE>

2.1. INTRODUCTION

Anomaly detection in time series (TS) data, the process of identifying unusual patterns that deviate from the expected norm, has become increasingly important across various domains [2, 3]. The rapid advancement of data-driven decision-making and real-time analytics has opened opportunities for developing more accurate anomaly detection methods. Such developments often lead to models competing to claim the status of 'State-of-the-Art' (SOTA). Achieving this status is not just a matter of academic prestige; it often directs the focus of future research, influences industry adoption, and guides the development of practical applications. However, choosing an appropriate evaluation metric is critical to avoid incorrect conclusions about a model's performance. Relying on evaluation metrics that do not accurately reflect the true effectiveness of the models can lead to flawed decisions in real-world applications. This is particularly consequential in critical domains, such as medical diagnostics or financial fraud detection, where relying on a poorly evaluated model can have serious repercussions.

Standard evaluation metrics such as Precision and Recall [4] are effective for point-based anomaly detection as they assess the accuracy of detecting isolated iid events. In this context, each data point is evaluated independently, allowing for straightforward calculation of these metrics. However, in TS data, events and anomalies typically occur in time *intervals*. This complexity causes several situations: 1) *Early Detection*, when potential anomalies are identified before they fully manifest, based on subtle changes in the data pattern over time. Figure 2.1 shows an example of early detection where prediction p_{11} detects the anomaly event a_1 earlier than its actual occurrence. Although p_{11} does not align exactly with a_1 , such early detection is valuable for early response actions and should be appropriately appreciated in evaluation metrics. 2) *Delayed Detection*, occurs when an anomaly event is not detected immediately but is identified at a later time, even after its actual occurrence. In Figure 2.1, the anomaly event a_1 is detected with a delay by prediction event p_{12} . Although p_{12} does not align precisely with a_1 , this type of delayed detection should be accounted for in the evaluation process, as it reflects the model's ability to eventually identify anomalies, even after some delay.

Another situation, 3) *Onset Response Time*, refers to how close the detection of an anomaly is to the start of the event. Timely detection is valuable, especially in scenarios where immediate action is required. In Figure 2.1, anomaly event a_2 is

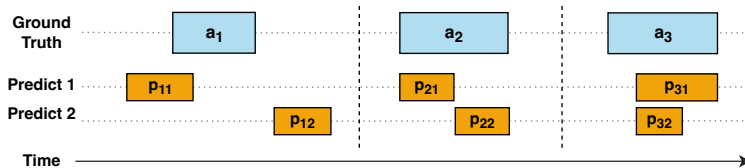


Figure 2.1: *Illustration of anomaly detection in time series data.* a_{1-3} represent the actual anomalies as ground truth. Predictions are denoted by p . The durations of both events are indicated by the length of the boxes. Overlapping areas between p and a demonstrate where the model has correctly identified anomalies.

detected by p_{21} and p_{22} . However, p_{21} aligns more closely with the beginning of the anomaly event a_2 , indicating a faster response than p_{22} . Evaluation metrics should reward those that occur promptly after the onset of an anomaly. Finally 4) *Coverage level of Predictions*, refers to the range that a prediction covers an actual anomaly. The effectiveness of a prediction can be measured by how much of the anomaly it successfully captures. In Figure 2.1, predictions p_{31} and p_{32} both detect anomaly event a_3 , but p_{31} covers a_3 more than p_{32} . This more extensive coverage by p_{31} makes it a more effective prediction for a_3 . Accordingly, evaluation metrics need to consider the coverage range of the predictions over the duration of the anomalies.

Various metrics have been developed that are specifically tailored to the sequential nature of time series data (referred to as *Sequential Adaptability*). For instance, Range-based Precision and Recall metrics, hereafter denoted as *R-based* [5], expand upon traditional metrics by incorporating factors such as existence (detecting the anomaly range with at least one point), size and position (reflecting the number and relative position of correctly detected anomaly ranges), and cardinality (penalizing fragmented predictions for a single anomaly). The Time Series Aware Precision and Recall, hereafter denoted as *TS-Aware* [6], follows a similar approach but omits cardinality and position considerations. This metric requires a prediction to cover a minimum percentage θ of an anomaly for it to be considered a true detection. They also add a buffer zone δ to give some credit for delayed detection in a decreasing manner. An enhanced version, denoted as *ETS-Aware* [7], further refines the evaluation by combining detection and overlap scores for improved accuracy in scoring overlapped detections. Further, the *Affiliation* metric [8], introduces a different perspective by focusing on the distance between prediction and actual anomaly ranges. It assesses the proximity of predicted anomalies to actual ones by measuring the duration between their respective ranges.

Another widely used method is the Point Adjusted F1 Score metric, which we will denote as *PA-F1* [9]. This approach assumes that detecting a single point in an anomaly range is sufficient for human experts to identify the entire range. Thus, it considers all observations within the corresponding anomaly range as correctly detected anomalies. However, it has been criticized for potentially generating optimistic scores. For example, [10] revealed that random anomaly scores from a uniform distribution outperform state-of-the-art methods when evaluated using this metric. To address this, [10] proposed a modified version that requires a portion of $K\%$ of the anomaly range to be detected before making any adjustments.

While all these metrics represent advancements in time series anomaly detection evaluation, they do not fully consider all the critical factors of early and delayed detections, or onset response timing. In addition to these limitations, the aforementioned metrics also require the setting of a threshold, a value where data points with anomaly scores exceeding this value are classified as anomalies. Selecting this threshold adds additional complexity and leads to subjectivity and inconsistency in evaluations. Metrics such as the Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the Area Under the Precision-Recall curve (AUC-PR) eliminate the need for thresholding by evaluating the performance of the model across a range of thresholds. However, they fall short in time series contexts due to not

considering the order of the data points and the temporal correlation between them. In response to this issue, Volume Under the Surface (VUS) metrics, *VUS-ROC* and *VUS-PR*, are proposed [11]. These metrics acknowledge the need to accommodate close predictions to the true anomaly ranges by adjusting the labels to be between 0 and 1 on a range over both sides of the actual anomaly range. Although the method is threshold-free, it does not pay attention to early and delayed detection, and onset response time. Furthermore, by changing the original labels, the metric gives unrealistic scores, as reaching the maximum detection score of 1 is not possible.

This paper introduces the Proximity-Aware Time series anomaly Evaluation metric, PATE (/peIt/). Our novel metric integrates buffer zones around the anomaly events and utilizes a special proximity-based weighting mechanism, enabling a detailed assessment of both early/delayed detections and addressing the onset response time challenge. PATE avoids the subjectivity of threshold-dependent metrics by integrating over the range of thresholds, offering a fair and unbiased evaluation, especially in research settings where expert knowledge might not be available for setting the exact desirable parameters based on the application. Table 2.1 illustrates a comparison between existing metrics and PATE, highlighting the comprehensive adaptability reconsideration of PATE in evaluating the TS anomaly detection.

Table 2.1: *Comparison of Anomaly Detection Evaluation Metrics*. Key features: Sequential Adaptability (SA); Early Detection (ED); Delayed Detection (DD); Onset Response Time (ORT); Coverage Level (CL) and Threshold-Free (TF)

Metric	SA	ED	DD	ORT	CL	TF
Precision/Recall (F1 Score)	-	-	-	-	-	-
R-based	✓	-	-	-	-	-
TS-Aware/ETS-Aware	✓	-	✓	-	-	-
Affiliation	✓	-	-	-	✓	-
PA-F1	✓	-	-	-	-	-
AUC-ROC/PR	-	-	-	-	-	✓
VUS-ROC/PR	✓	-	✓	-	-	✓
PATE	✓	✓	✓	✓	✓	✓

2.2. PROPOSED EVALUATION METRIC - PATE

A time series is denoted as a sequence of observations $\mathcal{X} = \{x_t\}_{t=1}^T$, where T represents the length of the time series, and each x_t is the observed data point at time t .

An actual anomaly event (labeled as positive in the ground truth labels) is a subsegment within the time series, denoted as $\mathbf{a}_k = (i_k, n_k)$ for points i_k and n_k with $1 \leq i_k \leq n_k \leq T$. The set of all anomaly events in the time series is represented as $\mathcal{A} = \{\mathbf{a}_k\}_{k=1}^N$, where N is the number of anomaly events present in the time series.

In practice, the detection models output continuous anomaly scores, denoted as $\mathcal{S} = \{s_t\}_{t=1}^T$, representing the likelihood of each observation x_t to be anomalous. These scores are then converted into binary predictions by applying a threshold θ , where scores equal to or exceeding the threshold are classified as anomalies. We define a prediction event as a subsegment identified by these binary predictions to be anomalous, denoted as $\mathbf{p}_l(\theta) = (m_l, j_l)$ for points m_l and j_l with $1 \leq m_l \leq j_l \leq T$. The

set of all prediction events is represented as $\mathcal{P} = \{p_l(\theta)\}_{l=1}^M$, where M is the number of prediction events identified by the model.

The effectiveness of the anomaly detector is determined by how well these $p_l(\theta)$ events align with the a_k events. PATE distinguishes several categories of matches between ground truth and predictions based on their temporal relationships and assigns proximity-specific weights to each point in each category. These weights are then used to compute a weighted version of Precision and Recall scores. The final measure of PATE is a weighted AUC-PR, which is derived from these weighted Precision and Recall scores. Further details on these computations are provided in the following sections.

2.2.1. CATEGORIZING THE EVENTS

Figure 2.2 illustrates the different categories of anomaly and prediction events in relation to each other. In assessing each $p_l(\theta)$, we consider its overlap, proximity, or distance (temporal relation) from each a_k . This approach allows for the clear differentiation of the diverse scenarios: complete and partial detection of anomalies, early or delayed detection, and instances where anomalies are either partially or entirely missed. Specifically, we categorize the anomaly and prediction events as follows:

PREDICTION EVENTS CATEGORIES:

- **True-Detection:** Sub-segments of the prediction event $p_l(\theta)$ that overlap with an anomaly event a_k , indicating anomalies that are accurately identified and not missed. Examples are segments p_1 , p_5 , and p_{6-2} in Figure 2.2.

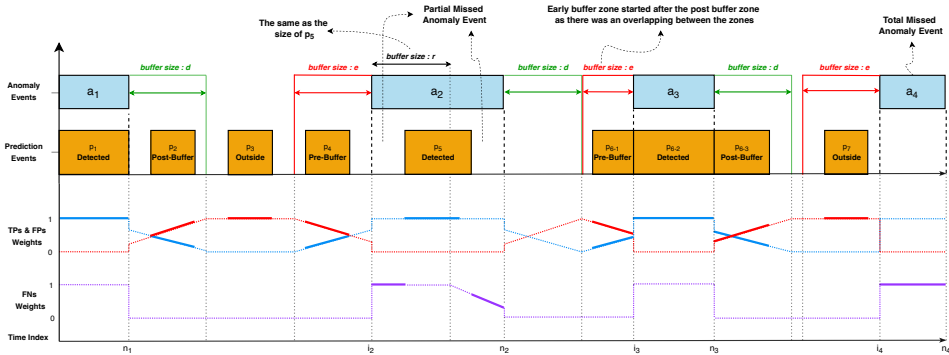


Figure 2.2: *Illustration of the Categorization and Weighting Mechanism in the PATE Method.* Prediction events ($p_1 - p_7$) are represented by orange boxes, while anomaly events ($a_1 - a_4$) are depicted by blue boxes. TP weights are illustrated with a blue line —, FP weights with a red line —, and FN weights with a purple line —. Note that the solid segments of the lines, in contrast to the dotted segments, indicate the activated weights for the example scenario depicted in the figure.

- **Post-Buffer Detection:** Sub-segments of the prediction event $p_l(\theta)$ that fall into a buffer zone immediately following an anomaly event a_k (See segments p_2 and p_{6-3} in Figure 2.2). This category highlights the capacity of the model for delayed detection. The post-buffer zone size, denoted by d , can be adjusted by experts based on specific application needs. When d is unknown for a specific application, we can consider a range of values for d rather than a fixed one $D = \{0, 1, \dots, d_{\max}\}$. This approach allows for a comprehensive assessment of the model's performance across different scenarios, as each buffer size can provide a different perspective on the performance of the model. Details on how these buffer sizes contribute to the overall PATE score will be discussed in the following sections.

- **Pre-Buffer Detection:** Sub-segments of the prediction event $p_l(\theta)$ that fall into a zone that precedes the start of an anomaly event a_k . This category highlights the capacity of the model for early detection, signaling potential anomalies ahead of time. Similar to the post-buffer zone, the size of the pre-buffer zone, denoted by e , varies within the set $E = \{0, 1, \dots, e_{\max}\}$ with the same approach for the assessment. The assignment of points to this category is conditional on not overlapping with the Post-Buffer zone of a preceding anomaly a_{k-1} , ensuring that the model early warning is distinct from a delayed detection of the previous event. In other words, the Post-Buffer category has priority, and therefore, if $i_k - e < n_{k-1} + d$ then the Pre-Buffer zone starts at $n_{k-1} + d + 1$ instead of $i_k - e$. Furthermore, Pre-Buffer detection is dependent on the successful detection of the subsequent anomaly event a_k . In situations where no part of the subsequent event a_k is detected by a True-Detection, this Pre-Buffer detection is considered a false alarm rather than a meaningful early detection. Consequently, this early prediction $p_l(\theta)$ is reclassified as False Positive (the Outside category, which is discussed below). Further details are given in Appendix 2.C. In Figure 2.2, p_4 and p_{6-1} are the examples of pre-buffer detection category, whereas p_7 is not considered in this category.

- **Outside:** Sub-segments of the prediction event $p_l(\theta)$ located outside the ranges of anomaly event a_k and its buffer zones. These are instances where the model incorrectly flags normal behavior as anomalous (False Positive), like segments p_3 and p_7 in Figure 2.2.

ANOMALY EVENTS CATEGORIES:

- **Total Missed Anomalies:** When an entire anomaly event a_k is not detected by any segments of the prediction event $p_l(\theta)$, that is, all detections are before $i_k - e$ or after $n_k + d$. This category indicates a complete failure (False Negative) of the model to identify the anomaly. See segment a_4 in Figure 2.2.

- **Partial Missed Anomalies:** This category is assigned when only a part of anomaly event a_k is detected by the prediction events $p_l(\theta)$'s, but there are segments within the anomaly range of a_k that remain undetected. This category not only highlights the model's capability to detect parts of an anomaly but also its inability to identify

the anomaly event in its entirety. For instance, segment a_2 in Figure 2.2, where a part of it is detected by p_5 but before and after p_5 we have partially missed segments.

2.2.2. WEIGHTING PROCESS

After each individual time point is assigned to its category, we define weights for each of these points to determine their contribution to the True Positive (TP), False Positive (FP), and False Negative (FN) metrics of the detector. It is important to note that time points at which no anomaly is present and no prediction is made, True Negatives (TN), do not actively contribute to the performance metrics and are, therefore, implicitly assigned a weight of zero, reflecting their non-contribution. The bottom half of Figure 2.2 visually represents the variations in weights across all different categories.

- **True-Detection Weights:** Each point t from the True-Detection category, lying within the range of an anomaly event $[i_k, n_k]$, is considered correctly identified. Thus, such points are assigned the maximum weight of 1 as True Positives:

$$w^{\text{TP}}(t) = 1 \quad \text{for } t \in \text{TrueDetection } p_l(\theta) \quad (2.1)$$

- **Post-Buffer Detection Weights:** Each point t from the post-buffer category, in the range of $(n_k, n_k + d]$, is evaluated in relation to the anomaly event a_k . These points, while not being true positives in the traditional sense, receive a weight based on their proximity to the a_k , which captures the diminishing influence of an anomaly over time as the distance from the anomaly event increases.

$$w^{\text{TP}}(t) = 1 - \frac{\sum_{y=i_k}^{n_k} |t - y|}{\sum_{y=i_k}^{n_k} |(n_k + d) - y|} \quad \text{for } t \in \text{Post-Buffer } p_l(\theta) \quad (2.2)$$

Here, the numerator calculates the distance of t from each point within the anomaly event, and the denominator normalizes this against the total potential spread within the buffer zone. With this method, we account for the proximity to the entire anomaly, not just its endpoint. Thus, we address the delayed detection by recognizing that any point within the actual anomaly range might influence predictions in the buffer zone, not just the most immediate or final points of the anomaly. This also implies that the lengths of the anomalies influence the weights. For smaller anomalies, points in the Post-Buffer zone are closer to the anomaly onset, and will therefore be assigned with higher true positive weights. Further details, regarding the impact of anomaly length on the weights, are given in Appendix 2.B.

In the Post-Buffer zone, as the distance from a_k increases, the likelihood of a detection being a False Positive rises. Thus, the weights assigned to false positives are calculated as the complement of the TPs weights, acknowledging the reduced significance of detections further from the actual anomaly. Figure 2.2 visually shows the variations in TP and FP weights across the Post-Buffer categories (p_2 and $p_{6(3)}$).

$$w^{\text{FP}}(t) = 1 - w^{\text{TP}}(t) \quad \text{for } t \in \text{Post-Buffer } p_l(\theta) \quad (2.3)$$

- **Outside Weights:** Each point t from the Outside category indicates a situation where the model incorrectly identifies normal behavior as anomalous. Given the lack of proximity to any real anomaly, these points are considered FPs with a maximum weight of 1, reflecting a significant deviation from accurate detection.

$$w^{\text{FP}}(t) = 1 \quad \text{for } t \in \text{Outside } p_l(\theta) \quad (2.4)$$

- **Pre-Buffer Detection Weights:** Each point t in the pre-buffer category, in the range of $[i_k - e, i_k)$, is assessed for potential early detection in relation to the preceding a_k . These points, while not being true positives in the conventional sense, are evaluated for their proximity to the upcoming anomaly:

$$w^{\text{TP}}(t) = 1 - \frac{\sum_{y=i_k}^{n_k} |y - t|}{\sum_{y=i_k}^{n_k} |(i_k - e) - y|} \quad \text{for } t \in \text{Pre-Buffer } p_l(\theta) \quad (2.5)$$

Here, the numerator represents the distance of t from every point in a_k , capturing how early t occurs relative to the anomaly. The denominator provides normalization against the total potential spread within the pre-buffer zone. This mechanism recognizes that any point within the anomaly event might have an influence on the zone.

Similar to the Post-Buffer zone, the likelihood of a point being a False Positive increases as the distance from the i_k increases. Thus, the weights assigned to FPs are calculated as the complement of the TPs weights, reflecting the reduced relevance of premature detections. Figure 2.2 shows the variations in weights of the Pre-Buffer categories (p_4 and $p_{6(1)}$).

$$w^{\text{FP}}(t) = 1 - w^{\text{TP}}(t) \quad \text{for } t \in \text{Pre-Buffer } p_l(\theta) \quad (2.6)$$

- **Total Missed Anomalies Weights:** When the entire range of a_k is undetected, each t within its interval receives a maximum False Negative weight of 1. This assignment underscores the complete failure of the model in detecting the anomaly event. Figure 2.2 shows the variations in FN weight across a_4 as a total missed event.

$$w^{\text{FN}}(t) = 1 \quad \text{for } t \in \text{Total-Missed } a_k \quad (2.7)$$

- **Partial Missed Anomaly Weights:** When a_k is only partially detected, the undetected points t within a_k , are evaluated based on their proximity to the start of the anomaly event. The closer the points are to the anomaly onset the higher the FN weight, emphasizing the onset response time in detection. Here for $t \in \text{Partial Missed } a_k$, we have:

$$w^{\text{FN}}(t) = \begin{cases} 1 & \text{if } t \leq i_k + r \\ 1 - \frac{\sum_{y=i_k}^{i_k+r} |t - y|}{\sum_{y=i_k}^{n_k} |n_k - y|} & \text{otherwise} \end{cases} \quad (2.8)$$

Here, r is the size of the buffer that starts from the onset of the anomaly event. Undetected points in this buffer are penalized with a maximum FN weight of 1. Undetected points outside the buffer received a reduced FN weight, weighted by the distance to the buffer. The rationale behind this design is that more comprehensive coverage of an anomaly by a prediction justifies a more lenient assessment of its exact timing accuracy. In other words, when a prediction successfully captures a larger portion of \mathbf{a}_k , the precision of its onset timing becomes less critical. Therefore, r is defined as the fraction of coverage of \mathbf{a}_k by its corresponding $\mathbf{p}_l(\theta)$. Figure 2.2 shows the variations in FN weight across the Partial Missed category where some segments of \mathbf{a}_2 are missed.

2.2.3. PATE FINAL SCORE

The PATE final metric is designed to comprehensively evaluate anomaly detection by considering a full range of combinations of pre-buffer (e) and post-buffer (d) sizes. For each combination of e and d , we apply a range of thresholds (θ) to convert the continuous anomaly scores (\mathcal{S}) into binary predictions, capturing the model's performance across different sensitivity levels. Based on these binary predictions, we identify the prediction events \mathcal{P} and then categorize all prediction and anomaly events. Based on this categorization, we assign appropriate weights to each observation.

We calculate weighted Precision and Recall across all thresholds in the considered range for each specific combination of e and d . Using these calculations, we construct the Precision-Recall curve for each combination and compute the area under the curve (AUC-PR). Note that the weights $w^{\text{TP}}(t)$, $w^{\text{FP}}(t)$, and $w^{\text{FN}}(t)$ are assigned based on the categorization of each time point t . For time points that do not fall into any specific category, the weights are considered to be 0. Thus, the summation in the formulas for Precision and Recall effectively includes only those time points that have been categorized.

$$\text{Precision}_{e,d}(\theta) = \frac{\sum_{t=1}^T w^{\text{TP}}(t)}{\sum_{t=1}^T w^{\text{TP}}(t) + \sum_{t=1}^T w^{\text{FP}}(t)} \quad (2.9)$$

$$\text{Recall}_{e,d}(\theta) = \frac{\sum_{t=1}^T w^{\text{TP}}(t)}{\sum_{t=1}^T w^{\text{TP}}(t) + \sum_{t=1}^T w^{\text{FN}}(t)} \quad (2.10)$$

Finally, the overall PATE score is determined by averaging the computed AUC-PRs across all combinations of e and d :

$$\text{PATE} = \frac{1}{|E| \times |D|} \sum_{e \in E, d \in D} \text{AUC-PR}_{e,d} \quad (2.11)$$

Here, $|D|$ and $|E|$ represent the number of distinct values for d and e within their respective sets.

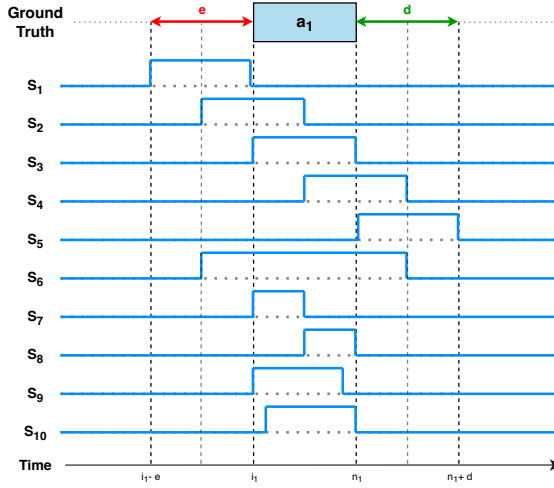
2.3. EXPERIMENTS AND RESULTS

2.3.1. SYNTHETIC DATA EXPERIMENTS

To highlight the merits of PATE, we first compare PATE with alternative evaluation metrics on a synthetic time series with a binary anomaly detector. The alternative measures can be threshold-dependent or independent. Threshold-independent metrics are inherently evaluated across a range of possible thresholds. For this example, we consider thresholds $\theta = \{0, 1\}$ to distinguish between normal and anomalous predictions. For threshold-dependent metrics, we define the optimal threshold as $\theta = 1$, identifying points predicted as '1' (anomalous) for evaluation.

Figure 2.3 shows anomaly a_1 with its pre and post-buffer zones. Below, ten different detection scenarios are shown, S_1, \dots, S_{10} . Results in Table 2.2 demonstrate that PATE effectively distinguishes the scenarios based on temporal proximity, duration, coverage level, and response timing. For instance, although S_1 is temporally close to the anomaly event, it fails to detect any part of it. In the context of time series, where past data is crucial for prediction, the inability to detect any part of the anomaly after it starts suggests that the prediction might be a true false alarm rather than a meaningful early detection. A low score for S_1 reflects a metric that appropriately penalizes lucky guesses or irrelevant detections. On the other hand, S_2 gets a higher score as it captures part of the anomaly itself, and then the non-overlapping part can be recognized as relevant early detection, which should be valued. Note that the PATE score of 0.03 for S_1 is not exactly zero because it considers a range of thresholds, including zero. At a threshold of 0, every point is labeled as a potential anomaly, thus increasing both true and false positives. This broad consideration prevents the PATE score from being zero for this specific example.

Meanwhile, S_1 and S_2 should be evaluated differently from delayed detections S_4 and S_5 . Although S_4 's coverage level is the same as that of S_2 , due to response timing, it gets a lower score. Similarly, the evaluation of S_5 is completely different from S_1 as it occurs after the anomaly event. This late detection might indicate that the model is responding to the anomaly, albeit with a significant delay. Hence, it is reasonable to evaluate S_5 higher than S_1 as it could reflect some response to the actual anomaly, even though it is late and fails to detect any part of the anomaly. Other metrics, while effective in certain scenarios, do not distinguish between the finer details of anomaly detection. For instance, these metrics just mirror the results of S_1 and S_2 for S_4 and S_5 without considering the early and delayed context. Moreover, S_3 , as an example of accurate detection, is expected to get the maximum score of 1 by all evaluation metrics, and S_6 is expected to get a lower score than S_3 . However, the VUS-ROC/PR metrics fail to evaluate these scenarios correctly. The scenarios S_7 , S_8 , S_9 , and S_{10} further exemplify the importance of the coverage level and response timing in detection. In each pair, S_7 and S_9 detect the anomaly right from the start; thus they should get scored higher than S_8 and S_{10} . While other metrics tend to score these pairs similarly, PATE recognizes the earlier detections in S_7 and S_9 and gives them higher scores. Moreover, in scenarios like S_9 and S_{10} , where the anomaly is covered more extensively, PATE assigns less penalties for response timing inaccuracies. This is seen in the smaller score difference between early and late detections in scenarios with greater coverage.



2

Figure 2.3: *Illustration of examples with synthetic data.* The figure shows the placement of different anomaly scores S from a binary anomaly detector.

Table 2.2: *Comparison of evaluation metrics for synthetic data examples depicted in Figure 2.3.* 'F1' refers to the F1 Score. 'Standard-F1' specifically denotes the conventional F1 Score calculated from standard Precision and Recall.

Scenarios	Threshold-independent Metrics					Threshold-dependent Metrics				
	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR	Standard-F1	PA-F1	R-based-F1	ETS-Aware-F1	Affiliation-F1
S_1	0.03	0.63	0.37	0.48	0.02	0.00	0.00	0.00	0.00	0.94
S_2	0.76	0.79	0.72	0.74	0.51	0.50	0.80	0.60	0.75	0.98
S_3	1.00	0.87	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_4	0.69	0.79	0.70	0.74	0.51	0.50	0.80	0.60	0.75	0.98
S_5	0.31	0.63	0.34	0.48	0.02	0.00	0.00	0.00	0.00	0.94
S_6	0.87	0.99	0.91	0.98	0.75	0.67	0.67	0.75	0.86	0.98
S_7	0.85	0.69	0.71	0.75	0.76	0.67	1.00	0.75	0.86	0.99
S_8	0.77	0.69	0.71	0.75	0.76	0.67	1.00	0.75	0.86	0.99
S_9	0.95	0.78	0.79	0.88	0.88	0.86	1.00	0.89	0.93	1.00
S_{10}	0.88	0.78	0.79	0.88	0.88	0.86	1.00	0.89	0.93	1.00

2.3.2. REAL-WORLD DATA EXPERIMENTS

To validate the practicality and effectiveness of PATE in real-world applications, we extracted some examples from the publicly available and widely used datasets, UCR-KDD21 [12] and MIT-BIH Arrhythmia (MBA) ECG [13]. The goal is to evaluate how well PATE, alongside other evaluation metrics, distinguishes between various detection models. To ensure a fair comparison, we compare PATE with threshold-independent evaluation metrics, guaranteeing an unbiased comparison of metrics performances.

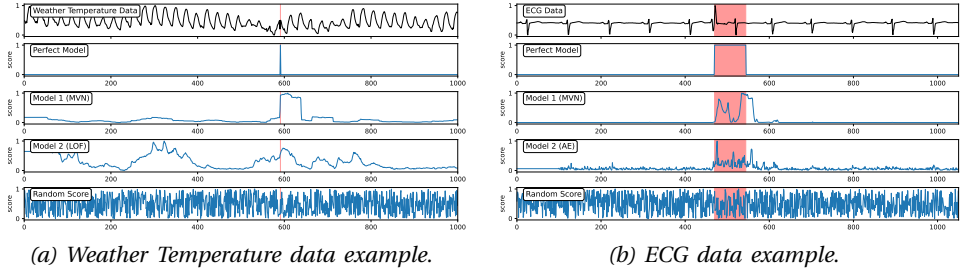


Figure 2.4: *Real-World Datasets and Anomaly Scores of Different Models.* The anomalous segment and its corresponding region (labeled by an expert), against which the models' predictions are compared, is highlighted in red

Table 2.3: *Quantitative Evaluation of Anomaly Detection Models.* Evaluation score for different anomaly detection models in detecting the anomalous region in examples of Figure 2.4.

Datasets	Weather Temperature					ECG				
	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR
Perfect Model	1.00	0.55	0.57	1.00	1.00	1.00	0.90	0.91	1.00	1.00
Model 1	0.88	0.98	0.71	0.98	0.02	0.83	0.99	0.89	0.98	0.69
Model 2	0.07	0.86	0.14	0.83	0.01	0.79	0.98	0.81	0.97	0.69
Random Score	0.02	0.67	0.08	0.66	0.01	0.07	0.56	0.11	0.43	0.06

We analyzed the anomaly scores generated by 1) a Perfect Model, which serves as the benchmark by perfectly identifying anomalies; 2) established models like MultiVariate Normal distribution (MVN) [14], Autoencoder (AE)[15], and Local Outlier Factor (LOF)[16]; 3) a baseline Random Score that assigns scores uniformly at random from a [0, 1] distribution. This selection covers a spectrum from theoretically ideal to practically random, offering a comprehensive view of the metrics' potential evaluation range. Detailed implementation of the models is available in our public code repository.

Figure 2.4 showcases two real-world examples: (a) Weather Temperature data from UCR-KDD21 and (b) ECG data. The top row of each example shows the time series data with actual anomalies highlighted in red. The next rows illustrate the output of the Perfect Model, and Models 1 and 2 (represented by MVN, LOF, or AE), demonstrating their respective detection scores. The final row displays a random score for baseline comparison. Table 2.3 quantitatively compares various metrics. PATE consistently rates the Perfect Model highest and the Random Score lowest, showing its capability to recognize optimal detection and effectively penalize poor performance. In contrast, VUS-ROC/PR and AUC-ROC metrics seem less capable of such differentiation with the baselines.

Moreover, PATE accurately takes into account the time series context and delayed

detection effect, offering a more realistic and conservative assessment compared to VUS-ROC and AUC-ROC metrics, which appear to overestimate the performance of Models 1 and 2. This overestimation is evident in the Weather Temperature data, where Model 2 is inaccurately scored high by VUS-ROC and AUC-ROC despite its poor detection. Additionally, AUC-PR is also not sensitive in evaluation. For instance, in the Weather Temperature data, Model 1's delayed yet successful detection is incorrectly evaluated with a very low score, similar to the detection of Model 2. Similarly, in the ECG data, PATE's evaluation reflects the inconsistent anomaly detection pattern of Model 2 (AE) compared to Model 1 (MVN). However, AUC-ROC/PR and VUS-ROC do not effectively consider this difference. Overall, PATE's assessments across both examples underscore its effectiveness in real-world applications.

2.3.3. IMPACT ANALYSIS: SOTA MODELS

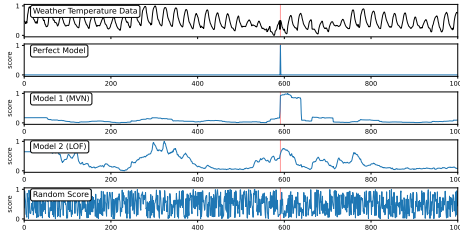
We re-evaluated several recent SOTA anomaly detection methods to not only assess their true performance but also to examine the stability of their ranking across various benchmark datasets when evaluated with different metrics, including PATE. Our comparative analysis includes models such as DCdetector [17], AnomalyTrans [18], and USAD [19], all of which have been recognized for their high performance in recent studies, alongside a Transformer and LSTM model, as simpler reconstruction-based anomaly detector baselines. These models are tested across the benchmark datasets of SMD [20], MSL [21], SWaT [22], and PSM [23], used in previous works. Implementation details are available in our public code repository.

In the literature on SOTA models, the PA-F1 is the most frequently used and widely accepted metric. Additionally, in some cases, the standard F1 Score and Point-Adjusted variant of AUC-ROC (PA-AUC-ROC) are also employed. For a comprehensive comparison, we included these metrics in our comparative analysis. Results, shown in Table 2.4, highlight a significant discrepancy between PATE scores and those obtained from other metrics like PA-F1, Standard F1 Score, and PA-AUC-ROC. Notably, models that performed exceptionally well under PA-F1 and PA-AUC-ROC, such as AnomalyTrans and DCdetector, exhibit markedly lower scores when evaluated with PATE. For instance, for the SMD dataset, AnomalyTrans achieves a PA-F1 score of 0.91, showcasing high performance, yet its PATE score is only 0.06, indicating a substantial reduction in performance. To visually illustrate the differences in detection quality, Figure 2.5 shows a portion of the anomaly scores for the SWaT and SMD. The figures show that AnomalyTrans and DCdetector models struggle with consistent detection. In particular, for the SWaT, the peaky detections by these models hardly align with the expert-labeled anomaly intervals, and the high values reported for PA-F1 and PA-AUC-ROC do not reflect this detection pattern. This suggests that these metrics may overestimate model effectiveness.

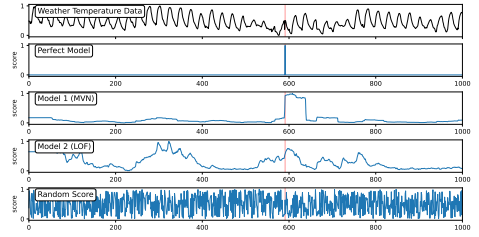
Next, Table 2.4 shows that the Standard F1 Score, AUC-ROC, and VUS-ROC, do not exhibit such overestimations. However, they lack sensitivity to the finer aspects of detection as discussed in section 2.2.1. For instance, on the SWaT dataset, the Standard F1 Score is not able to distinguish between the good performing LSTM and Transformer and the poorly performing AnomalyTrans and DCdetector, see also Figure 2.5 (a). Furthermore, AUC-ROC does not reflect the small differences between

Table 2.4: Comparison of SOTA anomaly detection model using different evaluation metrics across various benchmark datasets.

Datasets	SMD						MSL						SWaT						PSM					
Models			Standard-F1		PA-AUC-ROC				Standard-F1		PA-AUC-ROC				Standard-F1		PA-AUC-ROC				Standard-F1		PA-AUC-ROC	
	PATE	PA-F1					PATE	PA-F1					PATE	PA-F1					PATE	PA-F1				
AnomalyTrans	0.06	0.91	0.03	0.96	0.49	0.50	0.13	0.94	0.02	0.97	0.49	0.52	0.19	0.94	0.02	0.97	0.53	0.54	0.33	0.98	0.02	0.99	0.51	0.52
DCDetector	0.07	0.87	0.01	0.94	0.50	0.51	0.14	0.97	0.02	0.98	0.50	0.58	0.12	0.96	0.02	0.99	0.49	0.50	0.32	0.98	0.02	0.99	0.50	0.52
USAD	0.16	0.94	0.13	0.91	0.63	0.72	0.17	0.91	0.06	0.92	0.53	0.58	0.73	0.85	0.25	0.83	0.82	0.61	0.45	0.89	0.07	0.91	0.60	0.61
LSTM	0.25	0.80	0.14	0.87	0.76	0.81	0.19	0.82	0.08	0.87	0.57	0.64	0.71	0.82	0.03	0.85	0.82	0.60	0.55	0.93	0.15	0.94	0.73	0.73
Transformer	0.27	0.75	0.14	0.84	0.74	0.80	0.20	0.40	0.07	0.63	0.60	0.66	0.72	0.82	0.03	0.85	0.82	0.57	0.56	0.91	0.14	0.92	0.72	0.72



(a) Anomaly Scores of SOTA models for SWaT dataset.



(b) Anomaly Scores of SOTA models for SMD dataset.

Figure 2.5: Segments of anomaly scores of SOTA models for SWaT and SMD dataset. The highlighted regions in red indicate the true anomaly periods (labeled by an expert).

USAD, LSTM, or Transformer. The scores of this metric suggest that all models have an identical performance, that does not match the reality of their output. Moreover, while VUS-ROC offers a slightly better distinction among models than AUC-ROC, its limited scoring range (e.g., 0.54 for AnomalyTrans and 0.57 for Transformer) makes it challenging to clearly identify models that perform exceptionally well from those that do not. Meanwhile, PATE offers a more consistent and transparent assessment. It can be seen that PATE gives a relatively higher score to USAD (0.73), Transformer (0.72), and LSTM (0.71) according to their better detection pattern. PATE even slightly prefers USAD over LSTM, although the difference is small.

We also explored the average rankings of the models for all metrics across all four benchmark datasets. Figure 2.6 presents these rankings, highlighting noticeable differences in the standings of the models when using different metrics. The average rankings based on the PA-F1 metric place DCdetector at the forefront with an average rank of 1.62, followed by AnomalyTrans (1.88), USAD (3.00), LSTM (3.88), and Transformer (4.62). However, when evaluated with PATE, a significant shift occurs: Transformer and LSTM emerge as the top-performing models with ranks of 1.38 and 2.12, respectively, while AnomalyTrans and DCdetector drop to the bottom ranks of 4.50 each. This variance underscores the critical impact of the chosen evaluation metric and the importance of selecting a proper metric such as PATE.

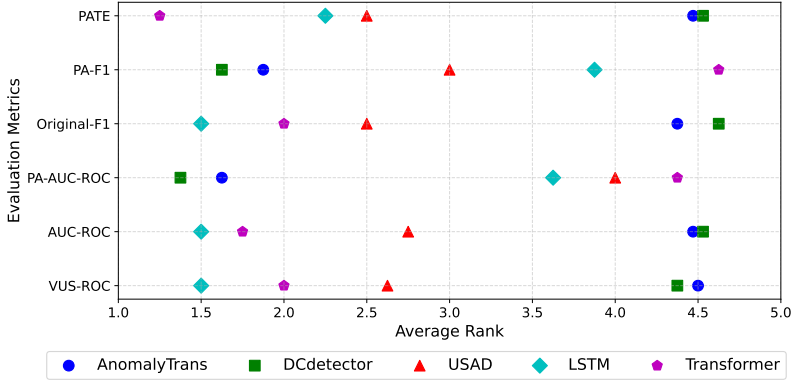


Figure 2.6: Average rankings of different models for various evaluation metrics across all benchmark datasets.

2.4. ABLATION ANALYSIS: BUFFER SIZES

The adaptability of PATE to accommodate different buffer sizes is one of its key strengths. This flexibility allows for an expert-driven and context-specific approach to model evaluation, ensuring that the unique characteristics of each dataset are appropriately considered. Figure 2.7 illustrates the mean performance of DCdetector, AnomalyTrans, USAD, LSTM, and Transformer across all four benchmark datasets using PATE. Results show that PATE consistently ranks models such as Transformer and LSTM the highest across different buffer sizes. This consistency in model rankings, irrespective of buffer size, highlights PATE’s robustness as an evaluation metric, and showcases PATE’s reliability for diverse applications, ensuring a consistent and dependable assessment for anomaly detection models.

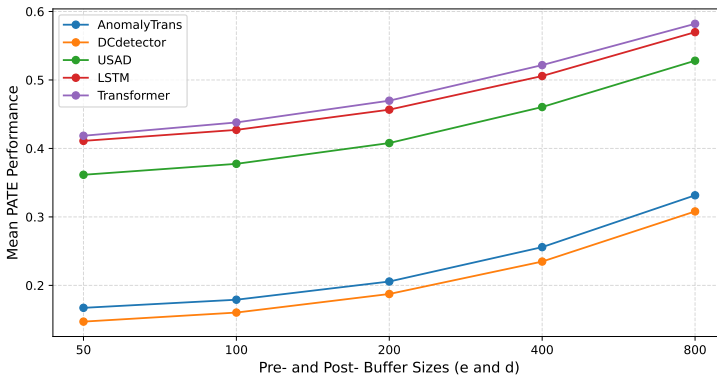


Figure 2.7: Mean PATE performance of all models across all datasets for different Pre and Post-Buffer sizes ($e = d$).

2.5. DISCUSSION AND CONCLUSION

We proposed PATE, a novel approach to evaluate anomaly detection models in time series data. PATE addresses the limitations of existing evaluation metrics by categorizing the anomaly and prediction events and assigning proximity-based weighting, considering different buffer zones around the anomaly event. PATE computes the area under the Precision-Recall curve, where the Precision and Recall are computed from weighted versions of True Positive, False Positive, and False Negative performances.

Our experiments with both synthetic and real-world data demonstrate that PATE effectively differentiates between models based on their actual performance, considering early and delayed detection, onset response time, coverage level of the anomaly event, and consistency in detection. The re-evaluation of SOTA anomaly detection methods using PATE reveals notable differences in performance assessments compared to other metrics. For instance, point-adjusted metrics often overestimate the performance of models. However, in practice, metrics such as ROC-AUC and VUS-ROC offer more reasonable estimates for SOTA models, though they might overlook subtle detection errors and sometimes lack discriminability between models. This analysis not only questions the true performance of current SOTA models but also indicates a shift in their rankings, challenging the prevailing understanding of the superiority of these models. PATE's ability to provide a more matching, context-sensitive, and transparent assessment highlights its potential as a more appropriate metric that can set a new standard for evaluating advancements in anomaly detection. Additionally, PATE's adaptability to various buffer sizes without compromising consistency and fairness in model evaluation further highlights its robustness and applicability across diverse applications.

To address the specific scenarios where either an expert has predetermined the threshold or models inherently output binary labels, we have developed *PATE-F1* as an essential extension of the original PATE framework. The methodology and experimental insights on *PATE-F1* are detailed in Appendix 2.D. *PATE-F1* effectively distinguishes between different scenarios based on temporal proximity, duration, coverage level, and response timing, setting it apart from other metrics that face limitations in capturing these aspects in evaluation. Additionally, our findings indicate that the original PATE framework, through strategic threshold application, naturally extends to effectively evaluate binary outputs. However, employing *PATE-F1* in such scenarios offers a more direct and simplified approach. This adaptation ensures PATE's methodology remains a versatile and applicable measure across a broader spectrum of anomaly detection approaches and contexts.

In conclusion, PATE represents a significant advancement in the evaluation of time series anomaly detection methods which has the potential to guide future research, influence industry adoption, and enhance the development of practical applications in critical domains such as healthcare and finance.

REFERENCES

- [1] R. Ghorbani, M. J. Reinders, and D. M. Tax. “PATE: Proximity-Aware Time Series Anomaly Evaluation”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 872–883.
- [2] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [3] R. Ghorbani, M. J. Reinders, and D. M. Tax. “Personalized anomaly detection in PPG data using representation learning and biometric identification”. In: *Biomedical Signal Processing and Control* 94 (2024), p. 106216.
- [4] C. C. Aggarwal and C. C. Aggarwal. *An introduction to outlier analysis*. Springer, 2017.
- [5] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich. “Precision and recall for time series”. In: *Advances in neural information processing systems* 31 (2018).
- [6] W.-S. Hwang, J.-H. Yun, J. Kim, and H. C. Kim. “Time-series aware precision and recall for anomaly detection: considering variety of detection result and addressing ambiguous labeling”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 2241–2244.
- [7] W.-S. Hwang, J.-H. Yun, J. Kim, and B. G. Min. “Do you know existing accuracy metrics overrate time-series anomaly detections?” In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 2022, pp. 403–412.
- [8] A. Huet, J. M. Navarro, and D. Rossi. “Local evaluation of time series anomaly detection algorithms”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 635–645.
- [9] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, *et al.* “Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 187–196.
- [10] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon. “Towards a rigorous evaluation of time-series anomaly detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 7194–7201.
- [11] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. “Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection”. In: *Proceedings of the VLDB Endowment* 15.11 (2022), pp. 2774–2787.
- [12] R. Wu and E. Keogh. “Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress”. In: *IEEE Transactions on Knowledge and Data Engineering* (2021).

- [13] G. B. Moody and R. G. Mark. "The impact of the MIT-BIH arrhythmia database". In: *IEEE engineering in medicine and biology magazine* 20.3 (2001), pp. 45–50.
- [14] S. Chakraborty. *An Intermediate Course in Probability*. 2011.
- [15] M. A. Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE journal* 37.2 (1991), pp. 233–243.
- [16] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.
- [17] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun. "DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection". In: *arXiv preprint arXiv:2306.10347* (2023).
- [18] J. Xu, H. Wu, J. Wang, and M. Long. "Anomaly transformer: Time series anomaly detection with association discrepancy". In: *arXiv preprint arXiv:2110.02642* (2021).
- [19] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga. "Usad: Unsupervised anomaly detection on multivariate time series". In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 3395–3404.
- [20] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei. "Robust anomaly detection for multivariate time series through stochastic recurrent neural network". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2828–2837.
- [21] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom. "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 387–395.
- [22] A. P. Mathur and N. O. Tippenhauer. "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)*. IEEE. 2016, pp. 31–36.
- [23] A. Abdulaal, Z. Liu, and T. Lancewicki. "Practical approach to asynchronous multivariate time series anomaly detection and localization". In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2485–2494.

APPENDICES

2.A. REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, the source code, along with comprehensive documentation, is publicly available at: <https://github.com/Raminghorbanii/PATE>.

This repository includes detailed instructions for using PATE, including how to set the buffer size, and complete descriptions of all models implemented for our experiments, covering configuration settings, training procedures, and experimental details to ensure accurate replication. Researchers seeking additional information are encouraged to contact the corresponding author.

2.B. EFFECT OF ANOMALY LENGTH ON BUFFER WEIGHTS

To explore the effect of anomaly length on the assignment of weights within the PATE framework, we consider three distinct anomaly events with varying durations: a_1 , a_2 , and a_3 , with a_1 being the longest and a_3 the shortest. Each was followed by a post-buffer zone of fixed size d . Figure 2.B.1 depicts the potential True Positive (TP) weights along the timeline, capturing the period before the anomaly, within its range, and throughout the post-buffer zone. The analysis of this figure indicates that TP weights for detections in the post-buffer zone are higher for a_3 , the shortest anomaly, and progressively lower for a_1 and a_2 , the longer anomalies. This observation underscores the direct correlation between the duration of an anomaly and the corresponding TP weights assigned to post-buffer detections. Higher TP weights for detections following shorter anomalies signify the critical nature of these detections, as they are in closer proximity to the anomaly onset. The PATE weighting mechanism accommodates this by adjusting the weights based on the distance from detections to the entire anomaly. This phenomenon also extends to the pre-buffer zone, where early detections are similarly influenced by the length of the forthcoming anomaly.

2.C. CLARIFICATION ON EARLY AND DELAYED DETECTIONS

To understand the distinct approaches PATE takes toward Early Detection (in the pre-buffer zone) and Delayed Detection (in the post-buffer zone), it is essential to consider the foundational goal of this evaluation metric.

For an anomaly detector, the ability to learn from past data and accurately predict future anomalies is essential. An early prediction that fails to correspond with an actual, subsequent anomaly suggests a fundamental modeling failure of the data's underlying structure—like sounding an alarm for an event that never happens. Ideally, if a model detects early signs of an impending anomaly, it should also

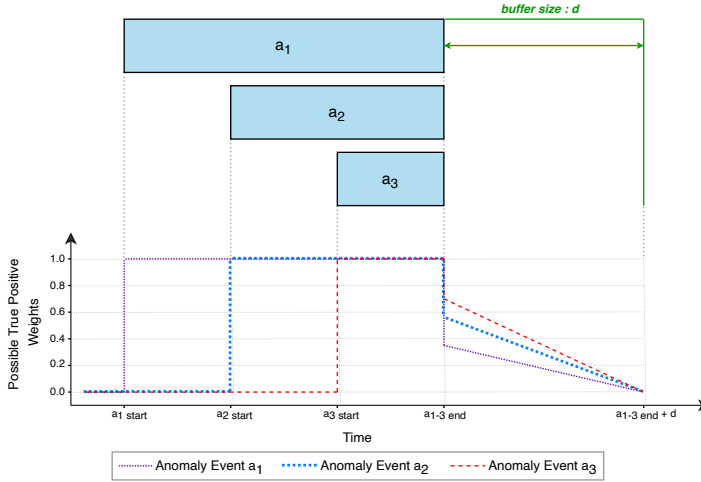


Figure 2.B.1: *Potential True Positive (TP) weights relative to the anomaly events with varying lengths.* The graph illustrates the higher TP weights for detections following the shortest anomaly event a_3 , and the progressively lower weights for the longer events a_1 and a_2 .

identify the anomaly when it occurs. The early signs—small changes or patterns of deterioration—lead to a larger and more evident departure from the norm. If the model has correctly identified these early signs, it should also recognize the anomaly itself, given the now more noticeable deviation. When the early detection is successfully followed by a true detection of the anomaly, the early detection is not considered just a lucky guess. It supports the model's predictive power and consistency.

In contrast, the context for delayed detection significantly differs as it showcases the capability of the model to identify anomalies post hoc. The model is apparently able to detect some deviation in the input, albeit a bit late. Such late detections still allow for the identification of the anomaly. Failing to have True Positive detections in the anomaly event is therefore not considered fatal for the Delayed Detection. Figure 2.C.1 shows the detection responses by three different models to an anomalous event, shown by the shaded area in red. Model 1 (top panel) reveals an early detection followed by True Positive detections, indicated by peaks aligning with the anomaly window. This pattern exemplifies an acceptable detection where the model preemptively and accurately identifies an anomaly. Model 2 (middle panel), however, demonstrates early detection without subsequent TPs during the actual anomaly, missing the critical deviation. This outcome might suggest a misinterpretation of the anomaly pattern by Model 2, potentially leading to a false alarm scenario. Conversely, Model 3 (bottom panel) shows a peak that arises post the onset of the anomaly, exemplifying a delayed detection. This detection is valued as it demonstrates the capacity of the model for retrospective analysis, acknowledging and learning from the anomaly event after its occurrence.

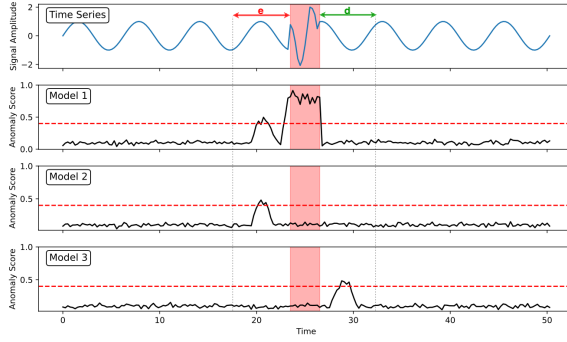


Figure 2.C.1: Comparative evaluation of model responses to an anomalous event in time series data.

2.D. PATE-F1 - ADJUSTED FOR BINARY SCORES

• **Methodology:** To enhance the applicability of PATE in scenarios where models use predetermined thresholds or where expert knowledge informs threshold determination, we propose an adapted version, *PATE-F1*. This adaptation leverages the core principles of PATE by assigning proximity-specific weights to categorized points and calculating weighted Precision and Recall. Unlike the original PATE, which evaluates a range of thresholds (θ), *PATE-F1* is tailored for binary scenarios, without the variation of thresholds but rather different combinations of buffer zones (e and d). For each combination, weighted Precision and Recall are calculated using equations 2.9 and 2.10 as detailed in Section 2.2.3. Subsequently, the F1 score for each combination is determined as follows:

$$\text{F1-Score}_{e,d} = 2 \times \frac{\text{Precision}_{e,d} \times \text{Recall}_{e,d}}{\text{Precision}_{e,d} + \text{Recall}_{e,d}} \quad (2.12)$$

The overall PATE-F1 score is then computed as the average of these F1 scores across all buffer zone combinations:

$$\text{PATE-F1} = \frac{1}{|E| \times |D|} \sum_{e \in E, d \in D} \text{F1-Score}_{e,d} \quad (2.13)$$

Here, $|E|$ and $|D|$ represent the number of distinct pre-buffer (e) and post-buffer (d) sizes, respectively.

• **Experimental Results:** We extend our analysis to *PATE-F1* by comparing the evaluations against threshold-dependent metrics, tailored for binary score predictions. Figure 2.D.1 shows 10 different detection scenarios shown by prediction events p_1, \dots, p_{10} . Table 2.D.1 shows that similar to the original PATE, *PATE-F1* effectively differentiates between scenarios based on temporal proximity, duration, coverage level, and response timing. This alignment with PATE's evaluation logic underlines the adaptability of our methodology to binary score scenarios without compromising the depth of analysis provided by the range of thresholds in the original framework.

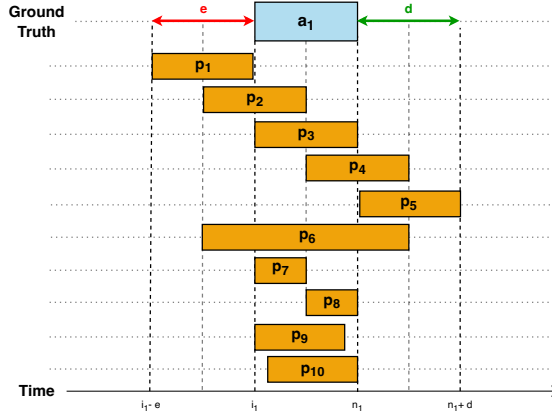


Figure 2.D.1: Examples with synthetic prediction events (binary scores). The figure shows the placement of different prediction events $p_l(\theta)$ from a binary anomaly detector.

Table 2.D.1: Comparison of evaluation metrics for synthetic prediction event examples depicted in Figure 2.D.1. 'F1' refers to the F1 Score.

Scenarios	Metrics						
	PATE	PATE-F1	Standard-F1	PA-F1	R-based-F1	ETS-Aware-F1	Affiliation-F1
p_1	0.03	0.00	0.00	0.00	0.00	0.00	0.94
p_2	0.76	0.75	0.50	0.80	0.60	0.75	0.98
p_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p_4	0.69	0.66	0.50	0.80	0.60	0.75	0.98
p_5	0.31	0.28	0.00	0.00	0.00	0.00	0.94
p_6	0.87	0.85	0.67	0.67	0.75	0.86	0.98
p_7	0.85	0.81	0.67	1.00	0.75	0.86	0.99
p_8	0.77	0.67	0.67	1.00	0.75	0.86	0.99
p_9	0.95	0.95	0.86	1.00	0.89	0.93	1.00
p_{10}	0.88	0.86	0.86	1.00	0.89	0.93	1.00

2.E. COMPLEXITY TIME ANALYSIS

We evaluated the computational efficiency of the PATE algorithm against established metrics like AUC-PR and VUS-PR through experiments on synthetic and real benchmark datasets when using a perfect anomaly detector. These experiments were conducted on a standard MacBook with a 2 GHz Quad-Core Intel Core i5 processor, Intel Iris Plus Graphics 1536 MB, and 16 GB RAM, reflecting the performance on commonly available hardware. Although PATE supports parallel execution to potentially decrease computation time, especially on High-Performance Computing (HPC) systems, we used a serial computation approach for consistent comparisons with other metrics.

• **Synthetic Data Experiments:** We generated synthetic time series data ranging from 1,000 to 100,000 points with anomaly ratios of 2%, 5%, and 10% to reflect various common scenarios. As shown in Figure 2.E.1, PATE's computation time increases linearly with data length and varies slightly with different anomaly ratios. Despite this, computation times remained under one second across all conditions, highlighting PATE's efficiency without parallel processing.

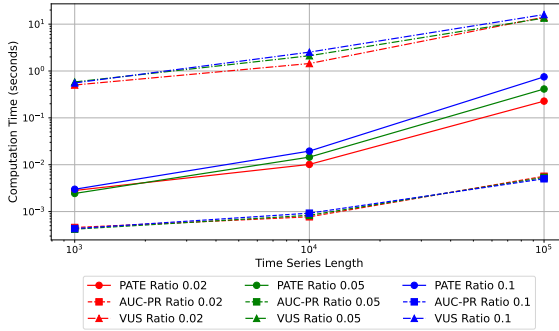


Figure 2.E.1: Computation time of PATE on synthetic data with varying lengths and anomaly ratios.

• **Benchmark Dataset Experiments:** We validated PATE on all standard benchmark datasets used in this study. As shown in Table 2.E.1, PATE's computation times are comparable to those of the AUC-PR metric and significantly faster than the VUS metric, remaining under one second for smaller datasets and under two seconds for larger ones. Note that further speed enhancements could be achieved on HPC systems or with parallel processing.

Table 2.E.1: Computation times (in seconds) for evaluation metrics across benchmark datasets.

Datasets	Time series Size	Anomaly Ratio	Evaluation Metrics		
			AUC-PR	VUS-PR	PATE
MSL	73700	10%	0.007	42.315	0.278
PSM	87800	4%	0.013	51.683	0.634
SWaT	449900	12%	0.267	249.573	1.895
SMD	708400	4%	0.064	462.252	1.796

3

SELF-SUPERVISED PPG REPRESENTATION LEARNING SHOWS HIGH INTER-SUBJECT VARIABILITY

With the progress of sensor technology in wearables, the collection and analysis of PPG signals are gaining more interest. Using Machine Learning, the cardiac rhythm corresponding to PPG signals can be used to predict different tasks such as activity recognition, sleep stage detection, or more general health status. However, supervised learning is often limited by the amount of available labeled data, which is typically expensive to obtain. To address this problem, we propose a Self-Supervised Learning (SSL) method with a pretext task of signal reconstruction to learn an informative generalized PPG representation. The performance of the proposed SSL framework is compared with two fully supervised baselines. The results show that in a very limited label data setting (10 samples per class or less), using SSL is beneficial, and a simple classifier trained on SSL-learned representations outperforms fully supervised deep neural networks. However, the results reveal that the SSL-learned representations are too focused on encoding the subjects. Unfortunately, there is high inter-subject variability in the SSL-learned representations, which makes working with this data more challenging when labeled data is scarce. The high inter-subject variability suggests that there is still room for improvements in learning representations. In general, the results suggest that SSL may pave the way for the broader use of machine learning models on PPG data in label-scarce regimes.

This chapter has been published as:

Ghorbani R, Reinders MJ, Tax DM. Self-supervised ppg representation learning shows high inter-subject variability in *Proceedings of the 2023 8th International Conference on Machine Learning Technologies (ICMLT)*, 2023 Mar 10, pp. 127-132. [1]

Code available at:

<https://github.com/Raminghorbani/SSL-PPG-shows-HighVariability>

3.1. INTRODUCTION

In recent years, wearables such as smartwatches and health trackers, equipped with a photoplethysmography (PPG) sensor, are becoming increasingly popular [2]. PPG is a non-invasive, low-cost optical measurement that can measure tissue blood flow over time following each pulse wave ejected from the heart. PPG works on the principle of pulse oximetry, wherein a sensor emits light to the skin and measures the intensity of light that is reflected or transmitted through the skin. Changes in arterial blood volume cause PPG signal variations [3, 4]. The cardiac rhythm corresponding to the PPG signal's periodicity can be used to obtain additional useful information from the users and predict various tasks. Some examples of research on PPG signals are related to Activity Recognition [5], Heart Rate Estimation [6], Blood Pressure Prediction [7], Biometric Identification [8], Sleep Staging Detection [9], and Atrial Fibrillation Detection [10].

In existing research, analyzing the PPG signals can be broadly categorized into signal processing and machine learning methods. The majority of machine learning solutions for PPG-based tasks utilize fully-supervised learning methods, which can be associated with several limitations. A fully-supervised learning setup usually requires considerable computational resources and time. Additionally, this setup requires large human-annotated datasets for high performance. Typically, obtaining labeled data is very costly and time-intensive, and the amount of labeled data is therefore insufficient in real-world applications, for instance, in the case of heart failure detection or fall detection. When automated detectors have to be trained on this type of problem, a good representation of the data with few numbers of informative features is essential [11]. Therefore, it is necessary to address the label-scarcity problem.

One approach to obtain a good informative representation is 'Self-Supervised Learning' (SSL). In SSL, two tasks are defined: a 'pretext' task and a 'downstream' task. The pretext task is the task of learning informative representations by itself. For instance, an auto-encoder tries to precisely reconstruct the input, squeezing the information through a bottleneck layer. It thereby learns a condensed, low-dimensional representation containing all necessary information to reconstruct the input exactly [12]. It is assumed that this learned low-dimensional representation reduces the complexity of the data by reducing anomalies and noise and, at the same time, improves the ability to detect patterns in the data simpler and better. Hence, learned representations from the pretext task should be helpful for learning a second-stage classifier on the downstream task, which is the actual task of interest that we want to solve.

The latest research in the field of machine learning shows the potential of SSL for finding generalized and robust representations [13–16]. The existing works in representation learning are generally concentrated on image-based applications where variations in the data could be visually observed. However, SSL is rarely applied to the field of time series data, especially biosignals. In recent years, some have applied SSL to time series data to show that this method can improve the representation, and they could confirm the potential of self-supervision in capturing important information even in the absence of labeled data. For instance, [17] introduced an Intra-inter Subject self-supervised Learning (ISL) model customized for ECG signals.

Their model integrates medical knowledge into self-supervision to effectively learn from intra-inter subject differences. Their results over different evaluation scenarios showed that the learned representations are information-rich and more generalizable than other state-of-the-art methods for diagnosing cardiac arrhythmias in label-scarce regimes. As another example, [16] investigated SSL to learn representations from EEG signals. They explored two pretext tasks based on temporal context prediction and contrastive predictive coding on two clinically EEG-relevant downstream tasks. The results show that linear classifiers trained on SSL-learned representations consistently outperform purely supervised deep neural networks in label-scarce regimes while reaching competitive performance when all labels are available. These findings are, however, not yet shown on noisy PPG signals, so it still remains to be shown whether self-supervision can bring improvements over standard supervised approaches on PPG signals.

In this paper, we focus on Human Activity Recognition (HAR) from PPG data. This is gaining interest since PPG data can be easily acquired from any of the widely available wearable devices [18]. Researchers have been exploring how SSL techniques can be either extended or explicitly designed for HAR tasks on accelerometer and gyroscope data. However, they have not yet looked into the PPG data specifically. In one of the early pioneering works, [19] used the task of identifying which signal transformation has been applied to a particular data sample as a pretext task using accelerometer and gyroscope data. The results show that SSL drastically reduces the requirement of labeled activity data, narrowing the gap between supervised and unsupervised techniques for learning meaningful representations.

Concluding, to the best of our knowledge, there are currently no studies using SSL on PPG data in label-scarce regimes. Therefore, we present the first detailed analysis of SSL tasks on PPG signals with attention to Activity Recognition as a downstream task. Our main contributions are: 1) Proposing a SSL framework for PPG data in label-scarce regimes, 2) Evaluating whether human activity recognition task can be done better when using SSL representations, and 3) Investigating the Inter-subject variability in PPG data and exploring how this is captured by the SSL representation.

3.2. PROPOSED FRAMEWORK

An overview of the proposed SSL framework is shown in Figure 3.1. We use an Autoencoder (AE) to learn a representation of the (unlabeled) data (unsupervised learning). Given an unlabeled dataset $D_U = \{\mathbf{x}_i\}_{i=1}^{N_u}$ where $\mathbf{x}_i \in \mathbb{R}^{1 \times T}$ is a vector of length T and N_u is the number of vectors (samples). The encoder maps each input vector into a latent space representation $\mathbf{h}_i = E_\phi(\mathbf{x}_i)$ where $\mathbf{h}_i \in \mathbb{R}^{1 \times d}$ where $d < T$. After that, \mathbf{h}_i is fed into the decoder component of the model, which follows the same approach to map \mathbf{h}_i to the output values $\hat{\mathbf{x}}_i = D_\theta(\mathbf{h}_i)$ where $\hat{\mathbf{x}}_i \in \mathbb{R}^{1 \times T}$. The encoder and decoder are parametrized by ϕ and θ , respectively. The AE is trained to minimize the mean squared error between \mathbf{x}_i and $\hat{\mathbf{x}}_i$ [20]:

$$\mathbf{L}_{Total} = 1/N_u \sum_{i=1}^{N_u} (1/T \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2) \quad (3.1)$$

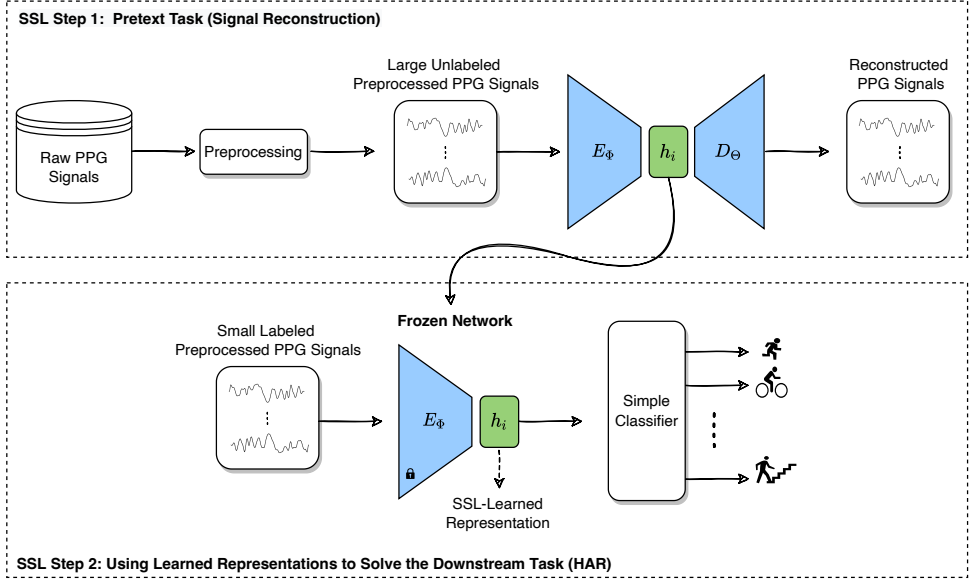


Figure 3.1: The overall proposed Self-Supervised framework

We have used a combination of Convolutional Neural Network layers (CNN) with the AE architecture, a Convolutional Neural Network AutoEncoder (CNN-AE). The main advantage of using a CNN-AE is a better reconstruction for the PPG signal as it exploits the correlations between time measurements in the PPG signal and thus captures the time-dependent information better.

For the downstream task, we use the original preprocessed PPG as the input to E_ϕ with frozen trained weights from the reconstruction task to get the related latent representation \mathbf{h}_i . This \mathbf{h}_i is used as input to train a simple classifier such as Logistic Regression (LR) or k-Nearest Neighbors (kNN). We do this because when the number of training samples is small, these simple classifiers often outperform more flexible and complex models. The classifier is trained on \mathbf{h}_i 's from several subjects and tested on a completely new subject.

3.3. EXPERIMENTAL SETUP

3.3.1. DATASET

We use the PPG-Dalia public dataset, which was collected by [21] for the PPG-based heart rate estimation task. This dataset contains recordings of 15 subjects performing daily activities such as *sitting*, *ascending/descending stairs*, *playing table soccer*, *cycling*, *driving a car*, *having lunch*, *walking*, and *working*. The measurements are obtained from wrist and chest-worn devices. Besides the activities, the transient periods between the activities are also recorded. We removed data from subject number 6 due to hardware issues during data recording. Note that *having lunch*, *driving*

a car, and *working* activities are categorized as concurrent or inter-leaved human activities, where actions of multiple activities are carried out simultaneously or where activities contain various activities while their actions can be interleaved in a shuffled manner [22, 23]. Therefore, we only considered the remaining five human activities for further study.

3.3.2. DATA PREPROCESSING

For the pretext task, a band-pass 2^{th} order Butterworth filter with low and high frequencies of 0.1 - 6 Hz is applied to the whole PPG signal of each subject individually. The filtered signal is normalized to zero mean and unit variance per subject. The final normalized filtered signals are segmented into a fixed window size of 8 seconds while two successive windows overlap by 6 seconds in the test dataset (this setting is common for PPG data). To increase the training set size, the two successive windows overlap by 7 seconds in the training dataset. For the downstream task, the PPG signals are split into 8 seconds windows while two successive windows overlap by 6 seconds in both training and test datasets. Activity labels are assigned to the corresponding PPG windows based on the available annotations.

3.3.3. IMPLEMENTATION

PROPOSED SSL

The hyperparameters and the architecture of the proposed CNN-AE are systematically determined by searching through all possible combinations to obtain the best performance. Eventually, we used a CNN-AE architecture deep learning model consisting of three convolution layers, followed by the Exponential Linear Unit (ELU) activation function, Batch Normalization, and MaxPooling layers. The decoder consists of the hidden layers in the reverse order of the encoder section. The Adam optimizer with a learning rate of 0.01, a decay rate of 0.001, and a clip-norm value of 0.9 are used. The batch size is 128, and training runs for 200 epochs. Finally, the parameters of all layers are randomly initialized. To assess the randomness of the deep learning framework, each training process for each test subject is repeated five times. Leave-One-Subject-Out cross-validation (LOSO) is used to evaluate the reconstruction performance.

For the downstream task, two simple classifiers are trained on the SSL-learned representations separately: a Logistic Regression and a kNN classifier (SSL-LR and SSL-kNN, respectively). The SSL-LR is regularized with the L2 penalty term and is solved using LIBLINEAR [24]. The SSL-kNN is trained with reweighted neighbors [25], where points are weighted by the inverse of their distance. Therefore, closer neighbors of a query point will have a larger influence than far away neighbors. Due to the different number of training samples which are 2, 5, 10, 50, and 1000 per class, the number of neighbors is selected as 8, 19, 39, 115, and 350, respectively. The LOSO is used to evaluate the AUC performance.

COMPARATIVE BASELINES

The performance of the SSL method is compared with two other baseline models: a simple and a more complex one. The simple baseline model (a typical baseline in SSL research) is trained directly on the original preprocessed PPG representations and consists of the encoder part of the CNN-AE from the pretext task, extended with one classification layer at the end. The encoder part is thus trained on the classification task immediately and not in a self-supervised setting. The Adam optimizer with a learning rate of 0.001 and a clip-norm value of 0.6 are used. The batch size is 128, and training runs for 200 epochs. The more complex baseline is a CNN-LSTM model also trained directly on the original preprocessed PPG representations. The architecture of the complex baseline consists of a convolution layer followed by a hyperbolic tangent function, Batch Normalization, MaxPooling layers, and then a LSTM layer with a hyperbolic tangent activation function, followed by a classification layer at the end. The Adam optimizer with a learning rate of 0.001 and a clip-norm value of 0.6 are used. The batch size is 128, and training runs for 200 epochs. To assess the randomness of these Deep Learning frameworks, each training process for each test subject is repeated five times. Both baseline models use the LOSO to evaluate the AUC performance.

BIOMETRIC IDENTIFICATION (BI) FOR EXPLORING INTER-SUBJECT VARIABILITY

If there is a large inter-subject variability, the subjects should be easily discriminated in the representation. To check if a subject can indeed be easily discriminated, we train and evaluate a kNN classifier with reweighted neighbors ($k = 20$) per activity. Note that this classifier is not optimized at all on the SSL-learned representations; the kNN fully relies on the metric that is induced by the CNN-AE latent representation \mathbf{h}_i . A good performance of the kNN for the BI task suggests that the learned representation from CNN-AE is heavily biased towards encoding different subjects and not so many other tasks like activities. In this experiment, PPG data is preprocessed with the same steps as the downstream task preprocessing. Afterward, the PPG windows of each activity are selected to sample a separate balanced training set over the subjects. The 4-fold cross-validation (75% for the training and 25% for the test set) is used to evaluate the AUC performance.

3.4. RESULTS

3.4.1. PRETEXT TASK

To determine a suitable dimensionality d of the learned representation \mathbf{h}_i , we compute the relative MSE (i.e., MSE in Eq. (1) normalized by the total variance across test subjects' data) by varying d using the CNN-AE. The results are shown in Table 3.1. It can be seen that the reconstruction error decreases with increasing dimensionality d . As the representation with a lower dimension is more suited for learning with limited labels, we chose $d = 64$ when proceeding with the downstream task. Also, later experiments show that $d = 64$ leads to better performance on the downstream task compared to other dimensionalities.

Table 3.1: Mean Relative MSE results of test subjects (LOSO) for the signal reconstruction task by varying d using the CNN-AE.

Dimension of \mathbf{h}_i	Relative MSE Results
$d = 2$	0.83 ± 0.02
$d = 8$	0.59 ± 0.03
$d = 32$	0.14 ± 0.03
$d = 64$	0.02 ± 0.00
$d = 128$	0.00 ± 0.00

3.4.2. DOWNSTREAM TASK

In Figure 3.2a, we show the AUC performances on the downstream task of predicting activity type for a varying number of training samples per class. The performance of the proposed SSL method is compared with the simple and complex baseline methods. As the number of training samples per class decreases, the performances of all methods drop, confirming the negative influence of when less and less samples with labels are available. The linear SSL-LR model fails to improve the performance compared to the baseline models when a few (< 10) training samples per class are available. However, SSL-kNN, as a non-linear solution, outperforms the baselines and the SSL-LR in the label-scarce regimes. This suggests that the SSL-learned representation is still too complex for a simple linear solution like LR. One reason for such a behavior could be the high Inter-Subject variability in the SSL-learned representations. Figure 3.2b shows the SSL-kNN performance of each of the individual test subjects, for a varying number of training samples per class. It can be observed that the AUC performance can vary between 0.5 and 0.7 for a small training size of $N = 2$, and even for very large training sizes of $N = 1000$, the AUC still varies between 0.55 and 0.8. This indicates that the data distributions of different subjects vary significantly, thus indicating large inter-subject variability.

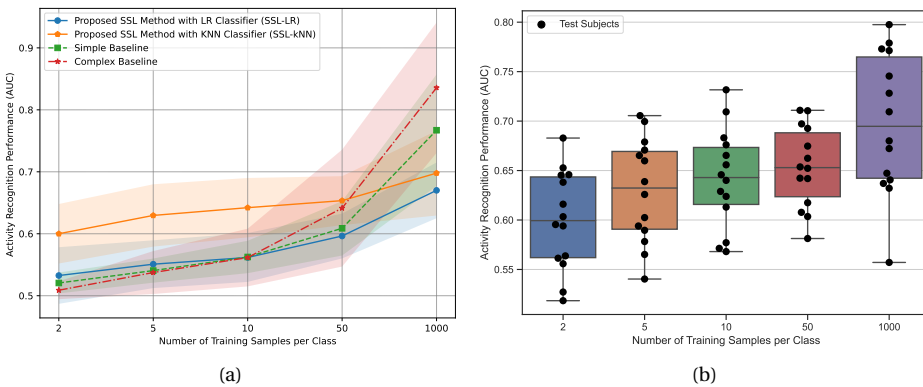


Figure 3.2: Results for Activity Recognition downstream task: (a) Mean AUC performance of test subjects over the different number of training samples per class and (b) variability in AUC performance among test subjects using SSL-kNN.

To explore the inter-subject variability more deeply, we also investigated the possibility of classifying subjects during each activity using a kNN model. Note that this is now different than our initial domain task; here, we are interested in whether subjects are still separable in the latent representations \mathbf{h}_i (which, in principle, is undesired when generalizing over subjects). The results of the BI task in Table 3.2 show that subjects can be discriminated perfectly for some activities such as *sitting* or *walking*. It can be seen that there are enough differences among subjects in the original as well as the SSL-learned representations. Moreover, the SSL-learned representation seems to highlight the inter-subject variation as its mean performance is consistently higher than on the original data. This suggests that the AE learned representation is more focused on encoding the different subjects and not so much on the domain tasks of interest, that is, predicting the different activities.

Table 3.2: Mean AUC Performance of test sets (4-fold cross-validation) for Biometric Identification task.

Activities	Input	
	Original Representation	SSL-Learned Representation
Sitting	0.84 ± 0.12	0.84 ± 0.01
Ascending/Descending Stairs	0.63 ± 0.02	0.64 ± 0.03
Playing Table Soccer	0.53 ± 0.01	0.61 ± 0.02
Cycling	0.59 ± 0.05	0.61 ± 0.06
Walking	0.72 ± 0.02	0.75 ± 0.02

3.5. DISCUSSION AND CONCLUSION

We have evaluated the usefulness of self-supervised representation for the activity recognition task when suffering from a label-scarcity in PPG data. The representation is not optimized on the downstream classification task (for which just a few labeled training samples may be available), but it is first optimized to perform a data reconstruction pretext task (for which no supervised information is needed). The results reveal that the SSL method can compete and outperform fully supervised baselines when a kNN model is trained on the SSL-learned representations in label-scarce regimes (with less than 50 samples per class). However, training a simple linear classifier like LR (instead of kNN) is not helpful since the inter-subject variability introduces too much non-linearities in the decision boundaries.

One should note that in the current study setup, fixed hyperparameters are used across the data regimes for all baseline models. When copious amounts of (unlabeled) data from all subjects would be available, all hyperparameters could be optimized for every different task.

The poor performance of the LR classifier on the SSL-learned representations shows that there is high inter-subject variability. High inter-subject variability makes the generalization more challenging. In this case, a subject-specific model could be a

solution for improving the performance over the learned representations. However, training a subject-specific model can be expensive because a large amount of (labeled) data that has to be obtained from each subject. Here, the SSL representation can come to the rescue, as we have shown that this representation can improve performance with respect to the original representation. However, there should be more focus on disentangling the inter- and intra-subject variability.

This matter opens the door for future research to learn more generalized informative PPG representations while addressing the inter-subject variability problem. For instance, removing the subject-specific factors in order to disentangle the inter-subject variations using a factor disentangling sequential autoencoder [26], or performing contrastive learning among subjects to learn distinctive representations [17] can be promising directions in learning informative PPG representations.

REFERENCES

- [1] R. Ghorbani, M. J. Reinders, and D. M. Tax. "Self-supervised ppg representation learning shows high inter-subject variability". In: *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*. 2023, pp. 127–132.
- [2] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran. "A review on wearable photoplethysmography sensors and their potential future applications in health care". In: *International journal of biosensors & bioelectronics* 4.4 (2018), p. 195.
- [3] A. Kamal, J. Harness, G. Irving, and A. Mearns. "Skin photoplethysmography—a review". In: *Computer methods and programs in biomedicine* 28.4 (1989), pp. 257–269.
- [4] T. Aoyagi and K. Miyasaka. "Pulse oximetry: its invention, contribution to medicine, and future tasks". In: *Anesthesia and analgesia* 94.1 (2002), S1–S3.
- [5] M. Boukhechba, L. Cai, C. Wu, and L. E. Barnes. "ActiPPG: using deep neural networks for activity recognition from wrist-worn photoplethysmography (PPG) sensors". In: *Smart Health* 14 (2019), p. 100082.
- [6] Z. Zhang, Z. Pi, and B. Liu. "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise". In: *IEEE Transactions on biomedical engineering* 62.2 (2014), pp. 522–531.
- [7] S. Ghosh, A. Banerjee, N. Ray, P. W. Wood, P. Boulanger, and R. Padwal. "Continuous blood pressure prediction from pulse transit time using ECG and PPG signals". In: *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*. IEEE. IEEE, 2016, pp. 188–191.
- [8] L. Everson, D. Biswas, M. Panwar, D. Rodopoulos, A. Acharyya, C. H. Kim, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte. "BiometricNet: Deep learning based biometric identification using wrist-worn PPG". In: *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. IEEE, 2018, pp. 1–5.
- [9] M. K. Uçar, M. R. Bozkurt, C. Bilgin, and K. Polat. "Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques". In: *Neural Computing and Applications* 29.8 (2018), pp. 1–16.
- [10] A. Aliamiri and Y. Shen. "Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. Las Vegas, NV, USA: IEEE, 2018, pp. 442–445.

- [11] P. K. Gyawali. *Learning with Limited Labeled Data in Biomedical Domain by Disentanglement and Semi-Supervised Learning*. Rochester Institute of Technology, 2021.
- [12] N. Tishby, F. C. Pereira, and W. Bialek. “The information bottleneck method”. In: *arXiv preprint physics/0004057* (2000).
- [13] M. Tagliasacchi, B. Gfeller, F. d. C. Quitry, and D. Roblek. “Self-supervised audio representation learning for mobile devices”. In: *arXiv preprint arXiv:1905.11796* (2019).
- [14] M. Kocabas, S. Karagoz, and E. Akbas. “Self-supervised learning of 3d human pose using multi-view geometry”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1077–1086.
- [15] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. “Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4006–4015.
- [16] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. “Uncovering the structure of clinical EEG signals with self-supervised learning”. In: *Journal of Neural Engineering* 18.4 (2021), p. 046020.
- [17] X. Lan, D. Ng, S. Hong, and M. Feng. “Intra-inter subject self-supervised learning for multivariate cardiac signals”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 4. 2022, pp. 4532–4540.
- [18] O. R. A. Almanifi, I. M. Khairuddin, M. A. M. Razman, R. M. Musa, and A. P. A. Majeed. “Human activity recognition based on wrist PPG via the ensemble method”. In: *ICT Express* (2022).
- [19] A. Saeed, T. Ozcelebi, and J. Lukkien. “Multi-task self-supervised learning for human activity detection”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–30.
- [20] M. A. Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2 (1991), pp. 233–243.
- [21] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. “Deep PPG: Large-scale heart rate estimation with convolutional neural networks”. In: *Sensors* 19.14 (2019), p. 3079.
- [22] M. S. Afzali Arani, D. E. Costa, and E. Shihab. “Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals”. In: *Sensors* 21.21 (2021), p. 6997.
- [23] N. A. Sakr, M. Abu-Elkheir, A. Atwan, and H. Soliman. “Data driven recognition of interleaved and concurrent human activities with nonlinear characteristics”. In: *Journal of Intelligent & Fuzzy Systems* 37.4 (2019), pp. 5573–5588.
- [24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. “LIBLINEAR: A library for large linear classification”. In: *the Journal of machine Learning research* 9 (2008), pp. 1871–1874.

- [25] S. A. Dudani. “The distance-weighted k-nearest-neighbor rule”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1976), pp. 325–327.
- [26] P. K. Gyawali, B. M. Horacek, J. L. Sapp, and L. Wang. “Sequential factorized autoencoder for localizing the origin of ventricular activation from 12-lead electrocardiograms”. In: *IEEE Transactions on Biomedical Engineering* 67.5 (2019), pp. 1505–1516.

4

REPRESENTATION LEARNING AND PERSONALIZATION FOR PPG ANOMALY DETECTION

Photoplethysmography (PPG) signals, typically acquired from wearable devices, hold significant potential for continuous fitness-health monitoring. In particular, heart conditions that manifest in rare and subtle deviating heart patterns may be interesting. However, robust and reliable anomaly detection within these data remains a challenge due to the scarcity of labeled data and high inter-subject variability. This paper introduces a two-stage framework leveraging representation learning and personalization to improve anomaly detection performance in PPG data. The proposed framework first employs representation learning to transform the original PPG signals into a more discriminative and compact representation. We then apply three different unsupervised anomaly detection methods for movement detection and biometric identification. We validate our approach using two different datasets in both generalized and personalized scenarios. Our results demonstrate significant improvements: for movement detection, in the generalized scenario, AUCs improved from barely 0.5 to above 0.9 with representation learning. Importantly, inter-subject variability was substantially reduced, from around 0.4 to below 0.1. In the personalized scenario, AUCs became close to 1.0, with variability further reduced to below 0.05, indicating the effectiveness of both representation learning and personalization for anomaly detection in PPG data. Similar enhancements were observed in biometric identification, emphasizing how our approach can minimize inter-subject variability and enhance PPG-based health monitoring systems.

This chapter has been published as:

Ghorbani, R., Reinders, M. J., & Tax, D. M. (2024). Personalized anomaly detection in PPG data using representation learning and biometric identification. *Biomedical Signal Processing and Control*, 94, 106216. [1]

4.1. INTRODUCTION

Photoplethysmography (PPG) data is a non-invasive, low-cost, optical physiological signal that measures the volume of blood flowing through the blood vessels and can be measured by a variety of wearable devices and smartwatches [2]. PPG data enables remote health monitoring and fitness tracking, which presents opportunities for identifying unusual patterns in the user data that may indicate potential health issues, like abnormal heart rate or irregular movement patterns [3].

The effectiveness of detecting anomalies largely depends on the availability of enough labeled data. Supervised machine learning methods, such as k-Nearest Neighbours (kNN), Random Forest, and Artificial Neural Networks (ANN), have been widely used in previous research to interpret PPG signals [4–8]. However, the process of data labeling is tedious, time-consuming, and costly, especially for anomaly detection problems, since anomalies seldomly occur in real-world applications. Furthermore, these supervised learning methods may be prone to bias and overfitting if the labeled dataset does not adequately represent the full range of normal and anomalous PPG signals. Moreover, these methods may not be adaptable to unknown or unexpected anomalies, as they learn to recognize patterns based on the examples provided in the training dataset, and may fail to effectively detect anomalies not represented in the dataset. To address these limitations, unsupervised anomaly detection methods can offer advantages over supervised approaches, as they do not rely on explicitly labeled examples of anomalous behavior and can be more adaptable to unknown or unexpected anomalies [9, 10].

Anomaly detection in PPG data can also be challenging due to other various factors that contribute to noise and inter-subject variability. These are factors like physical activity, stress, illness, measurement noise, age, gender, body composition, and genetic differences, as well as external factors such as sensor placement, sensor quality, and environmental conditions. This makes it difficult to develop generalized models that perform consistently across different individuals since each person's PPG signal may exhibit unique characteristics [11]. These complexities necessitate strategies to account for individual-specific characteristics.

Personalization can be a potential solution to help overcome the limitations of generalization by tailoring models to individual users [12]. However, the effectiveness of personalization hinges on accurate biometric identification. Inaccurate identification of individuals can lead to personalized models being trained on or applied to the wrong user's data, resulting in poor performance and potentially harmful outcomes. Hence, accurate biometric identification can enhance the reliability of personalized models, as it ensures that the models are based on the specific characteristics of each user.

In addition to unsupervised anomaly detection methods and personalization, representation learning can be particularly useful in enhancing performance [13]. Representation learning models are typically trained to learn from large amounts of unlabeled data, enabling them to extract a more compact, informative, and expressive representation of the data without the need for expensive and time-consuming manual labeling. By learning a lower-dimensional representation of the PPG data that captures its inherent structure and discriminative features, representation learning

can help overcome challenges posed by inter-subject variability, noise, and other factors affecting PPG signals. AutoEncoders, for example, are a type of self-supervised representation learning model that learns representations by encoding inputs into lower dimensions and then decoding them back to their original form, focusing on reconstructing the input [14]. Other representation learning models have been proposed with different tasks, such as contrastive learning or classification of augmented transformations of the original data [15, 16]. Representation learning has been increasingly used for anomaly detection in various domains, including image analysis [17–21], and time series data, such as bio-signals sensor data like EEG or ECG [22–25]. These studies have shown how representation learning can successfully extract meaningful features from complex bio-signals sensor data, leading to improved performance in classification tasks such as emotion detection or sleep stage classification. However, its application to PPG data for unsupervised anomaly detection and biometric identification remains underexplored, despite PPG being a commonly used bio-signal in health-monitoring applications.

In this paper, we present a two-stage framework for unsupervised movement detection and biometric identification in PPG data using representation learning. In the first stage, we train a deep neural network to obtain a lower-dimensional and informative data representation. In the second stage, we construct separate unsupervised anomaly detectors for both tasks using the learned representations from the first stage. Our approach not only investigates the effectiveness of representation learning in this context, but also explores the potential of personalization in enhancing anomaly detection performance. Additionally, we delve into biometric identification, aiming to improve the reliability of personalized anomaly detectors. To the best of our knowledge, this is the first study to jointly address these aspects for anomaly detection in PPG data. Summarizing, our contributions are:

1. We propose a two-stage framework for unsupervised anomaly detection and biometric identification in PPG data using representation learning.
2. We demonstrate the effectiveness of using the learned representations compared to the original representations in detecting difficult real-world anomalies and mitigating the subject variability.
3. We compare the effectiveness of generalization and personalization in anomaly detection, discussing the impact of tailoring models to individual users for enhancing the detection performance.
4. We investigate the unsupervised biometric identification task in PPG data to increase the reliability of personalized models.
5. We explore the impact of the dimensionality of the learned representation on the performance of our anomaly detection framework, demonstrating the robustness of representation learning across a wide range of dimensionalities.

4.2. PROPOSED FRAMEWORK

An overview of the proposed anomaly detection framework is shown in Figure 4.1. In the first step, we focus on obtaining a representation of the PPG data that captures the underlying structure of the data. Recent research shows that the task of classifying the original data and augmented transformed versions of the same data can outperform AutoEncoders and contrastive learning methods in learning better representation for the downstream task of interest [16]. Accordingly, we learn the representation by distinguishing original data from augmented transformed versions of the same data. This task is what we refer to as "Signal Transformation Classification."

Given the original signal $S(l)$, where $l = (1, 2, \dots, L)$ and L is the length of the time series, the augmented transformations of the data are described as:

- *Time reversal*: A time inverted version of the signal: $S'(l)$, where $l = (L, L-1, \dots, 1)$.
- *Amplitude reversal*: A amplitude inverted version of the signal: as $S'(l) = -S(l)$, where $l = (1, 2, \dots, L)$.
- *Both Time and Amplitude reversal*: We first perform the time reversal as described and then perform the amplitude reversal to obtain a time and amplitude inverted version of the signal: $S'(l) = -S(l)$, where $l = (L, L-1, \dots, 1)$.

To train the representations, we use a CNN model to classify PPG segments into four categories: Time reversal, Amplitude reversal, Both time and amplitude reversal, and the original signal. Given an unlabeled PPG dataset $D_U = \{\mathbf{x}_i\}_{i=1}^{N_u}$ where $\mathbf{x}_i \in \mathbb{R}^{1 \times T}$ is a vector of length T and N_u is the number of vectors (samples). $y_i \in \{1, 2, 3, 4\}$ is the class label for the i^{th} vector, where $y_i = 1, 2, 3$ represents the augmented data obtained by reversing the original PPG signal and $y_i = 4$ represents the original PPG signal. The CNN model consists of an encoder component that maps each input vector into a latent space representation $\mathbf{h}_i = E_\phi(\mathbf{x}_i)$ where $\mathbf{h}_i \in \mathbb{R}^{1 \times d}$ and $d < T$. After that, \mathbf{h}_i is fed into the classifier component of the model to predict the class label $\hat{y}_i = C_\theta(\mathbf{h}_i)$. The model is trained to minimize the cross-entropy loss between the predicted class label \hat{y}_i and the true class label y_i . The final learned representation is obtained by taking the latent space representation \mathbf{h}_i outputted by the encoder component. This learned representation is then used in the second stage of our proposed framework for anomaly detection.

In the second step of our proposed framework, we use the learned representation \mathbf{h}_i to detect anomalies. Specifically, we use three different methods to detect whether an input signal is an anomaly: Multi-Variate Normal distribution (MVN) [26], Isolation Forest (IF) [27], and PCA-Reconstruction [26]. For the MVN, the mean and covariance matrix are estimated on normal training samples. Given a test sample \mathbf{h}_{test} , we can then calculate the probability density function (PDF) of the test sample using the fitted Gaussian distribution as:

$$p(\mathbf{h}_{test}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{h}_{test} - \mu)^T \Sigma^{-1}(\mathbf{h}_{test} - \mu)\right) \quad (4.1)$$

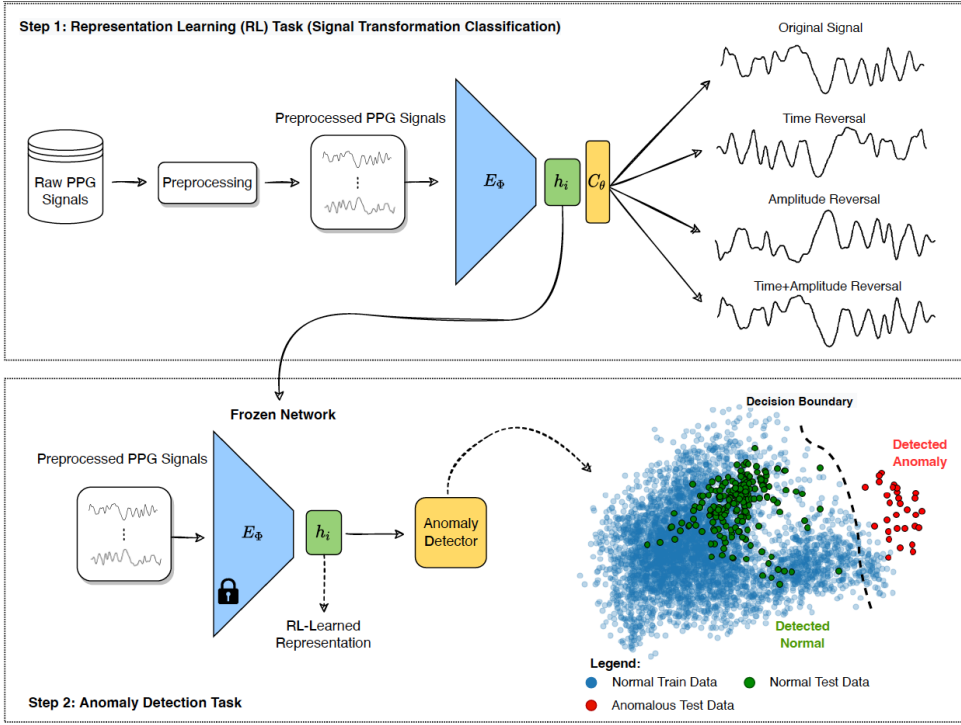


Figure 4.1: *Proposed framework for anomaly detection using Representation Learning (RL).* The framework consists of two steps: (1) A representation learning phase, where the model (consisting of an encoder and classifier component) is trained to discriminate between augmented transformations and the original data. The weights of the encoder are then frozen for the next step. (2) A anomaly detection phase, where the frozen encoder is used to extract features from the input data, which are then fed into an anomaly detector. The scatter plot illustrates an example of the distribution of points in a 2-dimensional feature space. The anomaly detector separates the normal and anomalous samples with the decision boundary (threshold) based on their anomaly scores.

where d is the dimension of the learned representation. It is expected that anomalous test samples have a lower probability compared to normal samples. Therefore, these points can be detected if the probability is below a set threshold.

For the Isolation Forest (IF) method, we first train an ensemble of decision trees on normal training samples. Given a test sample \mathbf{h}_{test} , the IF algorithm isolates the test sample from the others by recursively splitting the data with randomly selected features and split values. The number of splits, or the path length, required to isolate a sample is an indication of its anomaly score. Anomalous samples are expected to have shorter path lengths compared to normal samples. The anomaly score of a test sample \mathbf{h}_{test} using the IF algorithm is calculated as:

$$s(\mathbf{h}_{test}) = 2^{-\frac{E[L(\mathbf{h}_{test})]}{c(N)}} \quad (4.2)$$

where $E[L(\mathbf{h}_{test})]$ is the average path length of the test sample over all trees in the ensemble, $c(N)$ is the average path length of an unsuccessful search in a Binary Search Tree with N external nodes, and N is the number of samples in the training data. The anomaly score $s(\mathbf{h}_{test})$ ranges from 0 to 1, with higher scores indicating a higher likelihood of being anomalous. Anomalous test samples can be detected if the anomaly score is above a set threshold.

The PCA-Reconstruction method is a technique for detecting anomalies in high-dimensional data by reconstructing the original data from its principal components and evaluating the reconstruction error. Given a test sample \mathbf{h}_{test} , the reconstruction error can be calculated as the squared distance between the original sample and its reconstructed version (\mathbf{h}_{recon}) after mapping to a reduced PCA space. This is achieved by projecting the test sample \mathbf{h}_{test} onto the orthogonal basis vectors represented by the matrix describing the PCA mapping, W , and then transforming it back to the original space. The reconstruction error can then be expressed as:

$$e(\mathbf{h}_{test}) = \|\mathbf{h}_{test} - (WW^T)\mathbf{h}_{recon}\|^2 \quad (4.3)$$

Note that anomalous test samples can be detected if the reconstruction error is above a set threshold.

4.2.1. DEFINITION OF ANOMALIES

We define anomalies in the context of two specific tasks: activity movement detection and biometric identification.

ACTIVITY MOVEMENT DETECTION

In this particular setting, we train an anomaly detector on the recorded data during a specific activity (considered as the "normal" activity) and evaluate it on the data, which includes another activity (considered as an "anomalous" activity) in addition to the "normal" activity. We assume that the anomalous movement activity shows a different pattern than the normal activity and should be distinguishable from the "normal" movement activity. Accurately detecting movement can have significant practical implications in various applications, such as fitness health tracking, where identifying irregular patterns or deviations from expected behavior is crucial. By focusing on such a complex and practical problem, we can demonstrate the effectiveness and robustness of our proposed approach in handling real-world challenges associated with PPG data, including inter-subject variability, noise, and other factors affecting signal quality.

BIOMETRIC IDENTIFICATION

In the context of biometric identification, we aim to identify an individual (user) as an anomaly when compared to a given group of people or another individual as the

"intended" user(s). We train the anomaly detector on the recorded data from the intended group or individual during a specific activity and evaluate the anomaly detector when presenting new data, which includes another individual (considered as an "anomaly") during the same activity as the data from the intended user(s). Identifying such anomalies can be crucial in personalized health monitoring systems, where it is important to distinguish between users for accurate and safe health monitoring and assessments.

4.3. EXPERIMENTAL SETUP

4.3.1. DATASETS

We use two datasets in our experiments. The first one is the Pulse Transit Time PPG (PTT-PPG) public dataset [28], a high-resolution and time-synchronized dataset annotated with activity labels. It contains waveform records from multi-wavelength sensors measuring PPGs, attachment pressures, and temperatures. The recordings are from 22 healthy subjects ($M = 22$) performing different physical activities in random order. We selected *Sitting* and *Walking* activities for this study. We use the green wavelength recorded PPG from the proximal phalanx (base segment) of the left index finger palmar side (Frequency of 500 Hz).

The second dataset is the PPG-Dalia public dataset collected by [29] to perform PPG-based heart rate estimation. It has recordings of 15 subjects ($M = 15$) performing different daily activities. We have selected *Sitting* and *Walking* activities for this study. We removed data from subject number 6 due to incomplete data recording. The signals are recorded with a frequency of 64 Hz.

4.3.2. DATA PREPROCESSING

A band-pass 2^{nd} order Butterworth filter is applied to the whole PPG signal of each subject individually for both datasets, but with different frequency ranges of 0.35 - 20 Hz for the PTT-PPG dataset and 0.1-10 Hz for PPG-Dalia. To create different categories of signals, we used Time reversal, Amplitude reversal, and both Time and Amplitude reversal augmentations. All of the signals are then normalized to zero mean and unit variance across the whole signal per subject. The final normalized filtered signals are segmented into windows with a length of 8 seconds, while two successive windows overlap by 7.5 seconds (this setting is common for PPG data [29–31]). Since the PTT-PPG dataset frequency is 500 Hz, the input windows are resampled using the Fourier method from a size of 4000 to a fixed size of 512, which allows for more efficient processing during model training, and it is the same input window size as the PPG-Dalia dataset.

4.3.3. IMPLEMENTATION

REPRESENTATION LEARNING

The hyperparameters and the architecture of the proposed deep learning model are determined by systematically searching through all possible combinations to obtain the best performance on the classification task using Leave-One-Subject-Out

cross-validation (LOSO). Eventually, we used a CNN architecture deep learning model consisting of a 1D convolutional neural network layer with a series of five-layer blocks followed by a fully connected layer and a final classification layer. The layer blocks are composed of two 1D convolutional layers, each followed by the Exponential Linear Unit (ELU) activation function and, in the end, a MaxPooling layer. After the final layer block, there is a fully connected layer with a size of 64, which is the learned representation size, followed by the Rectified Linear Unit (ReLU) activation function. Finally, there is a classification layer (SoftMax activation function) with a size of 4, corresponding to the four categories. The final implemented CNN model details are available in Appendix A.

The model is optimized using categorical cross-entropy as the loss function. The Adam optimizer is used with a learning rate of 0.00001 and a decay rate of 0.0001 for the PTT-PGG dataset and a learning rate of 0.0001 and a decay rate of 0.001 for the PPG-Dalia dataset. The batch size is 64, and training runs for 400 epochs for both datasets. To assess the randomness of the deep learning framework, each training process for each test subject is repeated five times. To evaluate the signal transformation classification performance, we use the Area under the ROC curve (AUC-ROC) metric.

ANOMALY DETECTION

In our PCA-based anomaly detection approach, we optimize the number of principal components by ensuring they cumulatively account for 99% of the data variance. The Isolation Forest model was implemented with 100 base estimators in the ensemble. The number of base estimators was chosen based on our preliminary experiments, which showed good performance in this setting. The Multivariate Normal Distribution-based anomaly detector was implemented utilizing a Gaussian Mixture Model with a single component. The parameters of this distribution, namely the mean vector and the covariance matrix, are learned directly from the data. In the evaluation phase, we assess the performance of our anomaly detectors by calculating the AUC-ROC.

4.3.4. ANOMALY DETECTION EVALUATION SCENARIOS

We consider two evaluation scenarios for anomaly detection tasks: Generalization and Personalization.

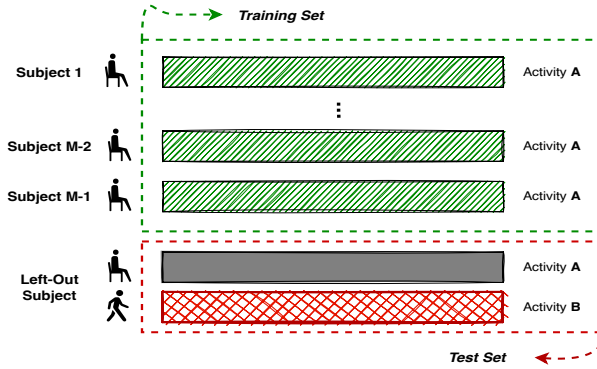
GENERALIZATION SCENARIO

In the generalization scenario, we aim to test the ability of the anomaly detection model to generalize across different individuals.

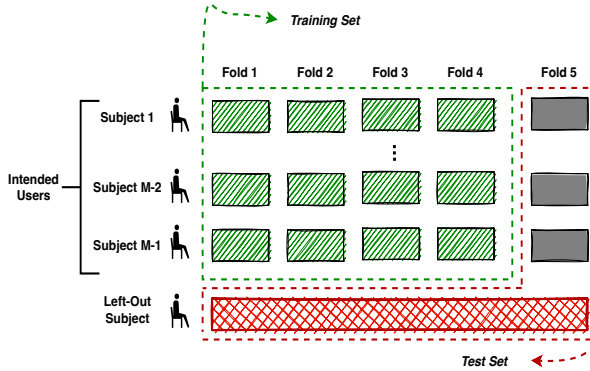
For activity movement detection, shown in Fig 4.2a, we train the model using data from all subjects performing *Sitting* activity as the main normal activity. This data is considered as 'normal training samples'. In the test phase, we introduce data from both a new activity, referred to as the 'anomalous activity' (in this case, *Walking*), and the main activity of a new subject (left-out) who was not part of the training data. We repeat this process for each subject, treating them as the test set (left-out subject),

using the LOSO setting. We then calculate the mean and standard deviation of the performance metrics across all test sets.

For biometric identification, shown in Fig 4.2b, we train the model on data from a group of subjects, who we refer to as the 'intended users'. This data forms our 'normal training samples'. We set aside 20% of the data from each subject for testing, using a 5-fold cross-validation approach. During the testing phase, we introduce 'anomalous data' from a new subject who is not part of the training data. This subject is referred to as the 'left-out' subject (user). To assess how well our model can differentiate the new user from the intended users, we use LOSO validation to treat each subject once as a 'left-out' user. We then calculate the mean and standard deviation of the performance metrics across all test sets.



(a)



(b)




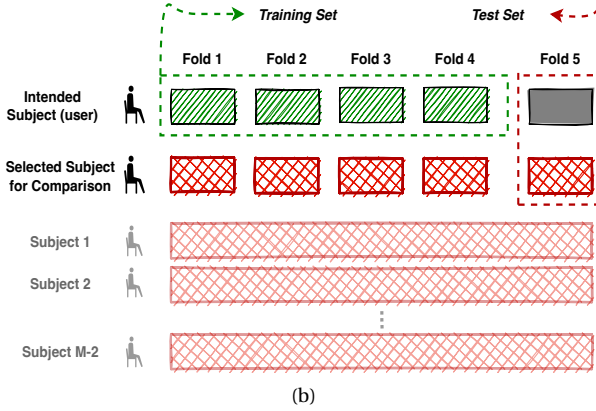
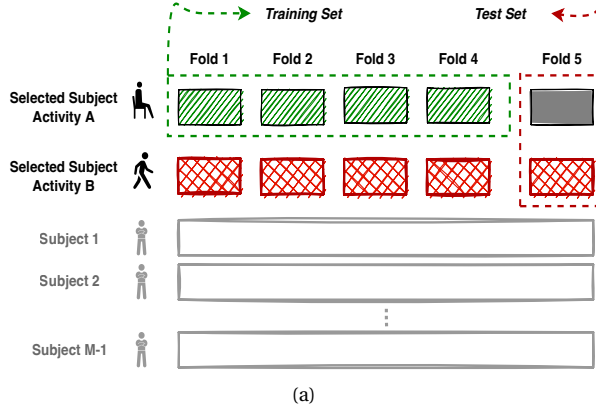
 Training samples (Normal),
  Test samples (Normal),
  Test samples (Anomaly)

Figure 4.2: Overview of Generalization Scenario: (a) Generalization in Movement Detection task, (b) Generalization in Biometric Identification task. Note that the distribution of the anomalous and normal samples in training and test sets follows the same ratios as depicted in the figures.



 Training samples (Normal),
  Test samples (Normal),
  Test samples (Anomaly)

Figure 4.3: Overview of Personalization Scenario a) Personalization in Movement Detection task. b) Personalization in Biometric Identification task. Note that the distribution of the anomalous and normal samples in training and test sets follows the same ratios as depicted in the figures.

PERSONALIZATION SCENARIO

In the personalization scenario, we aim to tailor the anomaly detection model to individual characteristics, both for movement detection and biometric identification. For activity movement detection, shown in Fig 4.3a, we select one subject and train the model on data related to the main activity, which is *Sitting*. This data forms our 'normal training samples'. We reserve 20% of this data for testing, using a 5-fold cross-validation approach. During the testing phase, we introduce 'anomalous activity data' from the same subject, in this case, *Walking* activity, and we use 20% of this data using 5-fold cross-validation for testing. Thereby, our test set includes 20% of 'Walking activity' data and 20% of the 'Sitting activity' data from one selected subject. We repeat this process for each subject. Finally, we calculate the mean performance and standard deviation across all test sets.

For biometric identification, shown in Fig 4.3b, we train the model using data from a single selected subject, who we refer to as the 'intended user'. This data forms our 'normal training samples'. We reserve 20% of this data for testing, using a 5-fold cross-validation method. During the testing phase, we introduce 'anomalous data' from a new subject who was not part of the training data, and we use 20% of its data using 5-fold cross-validation for testing. We compare each subject with the intended user in a pairwise manner. The average performance of these comparisons is taken as the performance of the intended user. We repeat this entire process for each individual, treating them as the intended user each time. Finally, we calculate the mean performance and standard deviation across all intended individuals.

4.4. RESULTS

4.4.1. REPRESENTATION LEARNING

The first step of the proposed framework is Representation Learning (see Fig 4.1). The overall performance of the signal transformation classification task for both datasets is calculated across all test subjects. Both datasets have a high mean AUC of 0.92 ± 0.09 for the PTT-PPG and 0.93 ± 0.06 for the PPG-Dalia. These results indicate that the model is able to generalize to new data (subject) and accurately classifies the augmented and original PPG segments in both datasets.

4.4.2. ANOMALY DETECTION

ACTIVITY MOVEMENT DETECTION

The second step of the proposed framework is Anomaly Detection (see Fig 4.1). Table 4.1 shows the results of movement detection for both datasets. In the generalized scenario, representation learning significantly improves the AUC performance for all three anomaly detection methods, suggesting its effectiveness in detecting anomalies in PPG data compared to the original data representation. For instance, the results of all anomaly detectors with PTT-PPG reveal that the AUC performance for anomaly detection barely reaches 0.5. However, the performance is increased towards 0.9 when using the learned representations.

Further, the original representation models demonstrate notable instability, indicative of high inter-subject variability. For example, the standard deviation in anomaly detectors without representation learning is around 0.4. However, employing representation learning substantially reduces this variability to below 0.1. These results underscore the capability of representation learning to facilitate better generalization across different individuals.

We also investigated the performance of the proposed methods in a personalized setting (Table 4.1). In the personalized setting, all methods show enhanced performance compared to the generalized scenario, with a moderate reduction in variability among individual performances. Notably, the integration of personalization with representation learning yields the most significant improvements. Here, the AUC performance approaches 1.0, and subject variability is markedly decreased to below 0.05. This demonstrates that the combined use of representation learning and

Table 4.1: *Mean Test AUC performance of Movement Detection in the Generalized and Personalized Scenario.* The 'RL-' prefix designates anomaly detectors that employ learned representations from Representation Learning (RL) instead of the original data representation.

Anomaly Detectors	Movement Detection AUC Performance			
	PTT-PPG Dataset		Dalia Dataset	
	Generalized	Personalized	Generalized	Personalized
MVN	0.40 ± 0.39	0.74 ± 0.29	0.78 ± 0.16	0.93 ± 0.08
RL-MVN	0.92 ± 0.09	0.98 ± 0.03	0.93 ± 0.10	0.97 ± 0.03
IF	0.28 ± 0.34	0.37 ± 0.31	0.56 ± 0.14	0.87 ± 0.15
RL-IF	0.91 ± 0.11	0.97 ± 0.05	0.88 ± 0.13	0.93 ± 0.03
PCA	0.44 ± 0.37	0.81 ± 0.25	0.76 ± 0.19	0.94 ± 0.07
RL-PCA	0.91 ± 0.09	0.97 ± 0.03	0.90 ± 0.18	0.95 ± 0.03

personalization not only improves performance but also ensures consistency across individuals, effectively capturing subject-specific characteristics of PPG data.

BIOMETRIC IDENTIFICATION

In light of the improved performance achieved through personalization in activity movement detection, biometric identification ensures that the detected anomalies are specific to the intended user. Table 4.2 shows the results of biometric identification in both generalized and personalized scenarios during *Sitting* Activity. In the generalized scenario, it can be observed that using representation learning is effective, and it improves the performance of all anomaly detectors across both datasets. While representation learning has been successful in improving performance, it still may not be perfect. This can be attributed to the inter-subject variability present in the data, as the model must distinguish between multiple people considered normal, which is challenging.

Considering the personalized scenario results, the performance of all methods is significantly higher compared to the generalized scenario, with substantially reduced variability: while in the generalized scenario, the standard deviations of the results are often around 0.2, in the personalized scenario, it is reduced to below 0.1. Moreover, representation learning continues to improve performance in the personalized setting, demonstrating the effectiveness of learned representations. These results indicate that minimizing inter-subject variability allows the model to better identify the anomalous person as it is easier to detect the anomalous individual from only one individual compared to a group.

4.4.3. ROBUSTNESS OF REPRESENTATION DIMENSIONALITY

One key aspect of our anomaly detection framework is the dimensionality of the learned representation, denoted as \mathbf{h}_i . Figure 4.4 illustrates the mean performance of anomaly detectors with varying \mathbf{h}_i dimensions ranging from 2 to 512 for both datasets in generalized and personalized scenarios.

Table 4.2: *Mean Test AUC performance of Biometric Identification in the Generalized and Personalized Scenarios.* Results are based on *Sitting* Activity. The 'RL-' prefix designates anomaly detectors that employ learned representations from Representation Learning (RL) instead of the original data representation.

Anomaly Detectors	Biometric Identification AUC Performance			
	PTT-PPG Dataset		Dalia Dataset	
	Generalized	Personalized	Generalized	Personalized
MVN	0.40 ± 0.26	0.76 ± 0.22	0.43 ± 0.26	0.56 ± 0.20
RL-MVN	0.60 ± 0.22	0.86 ± 0.08	0.55 ± 0.17	0.78 ± 0.09
IF	0.45 ± 0.36	0.58 ± 0.29	0.45 ± 0.29	0.56 ± 0.24
RL-IF	0.61 ± 0.24	0.86 ± 0.08	0.53 ± 0.18	0.74 ± 0.09
PCA	0.39 ± 0.34	0.67 ± 0.24	0.43 ± 0.26	0.55 ± 0.22
RL-PCA	0.59 ± 0.20	0.84 ± 0.09	0.55 ± 0.16	0.78 ± 0.09

As the \mathbf{h}_i dimensionality increases, the AUC also increases up to a certain point. This trend suggests that as the dimensionality rises, the representation captures more valuable information for anomaly detection. However, once we reach a certain dimensionality, further increases do not provide additional benefits, and the performance is stable.

In both scenarios, the learned representation improves the AUC compared to the original signal. Results show that using learned representation leads to better performance when the dimensionality of \mathbf{h}_i is reduced to extremely low levels. For instance, at the low dimensionality of 2 in PTT-PPG and 8 in the Dalia datasets, we can see the improvements in using learned representation over the original signal. Even when the dimensionality of the learned representation is the same as the original signal's dimensionality (512), it outperforms the original signal. This robust performance of the representation learning approach highlights its effectiveness in capturing the essential structure and patterns of the data and learning useful features across a wide range of low and high dimensionalities.

Choosing the right dimensionality depends on a balance between model performance and computational efficiency. Based on the results, the dimensionality of 64 for the PTT-PPG and 256 for the Dalia dataset seems to offer an ideal balance between computational efficiency and performance.

4.5. DISCUSSION AND CONCLUSION

This paper proposes a framework for anomaly detection in PPG data, consisting of two stages: representation learning and anomaly detection (Activity Movement Detection and Biometric Identification). We tested the ability of the proposed framework in generalized and personalized scenarios. Our research demonstrates that through representation learning and person-specific models, we can effectively address the key challenges in analyzing PPG signals, such as inter-subject variability, avoiding the influence of factors like color and skin thickness, weight, bone structure, etc. This

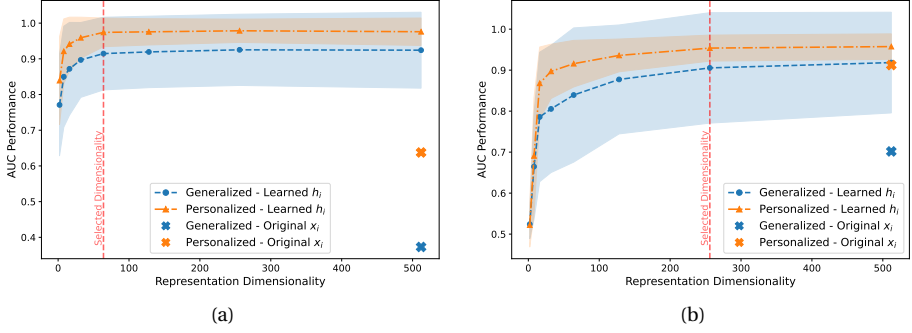


Figure 4.4: Overview of movement detection performance in the generalized and personalized scenarios for varying dimensions of the learned representation. (a) Results for the PTT-PPG Dataset and (b) results for the Dalia Dataset. The crosses indicate the performance obtained with the original representation.

significantly enhances the accuracy of anomaly detection, potentially allowing for the detection of rare and subtle anomalous patterns.

The results from the activity movement detection highlight the effectiveness of representation learning in improving AUC performance and decreasing inter-subject variability. However, it's important to note the variations in the extent of variability reduction between datasets. For instance, while representation learning significantly decreases variability in the PTT-PPG dataset, the reduction in the Dalia dataset is less pronounced. This difference could be attributed to various factors, such as the inherent complexities of each dataset, differences in signal quality, sensor types, environmental conditions during data collection, and participant demographics. Despite these challenges, representation learning not only consistently improved performance across datasets but also effectively reduced intra-subject variability. However, for further improvements, personalization combined with representation learning emerged as a critical factor. By customizing models to individuals, we achieve more consistent results and effectively capture subject-specific characteristics. These findings underscore the potential for further advancements in personalization to address the challenges in variability, especially in complex datasets and in building reliable and accurate PPG-based health monitoring systems. Note that the original representation AUC results, which are lower than 0.5, may indicate that flipping the label is actually beneficial. For example, it can be seen in Table 4.1 that an AUC of 0.28 in the generalized scenario from the PTT-PPG dataset would significantly improve to 0.72 by flipping, but it still remains worse than the 0.91 obtained by the RL representation. These results suggest that our approach may be beneficial in real-world applications.

In addition to the findings from the Activity Movement Detection task, our results in Biometric Identification further emphasize the key role of representation learning and personalization. Similar to the previous task, representation learning significantly enhanced the performance in both generalized and personalized scenarios for

biometric identification. Using personalization along with representation learning shows markedly higher performance and reduced variability. This improvement is attributed to the model's focus on individual characteristics. The consistency in these findings across both tasks underscores the efficacy of these methods separately and also in combination.

To further validate the effectiveness of our approach, we conducted a quantitative comparison with existing studies in PPG data for similar tasks and scenarios. For instance, in movement detection, a study [32], employing fully supervised learning, reported an AUC of 0.89 in personalized and 0.78 in generalized scenarios on the Dalia dataset (although in a more complex multi-class classification setting). Our unsupervised approach with representation learning achieved AUCs of up to 0.93 in the generalized scenario and 0.97 in the personalized scenario, suggesting higher overall performance. Similarly, for biometric identification, a supervised study [33] reported an AUC of 0.72 ± 0.14 in personalized scenarios (on a different dataset). Our framework, however, achieved a higher AUC and lower inter-subject variability, with an AUC of 0.86 ± 0.08 . Although these numbers cannot be fairly compared, the results underscore the potential of unsupervised learning methods in PPG anomaly detection, particularly when combined with representation learning and personalization.

Analyzing the robustness of representation dimensionality underscores its significance in anomaly detection frameworks. The performance of anomaly detection in relation to the dimensionality of the learned representation follows a pattern of initial gains followed by a plateau. This pattern suggests that while increasing dimensionality can enhance performance, there is a threshold beyond which additional increases do not yield further benefits. Interestingly, at the same dimensionality as the original signal, representation learning performs better. This suggests that learned representations can capture the (nonlinear) underlying patterns or structures in the data that may not be immediately apparent in the original signal. Furthermore, the fact that the learned representation can outperform the original signal even at extremely low dimensionalities signifies that representation learning can effectively extract and retain the most critical information from the original signal, thereby enhancing anomaly detection.

In all experiments of our framework, the performance difference between anomaly detection methods was relatively small. Therefore, we cannot draw a clear conclusion about which method performs better than the others overall. It seems that the crucial point in anomaly detection is not the method, but it is the representation and the personalization.

Despite the promising results in using representation learning and personalization, it is important to note that further research is needed to evaluate the effectiveness of RL on a wider range of different types of real-world anomalies in PPG. This is particularly important for practical applications, such as using smartwatches and self-monitoring for anomaly detection in healthcare, where the complexity and variability of real-world anomalies may be high. Exploring different algorithms or techniques to enhance the learned representations of PPG data can also be a future direction to further improve anomaly detection performance and decrease inter-subject variability.

In conclusion, our proposed framework provides a promising approach for different types of anomaly detection in PPG data. Combination of representation learning and personalization provides a more effective approach for developing reliable, robust, and accurate health monitoring systems.

REFERENCES

- [1] R. Ghorbani, M. J. Reinders, and D. M. Tax. “Personalized anomaly detection in PPG data using representation learning and biometric identification”. In: *Biomedical Signal Processing and Control* 94 (2024), p. 106216.
- [2] T. Aoyagi and K. Miyasaka. “Pulse oximetry: its invention, contribution to medicine, and future tasks”. In: *Anesthesia and analgesia* 94.1 (2002), S1–S3.
- [3] M. Boukhechba, L. Cai, C. Wu, and L. E. Barnes. “ActiPPG: using deep neural networks for activity recognition from wrist-worn photoplethysmography (PPG) sensors”. In: *Smart Health* 14 (2019), p. 100082.
- [4] Z. Zhang, Z. Pi, and B. Liu. “TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise”. In: *IEEE Transactions on biomedical engineering* 62.2 (2014), pp. 522–531.
- [5] S. Ghosh, A. Banerjee, N. Ray, P. W. Wood, P. Boulanger, and R. Padwal. “Continuous blood pressure prediction from pulse transit time using ECG and PPG signals”. In: *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*. IEEE. IEEE, 2016, pp. 188–191.
- [6] L. Everson, D. Biswas, M. Panwar, D. Rodopoulos, A. Acharyya, C. H. Kim, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte. “BiometricNet: Deep learning based biometric identification using wrist-worn PPG”. In: *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. IEEE, 2018, pp. 1–5.
- [7] M. K. Uçar, M. R. Bozkurt, C. Bilgin, and K. Polat. “Automatic sleep staging in obstructive sleep apnea patients using photoplethysmography, heart rate variability signal and machine learning techniques”. In: *Neural Computing and Applications* 29.8 (2018), pp. 1–16.
- [8] A. Aliamiri and Y. Shen. “Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor”. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. Las Vegas, NV, USA: IEEE, 2018, pp. 442–445.
- [9] R. Chalapathy and S. Chawla. “Deep learning for anomaly detection: A survey”. In: *arXiv preprint arXiv:1901.03407* (2019).
- [10] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly Detection: A Survey”. In: *ACM Comput. Surv.* 41.3 (July 2009). ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. URL: <https://doi.org/10.1145/1541880.1541882>.

- [11] M. Elgendi, R. Fletcher, Y. Liang, N. Howard, N. H. Lovell, D. Abbott, K. Lim, and R. Ward. “The use of photoplethysmography for assessing hypertension”. In: *NPJ digital medicine* 2.1 (2019), p. 60.
- [12] R. Kotorov, L. Chi, M. Shen, *et al.* “Personalized monitoring model for electrocardiogram signals: diagnostic accuracy study”. In: *JMIR Biomedical Engineering* 5.1 (2020), e24388.
- [13] Y. Bengio, A. Courville, and P. Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.
- [14] P. K. Gyawali. *Learning with Limited Labeled Data in Biomedical Domain by Disentanglement and Semi-Supervised Learning*. Rochester Institute of Technology, 2021.
- [15] N. Tishby, F. C. Pereira, and W. Bialek. “The information bottleneck method”. In: *arXiv preprint physics/0004057* (2000).
- [16] M. Tagliasacchi, B. Gfeller, F. d. C. Quiry, and D. Roblek. “Self-supervised audio representation learning for mobile devices”. In: *arXiv preprint arXiv:1905.11796* (2019).
- [17] M. Kocabas, S. Karagoz, and E. Akbas. “Self-supervised learning of 3d human pose using multi-view geometry”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1077–1086.
- [18] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. “Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4006–4015.
- [19] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. “Uncovering the structure of clinical EEG signals with self-supervised learning”. In: *Journal of Neural Engineering* 18.4 (2021), p. 046020.
- [20] X. Lan, D. Ng, S. Hong, and M. Feng. “Intra-inter subject self-supervised learning for multivariate cardiac signals”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 4. 2022, pp. 4532–4540.
- [21] O. R. A. Almanifi, I. M. Khairuddin, M. A. M. Razman, R. M. Musa, and A. P. A. Majeed. “Human activity recognition based on wrist PPG via the ensemble method”. In: *ICT Express* (2022).
- [22] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort. “Uncovering the structure of clinical EEG signals with self-supervised learning”. In: *Journal of Neural Engineering* 18.4 (2021), p. 046020.
- [23] A. Saeed, T. Ozcelebi, and J. Lukkien. “Multi-task self-supervised learning for human activity detection”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–30.
- [24] M. A. Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2 (1991), pp. 233–243.

- [25] P. Sarkar and A. Etemad. "Self-supervised ECG representation learning for emotion recognition". In: *IEEE Transactions on Affective Computing* 13.3 (2020), pp. 1541–1554.
- [26] D. M. J. Tax. "One-class classification: Concept learning in the absence of counter-examples." In: (2002).
- [27] F. T. Liu, K. M. Ting, and Z.-H. Zhou. "Isolation forest". In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.
- [28] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. "Deep PPG: Large-scale heart rate estimation with convolutional neural networks". In: *Sensors* 19.14 (2019), p. 3079.
- [29] M. S. Afzali Arani, D. E. Costa, and E. Shihab. "Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals". In: *Sensors* 21.21 (2021), p. 6997.
- [30] N. A. Sakr, M. Abu-Elkheir, A. Atwan, and H. Soliman. "Data driven recognition of interleaved and concurrent human activities with nonlinear characteristics". In: *Journal of Intelligent & Fuzzy Systems* 37.4 (2019), pp. 5573–5588.
- [31] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. "LIBLINEAR: A library for large linear classification". In: *the Journal of machine Learning research* 9 (2008), pp. 1871–1874.
- [32] M. S. Afzali Arani, D. E. Costa, and E. Shihab. "Human activity recognition: a comparative study to assess the contribution level of accelerometer, ECG, and PPG signals". In: *Sensors* 21.21 (2021), p. 6997.
- [33] D. Biswas, L. Everson, M. Liu, M. Panwar, B.-E. Verhoef, S. Patki, C. H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg, *et al.* "CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment". In: *IEEE transactions on biomedical circuits and systems* 13.2 (2019), pp. 282–291.

5

RESTAD: RECONSTRUCTION AND SIMILARITY BASED TRANSFORMER FOR TIME SERIES ANOMALY DETECTION

Anomaly detection in time series data is crucial across various domains. The scarcity of labeled data for such tasks has increased the attention towards unsupervised learning methods. These approaches, often relying solely on reconstruction error, typically fail to detect subtle anomalies in complex datasets. To address this, we introduce RESTAD, an adaptation of the Transformer model by incorporating a layer of Radial Basis Function (RBF) neurons within its architecture. This layer fits a non-parametric density in the latent representation, such that a high RBF output indicates similarity with predominantly normal training data. RESTAD integrates the RBF similarity scores with the reconstruction errors to increase sensitivity to anomalies. Our empirical evaluations demonstrate that RESTAD outperforms various established baselines across multiple benchmark datasets.

This chapter has been published as:

R. Ghorbani, M. J. T. Reinders and D. M. J. Tax, "RESTAD: Reconstruction and Similarity Based Transformer for Time Series Anomaly Detection," 2024 *IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, London, United Kingdom, 2024, pp. 1-6, doi: 10.1109/MLSP58920.2024.10734755. [1].

Code available at:

<https://github.com/Raminghorbani/RESTAD>

5.1. INTRODUCTION

Anomalies in time series data, i.e., unexpected patterns or deviations from normal behavior, can signify critical issues across various domains, from financial fraud to life-threatening health conditions. Hence, accurate anomaly detection is important. Given the rarity of anomalies and, thus, the lack of sufficient labeled data, fully supervised methods are less suited. Consequently, unsupervised learning methods have gained increasing attention [2]. These methods do not explicitly require labeled anomaly examples, making them ideal for the detection of unknown or unexpected anomalies [3].

Various classic unsupervised techniques like distance-based One-Class SVM (OC-SVM) [4] or density-based Local Outlier Factor (LOF) [5], have been widely used. However, they struggle with the temporal dependencies, high dimensionality, and complex generalization demands of time series data [6]. Recent developments in deep learning offer promising solutions for handling these challenges [7]. Architectures like Transformers and LSTMs excel at capturing temporal patterns and automatically learning hierarchical and non-linear features from time series data [8–10]. Building on these advancements, several effective anomaly detection methods have been developed, largely focusing on the reconstruction error as a primary anomaly criterion [11–13]. These methods typically assess the deviations between a given input and its reconstruction to identify anomalies. The underlying assumption is that typical data will have lower reconstruction errors, whereas anomalous data will exhibit higher errors due to the unfamiliarity of the model with these patterns [12, 14, 15].

A major issue of using the reconstruction error for anomaly detection is over-generalization [16]. Models fitted to capture the predominant patterns in the training data, generalize these patterns to include subtle variations as well. Therefore subtle anomalies can also be reconstructed well by these models. As a result, these anomalies are less distinguishable from typical patterns, reducing the model's detection sensitivity [17]. This effect is depicted in Figure 5.1.a, where the original signal includes a subtle anomaly at time point t_0 and a significant anomaly at t_1 . The

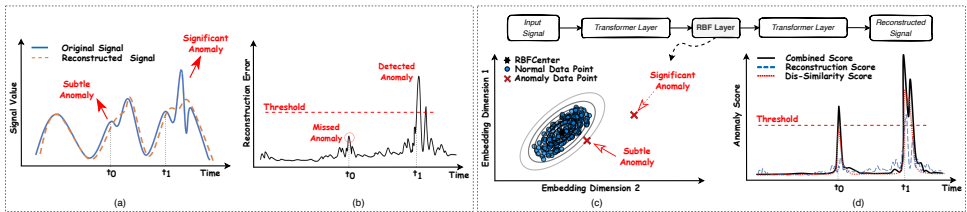


Figure 5.1: *Comparison of traditional reconstruction and RBF-enhanced anomaly detection:* a) Original signal with subtle and significant anomalies compared to its reconstruction. b) Reconstruction errors for the signals in (a), highlighting challenges in detecting subtle anomalies. c) Visualization of a model integrated with an RBF, shown via a 2D scatter plot that includes typical data, subtle and significant anomalies, and the RBF center with its influence radius, showing the RBF's ability to differentiate typical points from anomalies. d) Enhanced anomaly score using the RBF, which shows improved detection of subtle anomalies.

reconstructed signal is a slightly smoothed version of the original signal, and by using the reconstruction error alone, the subtle anomaly is missed as the reconstruction error remains below the detection threshold, as shown in Figure 5.1.b.

Efforts have been made to improve unsupervised anomaly detection by adding other types of scores to the conventional reconstruction error-based anomaly scores. For instance, AnomalyTrans [18] utilizes the concept of association discrepancy, which considers the similarity of a time point with its adjacent time points. It then reweights the reconstruction error accordingly to formulate the final composite anomaly score. However, in this method a normalization is performed, which can exaggerate the discrepancy scores for normal time points when no anomalies are present, potentially leading to false positives. This can misleadingly highlight normal data points as anomalies. Although this approach is effective for identifying clear outliers, it can inadvertently misrepresent subtle normal fluctuations as anomalies.

To overcome the challenges of scoring based on reconstruction error and the limitations of the association discrepancy method, we propose combining the reconstruction error with a specialized non-linear transformation like the Radial Basis Function (RBF) kernel [19]. The RBF kernel generates a similarity score that measures how close a data point is to a reference point or center, making it highly effective for anomaly detection. Anomalies, data points that deviate (are far away) from typical patterns, yield lower similarity scores with the RBF kernel, thus directly measuring how anomalous a point is. This score can effectively complement the reconstruction error and improve the sensitivity to subtle anomalies that might be overlooked by the reconstruction error. The effectiveness of combining RBF scores with reconstruction error is illustrated in Figure 5.1.c, where an RBF kernel is applied to typical data in the latent representation of a Transformer. By combining reconstruction error with RBF similarity scores, we create a comprehensive composite anomaly score that not only captures deviations from expected patterns but also ensures that subtle anomalies are still flagged. This composite anomaly score is shown in Figure 5.1.d, where the anomaly scores for both anomaly types are now above the detection threshold.

This paper presents an adaptation of the Transformer model, chosen for its ability to capture temporal dependencies in sequential data. By integrating the RBF neurons into the Transformer architecture, we develop a model that synergistically utilizes both similarity scores and reconstruction error to compute a distinctive anomaly score. Through an extensive evaluation, we show that this new REconstruction and Similarity based Transformer for time series Anomaly Detection, RESTAD, outperforms existing baselines across a range of benchmark datasets.

5.2. METHODOLOGY

Assume that the observed time series dataset consists of N sequences with length T . Each sequence in this dataset is denoted by $\mathcal{X}_i = \{\mathbf{x}_{i,t}\}_{t=1}^T$ where $\mathbf{x}_{i,t}$ represents the observed time point for i -th sequence at time t , having d dimensions, i.e., $\mathbf{x}_{i,t} \in \mathbb{R}^d$. Our task is to determine if a given $\mathbf{x}_{i,t}$ shows any anomalous behavior or not.

5.2.1. RESTAD FRAMEWORK

In our study, we incorporate the anomaly detection mechanism into the vanilla Transformer [9] through a specific layer of RBF neurons, see Figure 5.1.c. This RBF layer operates on the latent representations from the preceding layer, denoted by $\mathcal{H}_i = \{\mathbf{h}_{i,t}\}_{t=1}^T$, where $\mathbf{h}_{i,t} \in \mathbb{R}^{d_h}$. This layer computes the similarity of each data point $\mathbf{h}_{i,t}$ to a set of learnable reference points (centers), denoted by $\mathcal{C} = \{\mathbf{c}_m\}_{m=1}^M$, where $\mathbf{c}_m \in \mathbb{R}^{d_h}$. This computation results in the RBF output, $\mathcal{Z}_i = \{\mathbf{z}_{i,t}\}_{t=1}^T$, where $\mathbf{z}_{i,t} \in \mathbb{R}^M$, which then serves as the input to subsequent layer of the model. Specifically, the RBF similarity output for each data point relative to each center is defined by:

$$z_{i,t}^m(\mathbf{h}_{i,t}, \mathbf{c}_m) = \exp\left(-\frac{1}{2}e^\gamma \|\mathbf{h}_{i,t} - \mathbf{c}_m\|^2\right) \quad (5.1)$$

Here, the parameter γ controls the width of the RBF, influencing how it considers data points at varying distances from the center. This parameter is initialized and adjusted during training. Using the exponential of γ ensures the positivity of the scale parameter, simplifying the optimization process without enforcing a positivity constraint.

Anomaly Score: RESTAD is trained by minimizing the Mean Squared Error (MSE) to achieve accurate reconstruction. For anomaly detection, a composite anomaly score, $RESTAD_{score}$, is introduced by combining the normalized RBF similarity scores and reconstruction errors. The normalization is based on MinMax to ensure comparability. The RBF similarity score measures how closely $\mathbf{x}_{i,t}$ aligns with the learned centers. A higher similarity suggests normal behavior, whereas a lower similarity (or greater distance to the RBF centers) signals anomalies. This score is derived from averaging the RBF output $\mathbf{z}_{i,t}$ across all centers. The reconstruction error is the squared difference between the actual data $\mathbf{x}_{i,t}$ and its reconstruction $\hat{\mathbf{x}}_{i,t}$. The $RESTAD_{score}$ is formulated as:

$$RESTAD_{score}(\mathbf{x}_{i,t}) = \epsilon_r \times \epsilon_s \quad (5.2)$$

where $\epsilon_r = \|\mathbf{x}_{i,t} - \hat{\mathbf{x}}_{i,t}\|^2$ represents the reconstruction error, and $\epsilon_s = \left(1 - \frac{1}{M} \sum_{m=1}^M z_{i,t}^m\right)$ measures dissimilarity. This combination highlights subtle anomalies characterized by both low reconstruction errors and RBF scores, as well as significant anomalies with high reconstruction errors or low RBF scores.

Initialization of RBF Layer Parameters: Proper initialization of the RBF parameters, including the centers \mathbf{c} and scale parameter γ , is crucial for our methodology. We explore two initialization strategies: *Random* and *K-means*, to assess their impact on model performance. For *Random* initialization, parameters \mathbf{c} and γ are drawn from a normal distribution with zero mean and unit standard deviation. Although it is simple, it may lead to slower convergence, risk of local minima, and may not effectively represent the data distribution initially, possibly resulting in instability. In contrast, *K-means* initialization uses the inherent data structure for a more representative starting point. In this approach, initially, a base model (without the integrated RBF layer) is trained to minimize the MSE of reconstruction:

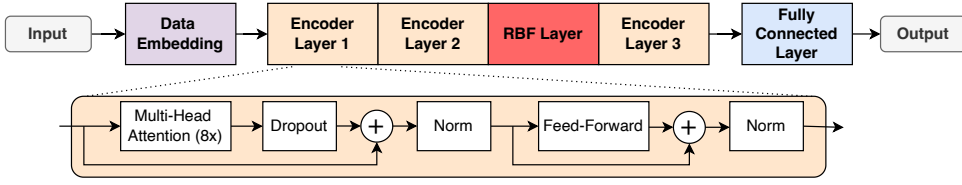


Figure 5.2: Overview of the proposed RESTAD model. Here, the RBF layer is added after the second encoder layer.

$$\mathcal{MSE} = \frac{1}{N} \sum_{i=1}^N \|\mathcal{X}_i - \hat{\mathcal{X}}_i\|_F^2 \quad (5.3)$$

After achieving satisfactory reconstruction accuracy from the base model, the latent representation is extracted from the specific layer where the RBF layer is intended to subsequently be integrated. This representation is then used to initialize \mathbf{c} via the K-means clustering algorithm. The scale parameter γ is initialized using $\tilde{\sigma}^2$, the mean squared distance from each data point to its nearest cluster center:

$$\tilde{\sigma}^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \min_m \|\mathbf{h}_{i,t} - \mathbf{c}_m\|^2, \quad \forall m \in [1, M] \quad (5.4)$$

Here, $\mathbf{h}_{i,t}$ denotes the latent representation vector of the i -th sample at the t -th time step, and \mathbf{c}_m is the m -th cluster center obtained from the K-means algorithm. This value, $\tilde{\sigma}^2$, is used to initialize γ as $\gamma = \frac{1}{\tilde{\sigma}^2}$, ensuring that the RBF function has a spread informed by the average dispersion of the data points around their respective centers.

5.3. EXPERIMENTAL SETUP

5.3.1. DATASETS AND PREPROCESSING

We use three public widely used benchmark datasets for our experiments: 1) Server Machine Dataset (SMD) [12], 2) Mars Science Laboratory (MSL) Rover [10], and 3) Pooled Server Metrics (PSM) [20]. Further information on each dataset is available in our code repository.

Data preprocessing involves normalizing each feature to zero mean and unit variance across the time dimension. Subsequently, the normalized signal is segmented into non-overlapped sliding windows [21] with a fixed length of 100 data points, a common setting based on previous related works [11, 18].

5.3.2. IMPLEMENTATION

RESTAD Model: The RESTAD model is an adaptation of a vanilla Transformer, incorporating an RBF kernel layer as detailed in Figure 5.2. It includes a DataEmbedding module that combines both token and positional embeddings, followed by an encoder with three layers. Each layer includes a multi-head

Table 5.1: *Performance metrics of baselines and RESTAD on test sets.* Initialization methods are denoted as (R) for *Random* and (K) for *K-means*. For all measures, a higher value indicates better anomaly detection performance.

Dataset	SMD					MSL					PSM				
	F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR	F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR	F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR
LSTM	0.12	0.74	0.17	0.79	0.20	0.06	0.56	0.14	0.63	0.19	0.11	0.73	0.50	0.72	0.51
USAD	0.13	0.63	0.11	0.72	0.14	0.06	0.53	0.14	0.59	0.18	0.07	0.60	0.41	0.61	0.43
PatchAD	0.01	0.50	0.04	0.61	0.08	0.03	0.50	0.10	0.57	0.15	0.02	0.50	0.28	0.55	0.33
Transformer	0.11	0.75	0.19	0.80	0.22	0.06	0.56	0.14	0.63	0.19	0.13	0.71	0.49	0.70	0.50
AnomalyTrans	0.03	0.49	0.04	0.50	0.07	0.02	0.49	0.10	0.52	0.14	0.02	0.51	0.30	0.53	0.34
DCDetector	0.01	0.50	0.04	0.51	0.08	0.02	0.50	0.11	0.58	0.15	0.02	0.50	0.28	0.52	0.32
RESTAD (R)	0.23	0.78	0.23	0.82	0.24	0.07	0.68	0.18	0.72	0.23	0.15	0.79	0.59	0.76	0.57
RESTAD (K)	0.20	0.79	0.24	0.83	0.25	0.07	0.66	0.18	0.71	0.23	0.14	0.79	0.57	0.76	0.56

self-attention mechanism and feed-forward networks. The model has a latent dimension of 32, an intermediate feed-forward network layer with a dimension of 128, and 8 attention heads. The RBF layer is placed after the second encoder layer (other placements are also possible, see section 5.4.1). Optimization is performed using the ADAM optimizer, and hyperparameters are determined through systematic search to optimize reconstruction task performance. Additional hyperparameter details are available in our code repository¹.

Evaluation: Anomaly scores (Eq. 5.2) exceeding a threshold δ are identified as anomalies. Performance is evaluated using the F1-score for threshold-dependent evaluation. Here, we follow [18] by setting δ to label a predefined proportion of data points as anomalies (0.5% for SMD, 1% for others). For threshold-independent analysis, we use AUC-ROC, AUC-PR, VUS-ROC, and VUS-PR metrics [22]. We exclude the point-adjustment method [23] due to its overestimation [24]. Our model is compared against baselines and state-of-the-arts models: LSTM [10], vanilla Transformer [18], USAD [14], PatchAD [13], AnomalyTrans [18], and DCdetector [11].

5.4. RESULTS

Our empirical results, as detailed in Table 5.1, highlight the effectiveness of the RESTAD for anomaly detection. RESTAD outperforms all baseline models across the benchmark datasets and evaluation metrics, regardless of the RBF initialization strategy. While there are slight performance differences between initialization methods, these variations are not significant enough to establish the superiority of one method over another.

To visually show detection differences, Figure 5.3 displays anomaly scores for a short segment of the SMD dataset. PatchAD, DCdetector, and AnomalyTrans models reveal many false detections, with DCdetector showing a pattern of repeated false positives and PatchAD resembling random scoring. LSTM, USAD, and Transformer models miss some anomalies or detect them weakly; for example, the first anomaly area is undetected by USAD, and only weakly detected by LSTM and Transformer. In

contrast, the RESTAD model demonstrates robust detection, effectively identifying all anomaly sections.

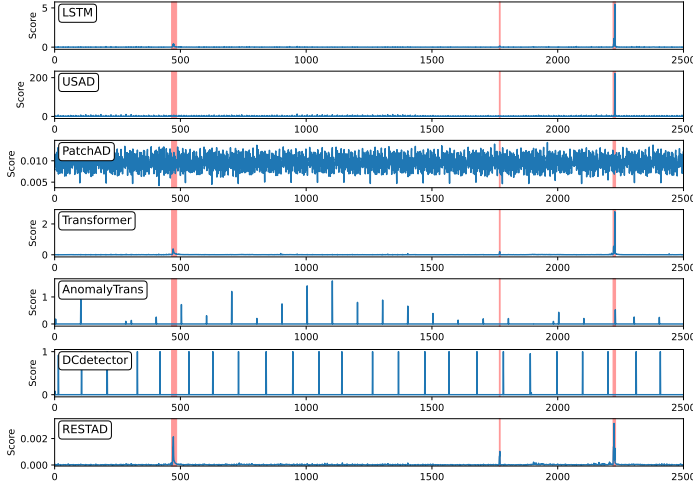


Figure 5.3: *Anomaly scores of different models for a segment of SMD dataset. The highlighted regions in red indicate the true anomaly periods (labeled by an expert).*

5

5.4.1. ABLATION ANALYSIS

The ablation experiments are based on the RBF layer with random initialization. This decision is based on our findings that random initialization is as effective as the K-means strategy (see Table 5.1), while offering greater simplicity and computational efficiency.

Anomaly Score Criterion: Table 5.2 highlights the impact of integrating the RBF score into anomaly detection. Multiplying the RBF layer’s dissimilarity score (ϵ_s) with the reconstruction error (ϵ_r) to form the composite anomaly score ($\epsilon_s \times \epsilon_r$) is found to be the most effective, consistently enhancing detection across all benchmarks and metrics. Adding the RBF layer to the vanilla Transformer with only reconstruction error ϵ_r as the anomaly score offers marginal improvements on some datasets. In contrast, using only the dissimilarity score (ϵ_s) or adding it directly to the reconstruction error ($\epsilon_s + \epsilon_r$) shows no significant benefits.

Figure 5.4 visually illustrates the superiority of our composite anomaly score over the traditional reconstruction score (ϵ_r) by showing subsets from all three datasets and the corresponding anomaly scores. Our anomaly score effectively identifies anomalies that are overlooked by the model relying solely on reconstruction error, with detections notably stronger and typically exceeding the threshold. Note that the thresholds depicted in the figures are the best optimized ones based on the entire dataset. Altering this threshold for the subset of data presented in the figures could diminish the overall performance and is therefore not possible.

Table 5.2: *Effect of integrating RBF layer and the choice of anomaly score.* For all measures, a higher value indicates better anomaly detection performance.

Architecture	Anomaly Criterion	SMD					MSL					PSM				
		F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR	F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR	F1-Score	AUC-ROC	AUC-PR	VUS-ROC	VUS-PR
Transformer	ϵ_r	0.11	0.75	0.19	0.80	0.22	0.06	0.56	0.14	0.63	0.19	0.13	0.71	0.49	0.70	0.50
RESTAD	ϵ_r	0.11	0.77	0.18	0.81	0.21	0.07	0.63	0.16	0.69	0.21	0.13	0.75	0.56	0.74	0.55
RESTAD	ϵ_s	0.01	0.44	0.03	0.52	0.07	0.01	0.43	0.08	0.48	0.12	0.01	0.32	0.20	0.37	0.25
RESTAD	$\epsilon_r + \epsilon_s$	0.04	0.57	0.06	0.60	0.10	0.07	0.61	0.16	0.65	0.20	0.01	0.68	0.49	0.59	0.45
RESTAD	$\epsilon_r \times \epsilon_s$	0.23	0.78	0.23	0.82	0.24	0.07	0.68	0.18	0.72	0.23	0.15	0.79	0.59	0.76	0.57

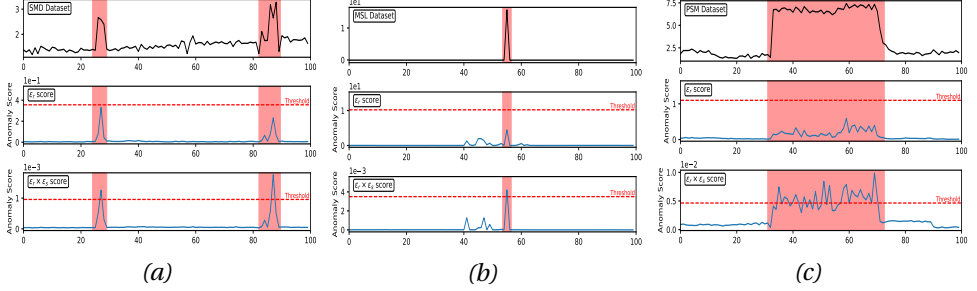


Figure 5.4: *Effect of our composite anomaly score ($\epsilon_r \times \epsilon_s$) compared to reconstruction error (ϵ_r) across segments of all datasets.* The highlighted regions in red indicate the true anomaly periods (labeled by an expert).

RBF Layer Placement: We explored the flexibility of RBF layer placement within the vanilla Transformer by integrating it after each of the three encoder layers. Figure 5.5a demonstrates that performance remains robust across all datasets, irrespective of the RBF layer’s location. Note that placing the RBF layer after the second encoder layer results in marginally better performance across all datasets. This slight advantage influenced our decision to position the RBF layer after the second layer in the final model architecture (see Figure 5.2).

Number of RBF Centers: Figure 5.5b represents the impact of the number of centers, ranging from 8 to 512, in the RBF layer of RESTAD. Results indicate that the optimal number of RBF centers is data-dependent. Additionally, beyond a certain threshold, increasing the number of centers does not enhance performance and may even reduce it.

5.5. DISCUSSION AND CONCLUSION

We introduced RESTAD, an adaptation of Transformers for unsupervised anomaly detection that improves on the limitations of using only reconstruction error as the anomaly score. By integrating an RBF layer into the Transformer, we combined RBF similarity scores with reconstruction error, enhancing the sensitivity to subtle anomalies. RESTAD consistently outperforms established baselines across various datasets and evaluation metrics.

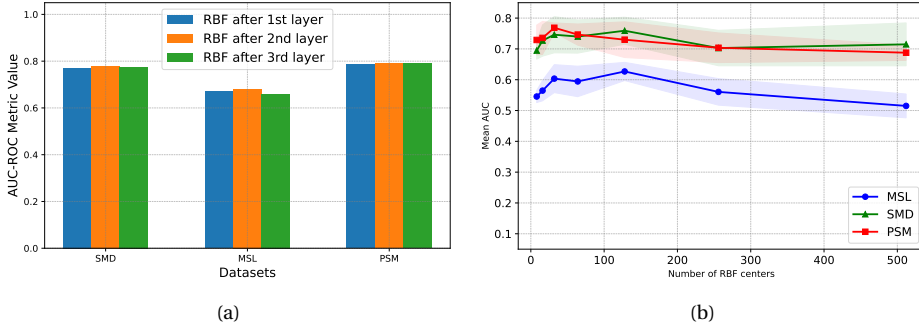


Figure 5.5: *Visualizations of RESTAD performance metrics: (a) with varying RBF layer placements and (b) across varying numbers of RBF centers. Here, shaded areas indicate \pm standard deviation, illustrating variability across multiple runs.*

Our findings reveal that RESTAD's performance is relatively invariant to RBF layer initialization methods, indicating robustness against initialization variability. The significant performance gains are primarily due to the multiplicative fusion of RBF similarity scores with reconstruction error, markedly improving anomaly detection capabilities. The RBF layer's placement within the architecture did not significantly affect performance, revealing architectural flexibility in integrating the RBF layer. However, the optimal number of RBF centers is data-dependent. These findings motivate future studies for the exploration of integrating RBF layers into other deep learning architectures for anomaly detection tasks.

REFERENCES

- [1] R. Ghorbani, M. J. Reinders, and D. M. Tax. “RESTAD: Reconstruction and Similarity Based Transformer for Time Series Anomaly Detection”. In: *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2024, pp. 1–6. DOI: 10.1109/MLSP58920.2024.10734755.
- [2] V. Chandola, A. Banerjee, and V. Kumar. “Anomaly detection: A survey”. In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.
- [3] R. Ghorbani, M. J. Reinders, and D. M. Tax. “Personalized anomaly detection in PPG data using representation learning and biometric identification”. In: *Biomedical Signal Processing and Control* 94 (2024), p. 106216.
- [4] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. “Support vector method for novelty detection”. In: *Advances in neural information processing systems* 12 (1999).
- [5] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. “LOF: identifying density-based local outliers”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.
- [6] N. Mejri, L. Lopez-Fuentes, K. Roy, P. Chernakov, E. Ghorbel, and D. Aouada. “Unsupervised Anomaly Detection in Time-series: An Extensive Evaluation and Analysis of State-of-the-art Methods”. In: *arXiv preprint arXiv:2212.03637* (2022).
- [7] K. Choi, J. Yi, C. Park, and S. Yoon. “Deep learning for anomaly detection in time-series data: review, analysis, and guidelines”. In: *IEEE Access* 9 (2021), pp. 120043–120065.
- [8] S. Tuli, G. Casale, and N. R. Jennings. “Tranad: Deep transformer networks for anomaly detection in multivariate time series data”. In: *arXiv preprint arXiv:2201.07284* (2022).
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [10] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom. “Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 387–395.
- [11] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun. “Dcdetector: Dual attention contrastive representation learning for time series anomaly detection”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 3033–3045.

- [12] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei. “Robust anomaly detection for multivariate time series through stochastic recurrent neural network”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2828–2837.
- [13] Z. Zhong, Z. Yu, Y. Yang, W. Wang, and K. Yang. “PatchAD: Patch-based MLP-Mixer for Time Series Anomaly Detection”. In: *arXiv preprint arXiv:2401.09793* (2024).
- [14] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga. “Usad: Unsupervised anomaly detection on multivariate time series”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 3395–3404.
- [15] D. Park, Y. Hoshi, and C. C. Kemp. “A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder”. In: *IEEE Robotics and Automation Letters* 3.3 (2018), pp. 1544–1551.
- [16] T. Zhao, L. Jin, X. Zhou, S. Li, S. Liu, and J. Zhu. “Unsupervised Anomaly Detection Approach Based on Adversarial Memory Autoencoders for Multivariate Time Series.” In: *Computers, Materials & Continua* 76.1 (2023).
- [17] H. Zhong, Y. Zhao, and C. G. Lim. “Abnormal State Detection using Memory-augmented Autoencoder technique in Frequency-Time Domain”. In: *KSII Transactions on Internet and Information Systems (TIIS)* 18.2 (2024), pp. 348–369.
- [18] J. Xu, H. Wu, J. Wang, and M. Long. “Anomaly transformer: Time series anomaly detection with association discrepancy”. In: *arXiv preprint arXiv:2110.02642* (2021).
- [19] M. J. Orr *et al.* *Introduction to radial basis function networks*. 1996.
- [20] A. Abdulaal, Z. Liu, and T. Lancewicki. “Practical approach to asynchronous multivariate time series anomaly detection and localization”. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021, pp. 2485–2494.
- [21] L. Shen, Z. Li, and J. Kwok. “Timeseries anomaly detection using temporal hierarchical one-class network”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13016–13026.
- [22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. “Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection”. In: *Proceedings of the VLDB Endowment* 15.11 (2022), pp. 2774–2787.
- [23] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, *et al.* “Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 187–196.

- [24] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon. “Towards a rigorous evaluation of time-series anomaly detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 7194–7201.

6

DISCUSSION AND CONCLUSION

We will discuss the contributions described in this thesis on three themes: 1) evaluation metrics, highlighting the development of Proximity-Aware Time Series Anomaly Evaluation (PATE); 2) representation learning, focusing on self-supervised learning approaches to address challenges such as label scarcity, high inter-subject variability, and long temporal dependencies; and 3) anomaly scoring mechanisms, centered on an innovative framework (RESTAD) for detecting subtle anomalies. Together, these themes provide a comprehensive overview of the thesis's contributions while identifying opportunities for future advancements.

6.1. EVALUATION METRICS FOR TIME SERIES ANOMALY DETECTION

Time series anomaly detection requires evaluation metrics capable of capturing temporal complexities, such as *Early* or *Delayed* detections, *Onset response time*, and *Coverage level* of detections over anomaly intervals. Existing metrics often fail to account for these aspects, leading to biased assessments that can misguide both researchers and experts. To address these shortcomings, Chapter 2 introduced PATE (Proximity-Aware Time Series Anomaly Evaluation), a novel metric that integrates proximity-based weighting with buffer zones around anomaly intervals and computes a weighted version of the Area Under Precision and Recall curve. Experimental results demonstrated PATE's ability to highlight meaningful performance differences across models and different scenarios.

PATE offers a more reliable benchmark for model evaluation, aligning results with user expectations and real-world requirements. This is particularly critical in high-stakes applications such as healthcare and finance, where misleading evaluations can delay responses or lead to costly errors. By addressing these shortcomings, PATE ensures that performance assessments are both practical and relevant, enabling the selection of models that are truly effective. For the research community, PATE's adoption reduces the risk of misaligned evaluation benchmarks, fostering innovation and guiding future advancements with meaningful and well-aligned evaluation criteria.

Despite its strengths, PATE has certain limitations. Its methodological complexity poses challenges for practitioners unfamiliar with advanced metric design, particularly

in non-academic or less technical environments. Interpreting PATE scores requires additional context or complementary metrics to define what constitutes a “good” performance, limiting its immediate practicality. Furthermore, PATE’s reliance on buffer zones and temporal proximity introduces subjectivity in parameter selection, complicating deployment across diverse applications. Although the method ensures consistent model rankings, customizing these settings for specific domains adds overhead. Lastly, evaluating multiple buffer settings and thresholds increases computational costs, making scalability a challenge, particularly for large datasets or systems prone to frequent false alarms.

Future work could focus on developing intuitive interfaces or tools that simplify PATE’s application and provide domain-specific default settings and evaluation criteria to streamline adoption across fields while maintaining flexibility for customization. In industrial contexts, domain experts could guide parameter adjustments to align PATE’s configurations with operational needs. Establishing standardized guidelines for interpreting PATE scores, including clear performance thresholds and visual diagnostics (e.g., visualizing the impacts of buffer zones and proximity weights), would improve accessibility and usability. Research into optimizing the evaluation process, such as pruning unnecessary configurations or parallelizing computations, could enhance scalability.

6

6.2. REPRESENTATION LEARNING FOR UNSUPERVISED TIME SERIES ANOMALY DETECTION

This thesis addresses the challenges of unsupervised time series anomaly detection through representation learning. Chapter 3 represented the first application of self-supervised learning (SSL) to photoplethysmogram (PPG) data in a label-scarce regime. Using reconstruction as a pretext task, the model learned low-dimensional, noise-robust representations from unlabeled data, improving downstream task performance and enabling the use of simpler models. However, reconstruction-based task struggled to learn invariant representations, limiting their ability to handle high inter-subject variability. This motivated the exploration of alternative pretext tasks.

Building on this, chapter 4 introduced an SSL approach using classification of augmented signal transformations as a pretext task. This method successfully learned robust and invariant representations, better addressing label scarcity and inter-subject variability. The chapter further demonstrated that personalization—tailoring models to individual users—significantly improved performance compared to universal models. These findings highlight the potential of SSL and personalized approaches to improve anomaly detection in real-world applications.

Despite the insights and advancements in representation learning, certain limitations remain. In Chapter 3, the representations learned through SSL were too complex for simple linear classifiers like Logistic Regression, suggesting that while they captured non-linear patterns, they did not simplify downstream task. Furthermore, the reliance on a single dataset and a specific task raises concerns about the generalizability of the framework. In Chapter 4, although the augmented transformation classification pretext task improved representation effectiveness, the reasons behind its success

remain unexplored. The study did not evaluate simpler designs, such as binary augmented transformation classification tasks, or the contribution of individual transformation classes. Additionally, reliance on Area Under the ROC curve for evaluation may introduce bias, especially in imbalanced datasets, highlighting the need for alternative metrics.

Optimizing pretext task design remains an open area of investigation, with promising directions including exploring additional augmentations such as Gaussian Amplitude Modulation, where the signal's amplitude is modulated by a Gaussian random factor. Our preliminary results suggest that such an augmentation can further improve downstream task performance; however, the model architecture may need to be adapted to perform well with the pretext task. Another idea is cut-paste augmentation, which introduces anomalous labels by rearranging parts of the signal. This binary classification pretext task, proven effective for images, could offer valuable insights into its applicability and impact on biomedical signals like PPG. Exploring multitask learning setups that combine multiple pretext tasks, such as reconstruction and classification, offers another promising avenue for optimizing representation learning. Additionally, examining the effects of binary classification for augmented transformations or testing various combinations of augmentations could clarify their individual contributions and further improve pretext task design. Studying the relationship between the dimensionality of learned representations and the design of pretext tasks models is another critical area. For example, investigating whether changes in model architecture, such as adding layers or altering complexity, affect downstream task performance while maintaining pretext task performance could enhance understanding of the representation learning process. Expanding evaluations to diverse datasets and downstream tasks is essential for validating the robustness and generalizability of these methods. Specifically, applying representation learning techniques to high-dimensional multivariate time series data could address the added complexity of handling multiple correlated signals, which is common in domains such as healthcare and industrial monitoring.

6.3. ANOMALY SCORING MECHANISM FOR SUBTLE TIME SERIES ANOMALIES

Detecting subtle anomalies in time series data is challenging, as traditional reconstruction-based methods often overgeneralize, missing small deviations from normal patterns. Chapter 5 introduces RESTAD (REconstruction and Similarity-based Transformer for Anomaly Detection), a framework that combines reconstruction errors with similarity-based scores. By leveraging a Radial Basis Function (RBF) layer within a Transformer architecture, RESTAD computes a composite anomaly score that enhances sensitivity to subtle deviations.

Benchmark evaluations demonstrate RESTAD's superior performance in detecting both subtle and prominent anomalies. This is particularly critical in high-stakes applications, where undetected anomalies can lead to severe consequences. RESTAD addresses the limitations of traditional reconstruction-based models and offers a practical foundation for more sensitive and reliable anomaly detection systems.

Despite its strong performance, RESTAD has limitations. The number of RBF centers is a sensitive hyperparameter that requires careful dataset-specific tuning. Beyond a certain point, increasing the centers does not improve performance and may degrade it, complicating deployment. While RESTAD integrates RBF layers within a Transformer architecture, its applicability to other architectures, such as LSTMs or CNNs, remains unexplored.

Future research could explore applying RESTAD's scoring mechanism to other architectures to broaden its utility across diverse modeling paradigms. Our preliminary experiments integrating the RBF layer into LSTMs have shown promising results, suggesting that this approach could enhance performance and encourage further investigation. Another promising direction involves refining the training process of the RBF layer by maximizing the likelihood under the RBF kernels while ensuring accurate data reconstruction. This may include designing a regularization term to prevent any single RBF center from disproportionately influencing the output, encouraging an effective distribution of centers that promotes high likelihood for the training data and enhances the RBF layer's representational efficiency.

6.4. FINAL WORDS

This thesis advances the field of time series anomaly detection by addressing key challenges through innovative methodologies and practical solutions. The contributions span the development of robust evaluation metrics, the application of representation learning for unsupervised anomaly detection, and novel scoring mechanisms for subtle anomalies. Together, these efforts lay the groundwork for more reliable, scalable, and generalizable anomaly detection systems.

At its core, this work emphasizes the importance of aligning research priorities with real-world challenges. By bridging theoretical innovation with practical applicability, it highlights how robust evaluation, representation learning, and anomaly scoring can redefine state-of-the-art practices. These contributions not only address current gaps but also set a trajectory for future advancements in anomaly detection research and applications.

Looking ahead, the challenges identified throughout this thesis provide a roadmap for continued exploration. This research serves as both a foundation and an invitation for the community to build upon, fostering innovation in a field where precision and reliability are critical.

ACKNOWLEDGEMENTS

Writing this thesis has been a long journey—one filled with discovery, doubt, growth, and many unforgettable moments. But above all, it was never something I could have done alone. I'd like to thank the people who made this possible, each in their own way.

I'd like to start by thanking my supervisor, **David**. From day one, you showed me what it truly means to be a researcher—how to ask questions that spark real curiosity, how to dive into them with passion, and how to stay open to the unexpected turns along the way. Your insistence on clarity and directness in my writing has made me a far better communicator. I'm deeply grateful for your steady support—whether we were celebrating a breakthrough or recovering from a setback.

I'm also sincerely thankful to **Marcel**, whose hands-on approach turned every challenge into an opportunity to grow. You taught me how to manage my time, sharpen my thinking, and tackle problems head-on. Your guidance, combined with David's, was the perfect balance—and I feel incredibly fortunate to have been mentored by you both.

I would also like to thank the people I worked alongside each day. I feel incredibly lucky to have been part of such a smart, kind, and supportive group of colleagues. **Tom**, working on interviews with you was genuinely fun—I think we made a great team. You always brought good energy, kept things light, and, of course, kept the BBQ tradition alive—thanks for that! **Jesse**, thanks for patiently answering all my causal inference questions and for the rides to Rotterdam. I truly enjoyed our conversations. **Hayley**, playing violin with you during the retreat is still one of my favorite memories. I'm grateful for your support and the introductions to industry. **Xucong**, your feedback on my resume and career advice were incredibly helpful—thank you. **Jan**, I've always admired your depth and practicality in research. **Chirag**, our time at the Gran Canaria summer school was a real highlight. Thanks for all the thoughtful advice over the years. **Bernd**, our collaboration was short but packed with learning. I'm glad we had the chance to work together—I learned a lot from your sharp thinking.

Mahdi, being office mates with you was one of the highlights of my PhD. We shared so many ideas, great conversations, and those tea breaks—still some of my favourite memories. You always liked to be called “The Legend,” and rightly so. Our talks—about research, politics, life—always left me with deeper questions, especially *why*. **Aurora**, thank you for organizing so many great activities and for the fun chats. **Alejandro**, thank you for all the great conversations—and I hope you're done with the endless MacBook drama! **Hesam**, our coffee chats were always great, and I still laugh thinking about the time we accidentally shot our teammate in paintball. Great memories! **Taylan**, your quick hellos and warm energy always brightened the day. **Rickard**, the tallest Swede in the Netherlands—thank you for your sharp takes on causality and career advice. **Ojas**, it's funny we first met at the gym, then connected more on the way to Rotterdam. Your drive and focus always stood out. **Thomas**, such a fun person to be around. I always

enjoyed joking with you and chatting. **Sayak**, thanks for your help with the UK visa and for the great summer school memories. Always enjoyed our chats—good luck in your new position! **Chengming**, hope you finally managed to collect your data—and that you're still enjoying the Mazda! **Alireza**, thanks for your STRAP collaboration and all the coding support. Working with you was always a pleasure. **Niels**, exploring PPG data with you—and our rides to Eindhoven—were a highlight of the PhD. I really valued our discussions.

And to all the others who made this experience so much richer—**Robert-Jan, Attila, Ombretta, Marco, Hadi, Osman, Jim, Nargis, Casper, Seyran, Ziqi, Tiffany, Amelia, Olaf, Sander, Xiangwei, Zhi-Yi, Jing, Gijs, Cheng, Yavuz, Myrthe, Chenxu, Vandana, Zonghuan, Ramin, Stavros, Mostafa, Yasin, Azza, Panos, Eelko, Eric, Pieter, Xi, Jhon, Rosalinde**—thank you all for being such amazing colleagues. I've truly appreciated your company, support, and all the moments we shared.

A big thank you to **Ruud** and **Bart**, our behind-the-scenes tech magicians—always ready to jump in whenever something broke or refused to work. And to **Saskia** and **Marunka**—the absolute best when it came to handling paperwork. From travel declarations to last-minute forms and scheduling, your quick responses and kind support made my life so much easier. Thank you for being so reliable, efficient, and thoughtful every step of the way.

Outside of work, I'm deeply grateful to the friends who made this journey lighter, brighter, and far more meaningful. **Morteza**, you've been a true friend from the very beginning. We faced the challenges of migration and starting over together, learned Dutch side by side, collaborated on papers, and shared countless conversations that helped me grow. I truly value the friendship we've built over the years. **Farhad**, even though we've only met a few times in person, your support has meant a great deal. Whether it was advice, encouragement, or just being there to talk, you've been someone I could always count on—throughout my life and especially during this PhD. Thank you. **Masoud**, my friend from Tehran to Rotterdam—you've been a constant source of comfort and joy. Talking to you always lifts my mood. Thank you for being such a generous and supportive presence—pure gold. **Miad**, I'm so glad we met that day at the gym. From that point on, we've had deep, honest conversations—about PhD life, the future, and everything in between. I'm thankful for your friendship. **Tessa** and **Jeroen**, seeing you always lifted my spirits. I loved every moment we shared—especially our trip to Germany. **Hamid**, even though we live far apart, every conversation with you has been uplifting. Your support has meant a lot to me, and I truly value our friendship. And to my other wonderful friends—**Parastou, Beheshte, Fatemeh, Kamran, Sepehr**—thank you for your kindness, warmth, and company. I've truly appreciated every moment we've spent together.

I'm also deeply grateful for the love and support of my family—both the one I was born into, and the one I found here. To **Theo, Barbara, Sanne**, and **Steven**—thank you for making me feel at home in a place far from home. You've extended your warmth, your care, and your family to me, and I've never taken that for granted. **Theo**, thank you for always being kind to me, and offering support whenever I needed it. I'm truly grateful for everything you've done. You're the most knowledgeable person I've ever met, and I admire how you're always learning and sharing more. It's marvellous. That said, I still have to disagree with the whole “7% rain” theory—we really need to settle that, once and

for all! **Barbara**, I can't express how much your support has meant to me. Thank you for sending me amazing food and teaching me how to pronounce Dutch properly—thanks to you, I finally know how to say *Zwolle*! **Sanne**, you're always full of energy and so kind to me. I'll never forget my first truly practical Dutch word—*Plint*—which came from you! Thank you for your support and for always making me feel welcome. **Steven**, it's always a pleasure talking to you—about cars, sheep, pigs, games, or whatever else comes up. Visiting the car museum together was a real highlight—I hope we do more of that! And thanks for your ongoing faith in my Dutch-learning journey—even if I've promised to start “next Monday” for years!

And to my own family—thank you for being the foundation of everything I've built. **Atefeh**, my most precious person—my mom. I promised I would call you every day, and I did my best. Even though you weren't physically here, I always felt your love and support. Everything I've achieved rests on the foundation of your care, strength, and unwavering belief in me. Thank you for being my greatest source of love and encouragement. **Reza**, my dad, my hero. I've learned so much from you. Everything I am and everything I've accomplished is rooted in your dedication, quiet strength, and constant support. You've always had my back—even from afar. Thank you for always caring, in your own thoughtful way. **Amin**, you've always meant more to me than just being a brother. You've been there to listen, to talk, and to help—no matter how busy you were. Thank you for keeping me connected to home and for handling everything so I could focus here without worry. I'll never forget what you've done for me. And yes—start training for FIFA, because next time, I'm showing you who's boss.

And finally, to **Marieke**, my love and the person who means the most to me—meeting you during my PhD has been my greatest gift. Your support, love, and belief in me carried me through this entire journey. You were there through all of it: the failures, the successes, and the moments I wanted to give up. You helped me to keep going. I couldn't have done this without you—and I wouldn't have wanted to. I'm endlessly grateful for your love, your patience, and for always being there.

And to everyone else I crossed paths with along the way—there are surely more of you who deserve to be mentioned. If you're reading this, know that I'm grateful—you were part of this journey too.

*Ramin Ghorbani
Delft, December 2024*

CURRICULUM VITÆ

Ramin GHORBANI

15-02-1995 Born in Esfahan, Iran.

EDUCATION

- 2020–2024 **PhD in Computer Science**
Delft University of Technology, Delft, Netherlands
Thesis: Unmasking the Unexpected | Towards Reliable Time
Series Anomaly Detection
Promoters: Prof. dr. ir. M.J.T. Reinders and Dr. D.M.J. Tax
- 2017–2019 **Master of Science in Industrial Engineering - System Optimization**
Iran University of Science and Technology, Tehran, Iran
- 2013–2017 **Bachelor of Science in Industrial Engineering**
Yazd University of Science and Technology, Yazd, Iran

WORK EXPERIENCE

- 2025–current **ABN AMRO Bank N.V., Amsterdam**
Machine Learning Engineer
- 2020–2024 **Erasmus Medical Center, Rotterdam**
Machine Learning Scientist

LIST OF PUBLICATIONS

In this thesis:

1. R. Ghorbani, M. J. Reinders, and D. M. Tax. "PATE: Proximity-Aware Time Series Anomaly Evaluation". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 872–883
2. R. Ghorbani, M. J. Reinders, and D. M. Tax. "Self-supervised ppg representation learning shows high inter-subject variability". In: *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*. 2023, pp. 127–132
3. R. Ghorbani, M. J. Reinders, and D. M. Tax. "Personalized anomaly detection in PPG data using representation learning and biometric identification". In: *Biomedical Signal Processing and Control* 94 (2024), p. 106216
4. R. Ghorbani, M. J. Reinders, and D. M. Tax. "RESTAD: Reconstruction and Similarity Based Transformer for Time Series Anomaly Detection". In: *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2024, pp. 1–6. DOI: 10.1109/MLSP58920.2024.10734755

Other publications:

1. M. Moradi, R. Ghorbani, S. Sfarra, D. M. Tax, and D. Zarouchas. "A spatiotemporal deep neural network useful for defect identification and reconstruction of artworks using infrared thermography". In: *Sensors* 22.23 (2022), p. 9361
2. R. Ghorbani and R. Ghousi. "Comparing different resampling methods in predicting students' performance using machine learning techniques". In: *IEEE access* 8 (2020), pp. 67899–67911
3. R. Ghorbani, R. Ghousi, A. Makui, and A. Atashi. "A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset". In: *IEEE Access* 8 (2020), pp. 141066–141079
4. R. Ghorbani and R. Ghousi. "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction". In: *International Journal of Data and Network Science* 3.2 (2019), pp. 47–70
5. R. Ghorbani, R. Ghousi, and A. Makui. "Location of compressed natural gas stations using multi-objective flow refueling location model in the two-way highways: A case study in Iran". In: *Journal of Industrial and Systems Engineering* 12.2 (2019), pp. 95–112