EXPLAINABLE ANOMALY DETECTION IN SENSOR-BASED REMOTE HEALTHCARE MONITORING WITH ADAPTIVE TEMPORAL CONTRAST

Nivedita Bijlani
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
n.bijlani@surrey.ac.uk

Gustavo Carneiro
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
g.carneiro@surrey.ac.uk

Payam Barnaghi
UK Dementia Research Institute Care Research and Technology Centre
Imperial College
London, United Kingdom
p.barnaghi@imperial.ac.uk

Samaneh Kouchaki
Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom
s.kouchaki@surrey.ac.uk

ABSTRACT

Sensor-based remote healthcare monitoring can be used for the timely detection of adverse health events in people living with long-term health conditions, and help reduce preventable hospitalization. Current anomaly detection approaches in a real-world setting are challenged by noisy data and unreliable event annotation. Inspired by the conceptual simplicity and recent applications of negative sample-free contrastive learning in computer vision, we propose a lightweight, self-supervised model to extract noise-adaptive representations from multidimensional sensor data. We use the contrastive loss between the more granular observation data and a corresponding learnable, lower temporal resolution augmentation, and use the learned representations for anomaly detection. Learning to adjust this "contrast factor" enables the model to identify and leverage the most informative temporal features at different scales, enhancing its ability to discern underlying patterns amidst noise. Our model outperformed comparable representation learning algorithms in detecting agitation and fall events across three distinct participant cohorts in a real-world study of people living with dementia in their homes. We further used the representations to create a "spatiotemporal attention map" to focus on the source of anomaly and offer explainability. Our approach is domain-agnostic and can be used in wider healthcare, industrial and urban sensor settings.

1 INTRODUCTION

Insights from sensor-based remote health monitoring data can be used to analyze temporal patterns and detect adverse conditions, with minimal intrusion and low cost. A real-world monitoring setting, however, poses a unique set of challenges. It is characterized by absence of reliable labeling (annotation can be resource-intensive or self-reported), data drift, noise and lack of regular patterns. An anomaly detection (AD) algorithm in this context must address these issues while also achieving high sensitivity, low alert rate, and explainability to a monitoring team. To achieve this, it is crucial to obtain discriminative, noise-resilient representations from sensor data.

A variety of deep learning approaches have been used for multivariate time series representation learning in unsupervised settings. Autoencoders and temporal convolutional networks are used in Franceschi et al. (2019) to capture patterns and dynamics inherent in time series data via triplet loss. Ma et al. (2019) integrate the temporal reconstruction and K-means objective to generate cluster-specific temporal representations. Some works have applied graph representation learning to time series data (Zheng et al., 2019; Zhao et al., 2019; Li & Jung, 2021; Bijlani et al., 2022). Recently, Li et al. (2023) used vision transformers on line graph images generated from irregularly sampled time series for time series classification. However, ViTs and Swin Transformers are inherently complex models, making them prone to overfitting and impacting their interpretability.

In recent years, the computer vision field has pioneered advancements in negative sample-free contrastive learning for obtaining high-quality representations of images using different augmentations of the underlying data, simplifying the learning process, reducing computational overhead, and achieving robust benchmark performances in downstream tasks. Negative sample-free contrastive learning is now being applied in various domains including NLP and audio processing (Niizumi et al., 2021; Elbanna et al., 2022; Sarkar et al., 2023; Xu et al., 2024).

In this work, we propose a lightweight, self-supervised model to extract noise-adaptive, discriminative representations from multidimensional data collected via real-world sensor-based remote health monitoring. Our key contribution is leveraging vision-inspired contrastive loss together with a learnable temporal augmentation to extract noise-adaptive embeddings directly from sensor data. We demonstrate the efficacy of this approach in a real-world study of 65 home-living people with dementia, achieving 84% average recall and 92% generalizability in detecting adverse events. In addition, we demonstrate how the model can be used to gain insight into the source of anomaly by combining temporal attention with sensor importance to create a spatiotemporal attention map. Our work could open up new possibilities to utilize computer vision techniques in sensor-based monitoring and the wider time series domain in a data-driven way.

2 METHOD

Model overview. Figure 1 illustrates our contrastive learning model. The raw sensor data is preprocessed and input into a Transformer encoder network. The Gumbel-Softmax distribution (Jang et al., 2016) performs a soft selection over a predefined set of "contrast factors", which refers to the granularity of temporal aggregation, e.g. 3, 4, 6 hours. The contrast factor is used to dynamically adjust the temporal aggregation of input features, influencing the embeddings and the model's sensitivity to anomalies. In this way, the model learns the optimal parameters for feature extraction, together with the optimal temporal resolution for processing the input data. The original data is downsampled to the learned resolution, and the two versions are fed to identical Transformer encoders. The network is trained using a self-supervised vision-inspired contrastive loss, e.g. DINO (Caron et al., 2021), SimSiam (Chen & He, 2021), BYOL (Grill et al., 2020) or Barlow Twins (Zbontar et al., 2021).

Learned augmentation. In order to mitigate the need for prior understanding of the data, we integrate the learning of augmentation into the training pipeline. By learning to adjust the contrast factor dynamically, the model can identify and leverage the most informative temporal features at different scales, enhancing its ability to discern discriminative patterns amidst noise. In computer vision, there exist commonly used operations, such as rotation, cutout, flip or crop that can be learned via AutoAugment (Cubuk et al., 2019) and related strategies (Hataya et al., 2020; Zheng et al., 2022), but there are no standard augmentations for sensor-received data. We use the Gumbel-Softmax distribution (Jang et al., 2016) to select contrast factors for data augmentation. For each contrast factor, we associate a learnable logit, representing the unnormalized log probabilities of selecting that particular factor. We apply the Gumbel-Softmax operation to these logits, incorporating Gumbel noise to simulate the sampling process. This operation is parameterized by a temperature τ, which controls the smoothness of the approximation to the discrete distribution. Its output is a differentiable approximation of a one-hot encoded vector, indicating the selected contrast factor.

Transformer encoder. The Transformer encoder, proposed in Vaswani et al. (2017), consists of identical layers of multi-head self-attention and a position-wise fully connected feed-forward network. The Transformer encoder produces a representation that assigns different weights to different temporal segments, which is particularly useful for continuous health monitoring as unusual activity in certain parts of the day can be more significant in informing anomaly detection than activity

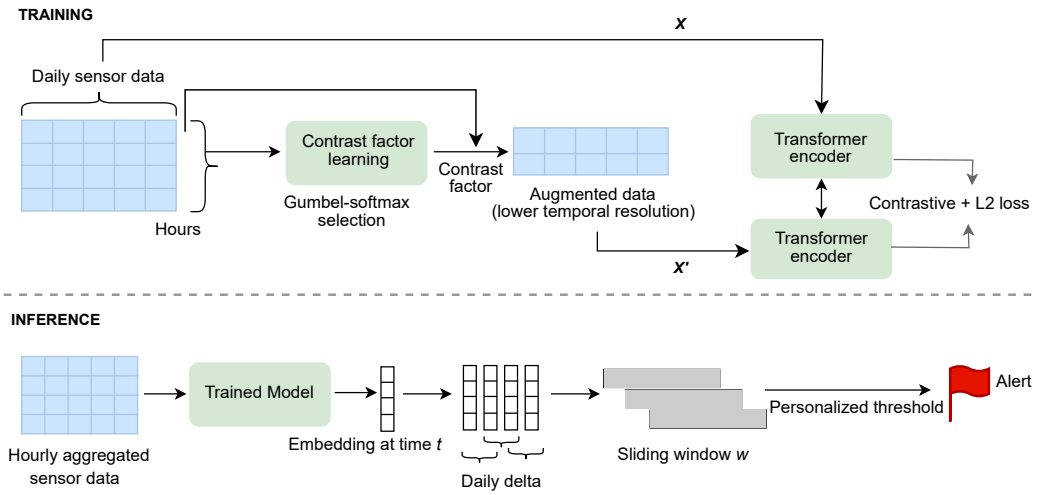

Figure 1: Contrastive learning-based anomaly detection pipeline. The model takes as input hourly-aggregated daily sensor data $\boldsymbol{X}$. The Gumbel-Softmax distribution is used to perform a soft selection over "contrast factors" for adaptive learning of the optimal temporal aggregation. The original data is accordingly downsampled and augmented to $\boldsymbol{X}'$. The network is trained using a vision-inspired contrastive loss, and L2 regularization. During inference, the daily delta of embeddings, or anomaly score, is conditioned by sliding window thresholding using the personalized threshold to raise alerts.

mixed with noise at other times. Our model uses a single layer Transformer encoder with two attention heads (Figure 2).

Contrastive loss. The network is trained via a vision-inspired contrastive loss, e.g. DINO, with optional regularization. The total loss is given by $L_{\text{total}} = L_{\text{DINO}} + \lambda||\mathbf{w}||_2^2$, where L_{DINO} is the DINO loss, the second term represents L2 regularization, and λ is the loss-balancing hyperparameter.

Inference. The raw daily sensor data are preprocessed and translated to a corresponding day-level embedding. Then, the cosine difference between successive daily embeddings, the "anomaly score" is calculated. It quantifies the consistency in day-to-day activity. The personalized alerting threshold ρ is computed for each individual based on their 7-day moving average and standard deviation of anomaly scores, and the target alert rate. Finally, an alert is raised if the anomaly score lies above ρ.

Explainability. We use the attention weights from the Transformer encoder to discover the most influential hour indicative of anomaly on a flagged day. We utilize Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) to find the most important sensor that contributed to the anomaly. To embed LRP within our Transformer architecture, we make two adaptations. The feedforward layers in the Transformer encoder and projection layers are re-engineered to compute the relevance scores, and the calculation of LRP scores for the attention and LayerNorm components of the Transformer is refined to ensure accurate feature attribution, in line with Ali et al. (2022). Finally, we merge the temporal attention weights with the global relevance score, to obtain a "spatiotemporal attention map". The spatiotemporal attention map visualizes the weights assigned by the model to different spatial locations and temporal intervals the model deems most influential for making predictions. It highlights areas within the multivariate sensor data from the participant's household that significantly deviate from typical patterns, thereby pinpointing specific times and sensors that contributed to the anomaly decision for that day. This allows for a targeted investigation into the participant's activity and health status, facilitating early intervention by the monitoring team.

3 Experiments and Results

Data. This work utilizes data from an active remote health monitoring program collected between August 2019 and April 2022, for 80 home-living individuals with dementia or mild cognitive im-

pairment. Participants, who consented in writing, have passive infrared sensors installed in the hallway, bathroom, bedroom, lounge and kitchen to track daily movement. The clinical team labels adverse health events such as falls, UTI (urinary tract infection) or agitation episodes via preset alerts, weekly check-in, or participant self-reports, within a 7-14 day window. We aggregated data for each household location hourly, log-transformed and normalized it to enable cross-sensor and cross-participant standardization. Multiple signals from the same sensor in quick succession were combined into a single signal.

Experimental setup. We evaluated each model on three distinct participant cohorts - two with agitation and one with fall events (Table 2), and tuned the experimental parameters on a validation set with confirmed UTI events. 500 days of cross-participant data were kept aside for training. Table 3 provides details of experimental parameters. The adaptive personalized alerting threshold for different participant households was computed using a clinician-specified target alert rate of 7%. Each model was evaluated with 5 different seeds, and executed on a 64-bit Intel i7-8700K CPU, 3.7 GHz Windows 10 machine with 32 GB RAM, using the Pytorch framework (Paszke et al., 2019).

Evaluation. We benchmark the efficacy of our self-supervised Transformer-based feature extraction model against SOTA models suitable for the AD pipeline. This includes (1) Advanced pretrained image feature extractors such as ConvNeXt Tiny (Liu et al., 2022), Xception Net (Chollet, 2017), and MobileNet V2 (Sandler et al., 2018) applied to images of hourly-aggregated sensor data, (2) self-supervised 1D Conv (convolutional) encoder models, and (3) the leading Graph Barlow Twins-based model on this dataset (Bijlani et al., 2024). We run our experiments with four vision-inspired contrastive loss functions - DINO (Caron et al., 2021), SimSiam (Chen & He, 2021), BYOL (Grill et al., 2020) and Barlow Twins (Zbontar et al., 2021), with both, fixed and adaptive temporal augmentation. Note that 1D Conv networks encounter compatibility issues with DINO and BYOL because the frameworks require precise alignment in data structure and network architecture, particularly kernel sizes. For fixed contrast models, we use the best contrast factor based on validation set recall. We report two evaluation metrics: average recall and generalizability, given a target alert rate. Generalizability is the percentage of participants for which the model yields sensitivity greater than 50%. This metric underscores the cross-participant applicability of the model. Due to limited and noisy annotation of anomalous events, accurate identification of false positives is challenging. Therefore precision would be a misleading metric here.

Table 1: Model performance, average (SD), with alert rate within $\pm 1\%$ of the target alert rate of 7%

Model	**Contrast**	**Recall%**	**Generalizability%**	**Parameters**
Pretrained ConvNeXt Tiny (image)	Fixed	80.11	88.83	28 M
Pretrained Xception Net (image)	Fixed	80.16	86.27	22.9 M
Pretrained MobileNet V2 (image)	Fixed	82.48	91.36	3.5 M
G-BT (Bijlani et al., 2024) (graph)	Fixed	81.03 (7.61)	87.96 (6.12)	40 K
1D Conv AE	Fixed	79.22 (5.3)	85.14 (5.66)	6.5 K
1D Conv (SimSiam)	Fixed	81.67 (4.74)	88.42 (5.86)	44 K
1D Conv (Barlow Twins)	Fixed	82.04 (5.89)	87.83 (6.1)	44 K
Transformer (BYOL)	Fixed	80.83 (2.55)	88.13 (2.66)	44 K
Transformer (Barlow Twins)	Fixed	81.85 (3.02)	87.93 (3.18)	88 K
Transformer (SimSiam)	Fixed	83.06 (4.7)	92.22 (4.44)	220 K
Transformer (DINO)	Fixed	83.96 (5.02)	90.03 (4.68)	44 K
Transformer (BYOL)	Adaptive	81.04 (5.28)	88.87 (5.43)	44 K
Transformer (Barlow Twins)	Adaptive	82.31 (4.37)	88.39 (4.36)	88 K
Transformer (SimSiam)	Adaptive	84.3 (3.15)	90.61 (3.52)	220 K
Transformer (DINO)	**Adaptive**	**84.64 (2.36)**	**92.16 (2.33)**	44 K

Results. Table 1 reports the average recall and generalizability on the anomaly detection task for the three participant cohorts, given a clinician-specified target alert rate of 7%. Cohort-wise results are detailed in Tables 4, 5 and 6. Pretrained image-based SOTA models show good out-of-the-box performance on sensor data visualized as heatmaps, but comprise millions of parameters. Custom convolutional encoders for non-image data achieve better recall and lower variance relative to pre-

vious work using Graph Barlow Twins, at a fraction of the size of image models. The Transformer-based model trained with DINO contrastive loss achieves top performance. While backbone models like SimSiam and BYOL exhibit superior performance in specific cohorts, the Transformer-based model utilizing DINO contrastive loss delivers a more balanced performance in terms of recall and generalizability. Importantly, we observe that on the whole, adaptive contrast models show superior performance compared to fixed temporal resolution models despite no prior knowledge of optimal augmentation. We use the temporal attention map and feature importance (Figure 3) and the merged spatiotemporal attention map (Figure 4) to provide insight into routine vs. anomaly, discovering the most influential sensor and hour on a given day. In the example shown, the bathroom sensor is the most significant contributor towards the model outcome during the identified hours. For this correctly flagged agitation event day, the participant's agitated state may be inferred from irregular patterns of bathroom use relative to the established baseline pattern of the previous week, without the need for direct comparison to the activities of the preceding days. We also note that lounge and hallway activity are contra-indicative of the model outcome, which aligns with our understanding that these areas reflect more communal, transitory or highly variable activity in the home, and are generally noisy or non-discriminative factors.

4 Conclusion

In this paper, we proposed a lightweight, computer-vision inspired contrastive learning model to extract salient, noise-adaptive features from sensor-based remote health monitoring data, and used it to detect anomalous home activity that might signal an adverse health event. The temporal contrast factor was dynamically learned through an end-to-end trainable framework, together with the parameters for feature extraction. The ability to adjust focus across temporal scales offers a form of noise adaptation and enables the model to learn stable, discriminative temporal patterns. We evaluated the model's performance in detecting adverse health events in individuals with dementia, achieving higher detection accuracy and cross-participant generalizability than comparable models. The model was used to generate a daily spatiotemporal map which can identify the time of day and sensor that most significantly influences the model's outcome. Our proposed approach, while framed in the context of healthcare monitoring, is domain-agnostic and applicable in other monitoring settings.

Author Contributions

NB: Conceptualization, Methodology, Software, Visualization, Writing - Original Draft; **GC:** Validation, Writing - Review, Supervision; **PB:** Writing - Review, Supervision, Funding Acquisition; **SK:** Writing - Review, Supervision, Funding Acquisition.

Acknowledgments

This study is funded by the UK Dementia Research Institute Care Research and Technology Centre funded by the Medical Research Council (MRC), Alzheimer's Research UK, Alzheimer's Society (grant number: UKDRI–7002), and the Engineering and Physical Sciences Research Council (EPSRC) PROTECT Project (grant number: EP/W031892/1). The infrastructure for this research is partially supported by the NIHR Biomedical Research Centre at Imperial College. We are grateful to the Minder research study participants and their families for contributing to this research.

References

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pp. 435–451. PMLR, 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Nivedita Bijlani, Oscar Mendez, and Samaneh Kouchaki. G-CMP: Graph-enhanced Contextual Matrix Profile for unsupervised anomaly detection in sensor-based remote health monitoring. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24,*

2022. BMVA Press, 2022. doi: https://doi.org/10.48550/arXiv.2211.16122. URL https://bmvc2022.mpi-inf.mpg.de/0854.pdf.

Nivedita Bijlani, Oscar Mendez Maldonado, Ramin Nilforooshan, Payam Barnaghi, Samaneh Kouchaki, CR&T Group, et al. Graph contrastive learning for anomaly detection and personalized alerting in sensor-based remote monitoring for dementia care. *Authorea Preprints*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak. Byol-s: Learning self-supervised speech representations by bootstrapping. In *HEAR: Holistic Evaluation of Audio Representations*, pp. 25–47. PMLR, 2022.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 1–16. Springer, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Gen Li and Jason J Jung. Entropy-based dynamic graph embedding for anomaly detection on multiple climate time series. *Scientific Reports*, 11(1):13819, 2021.

Zekun Li, Shiyang Li, and Xifeng Yan. Time series as images: Vision transformer for irregularly sampled time series. *arXiv preprint arXiv:2303.12799*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Qianli Ma, Jiawei Zheng, Sen Li, and Gary W Cottrell. Learning representations for time series clustering. *Advances in neural information processing systems*, 32, 2019.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Balaram Sarkar, Chandresh K Maurya, and Anshuman Agrahri. Direct speech to text translation: Bridging the modality gap using simsiam. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pp. 250–255, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jiahao Xu, Charlie Soh Zhanyi, Liwen Xu, and Lihui Chen. Blendcse: Blend contrastive learnings for sentence embeddings with rich semantics and transferability. *Expert Systems with Applications*, 238:121909, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21(9):3848–3858, 2019.

Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. Addgraph: Anomaly detection in dynamic graph using attention-based temporal gcn. In *IJCAI*, volume 3, pp. 7, 2019.

Yu Zheng, Zhi Zhang, Shen Yan, and Mi Zhang. Deep autoaugment. *arXiv preprint arXiv:2203.06172*, 2022.

A Appendix

A.1 Participant characteristics

Table 2: Participant cohort characteristics (M-Male, F-Female).

Cohort	**Participants**	**Average age (years)**	**Total days**	**Total observations**	**Event type**	**Event count**
Training and Validation	8(M)	85.13	9363	3780922	UTI	31
	7(F)	82.86				
Test 1	11(M)	82.64	8015	3580451	Agitation	71
	7(F)	85.86				
Test 2	19(M)	81.74	5411	2704927	Agitation	74
	5(F)	83.2				
Test 3	15(M)	83.73	5284	2319328	Fall	38
	8(F)	85.88				

A.2 EXPERIMENTAL PARAMETERS

Table 3: Experimental parameters

Model/Setting	**Parameter**	**Value**
Global	Target alert rate	7%
	Sliding window for thresholding	7 days
	Soft margin for label validation (agitation, UTI)	-10, +7 days
	Soft margin for label validation (fall)	± 7 days
	Default learning rate	0.001
	Default batch size	32
	Maximum epochs	500
	Default dropout rate	0.1
	Output dim	128
Transformer	Patience	20
	Dropout	0.1
	Layers	1
	Attention heads	2
	Forward expansion	32
	Projection layer size	128, 256
	Gumbel-softmax temperature	0.1
	L2 loss weight	0.01
	Weight decay (BYOL)	1e-4
	Momentum coefficient β (BYOL)	0.99
	Momentum coefficient β (DINO)	0.996
	Batch size (DINO, SimSiam)	32
	Batch size (BYOL, Barlow Twins)	256
G-BT	λ (for Barlow Twins loss)	0.005
	Masking aggregation factor	12 hours
	Default threshold	1.8
Conv1D AE (data)	Layers	3
	Padding	1
	Kernel size	(1, 4)
	Channels	32-16-8

A.3 COHORT-WISE MODEL PERFORMANCE

Table 4: Model performance for cohort Test 1, average (SD), alert rate within $\pm 1\%$ of 7% target alert rate

Model	**Contrast**	**Recall%**	**Generalizability%**
Pretrained ConvNeXt Tiny (image)	Fixed	86.06	100
Pretrained Xception Net (image)	Fixed	87.12	88.89
Pretrained MobileNet V2 (image)	Fixed	93.15	100
G-BT Bijlani et al. (2024) (graph)	Fixed	84.55 (5.02)	92.22 (4.97)
1D Conv AE	Fixed	75.25 (5.46)	82.22 (5.44)
1D Conv (SimSiam)	Fixed	81.24 (1.69)	90.00 (4.65)
1D Conv (Barlow Twins)	Fixed	81.36 (5.52)	90.00 (4.65)
Transformer (BYOL)	Fixed	84.95 (2.56)	93.33 (2.22)
Transformer (Barlow Twins)	Fixed	80.45 (1.12)	87.78 (2.22)
Transformer (SimSiam)	Fixed	84.62 (4.7)	92.22 (4.44)
Transformer (DINO)	Fixed	83.95 (2.52)	92.22 (3.04)
Transformer (BYOL)	Adaptive	84.68 (5.27)	93.33 (5.44)
Transformer (Barlow Twins)	Adaptive	81.24 (4.07)	90.00 (2.22)
Transformer (SimSiam)	Adaptive	84.29 (1.95)	92.22 (2.72)
Transformer (DINO)	Adaptive	86.58 (2.31)	94.44 (0)

Table 5: Model performance for cohort Test 2, average (SD), alert rate within $\pm 1\%$ of 7% target alert rate

Model	**Contrast**	**Recall%**	**Generalizability%**
Pretrained ConvNeXt Tiny (image)	Fixed	61.53	70.83
Pretrained Xception Net (image)	Fixed	80.89	91.67
Pretrained MobileNet V2 (image)	Fixed	81.84	95.83
G-BT Bijlani et al. (2024) (graph)	Fixed	81.15 (8.46)	91.67 (5.89)
1D Conv AE	Fixed	83.06 (4.89)	90.83 (3.12)
1D Conv (SimSiam)	Fixed	81.24 (1.69)	90.00 (4.65)
1D Conv (Barlow Twins)	Fixed	81.36 (5.52)	90.00 (4.65)
Transformer (BYOL)	Fixed	76.09 (2.21)	85.83 (2.04)
Transformer (Barlow Twins)	Fixed	81.92 (4.37)	91.67 (4.56)
Transformer (SimSiam)	Fixed	81.66 (4.39)	91.67 (2.63)
Transformer (DINO)	Fixed	82.14 (5.03)	89.17 (4.25)
Transformer (BYOL)	Adaptive	84.81 (3.77)	94.17 (4.25)
Transformer (Barlow Twins)	Adaptive	82.65 (3.71)	90.83 (3.12)
Transformer (SimSiam)	Adaptive	79.92 (1.78)	89.17 (2.04)
Transformer (DINO)	Adaptive	83.73 (2.08)	93.33 (2.04)

Table 6: Model performance for cohort Test 3, average (SD), alert rate within ±1% of 7% target alert rate

Model	Contrast	Recall%	Generalizability%
Pretrained ConvNeXt Tiny (image)	Fixed	92.75	95.65
Pretrained Xception Net (image)	Fixed	72.46	78.26
Pretrained MobileNet V2 (image)	Fixed	72.46	78.26
G-BT Bijlani et al. (2024) (graph)	Fixed	77.39 (8.78)	80.00 (7.28)
1D Conv AE	Fixed	78.84 (4.92)	80.87 (5.22)
1D Conv (SimSiam)	Fixed	81.89 (5.71)	86.09 (6.45)
1D Conv (Barlow Twins)	Fixed	80.00 (6.05)	83.48 (6.45)
Transformer (BYOL)	Fixed	81.45 (2.85)	85.22 (3.48)
Transformer (Barlow Twins)	Fixed	83.19 (2.65)	84.35 (2.13)
Transformer (SimSiam)	Fixed	82.9 (4.8)	86.96 (4.76)
Transformer (DINO)	Fixed	85.8 (5.49)	88.69 (5.22)
Transformer (BYOL)	Adaptive	73.62 (6.44)	79.13 (6.39)
Transformer (Barlow Twins)	Adaptive	83.04 (5.2)	84.35 (6.51)
Transformer (SimSiam)	Adaptive	88.7 (4.77)	90.43 (5.07)
Transformer (DINO)	Adaptive	83.62 (2.65)	88.69 (3.48)

A.4 TRANSFORMER ENCODER

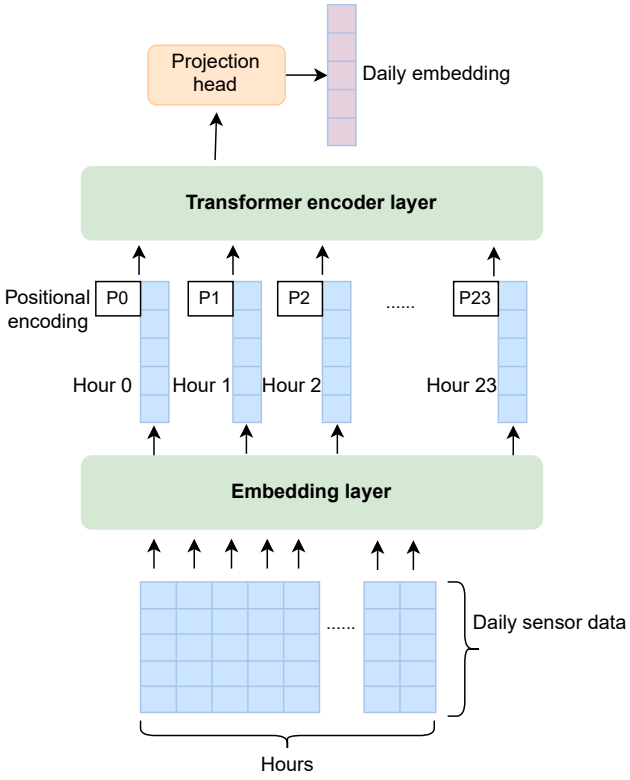

Figure 2: Transformer Encoder applied to temporal sensor data. Hourly data is fed into an embedding layer and position-encoded in time order. The Transformer encoder layer employs self-attention and position-wise feedforward layers, followed by a projection head to produce the daily embedding.

A.5 ATTENTION MAPS

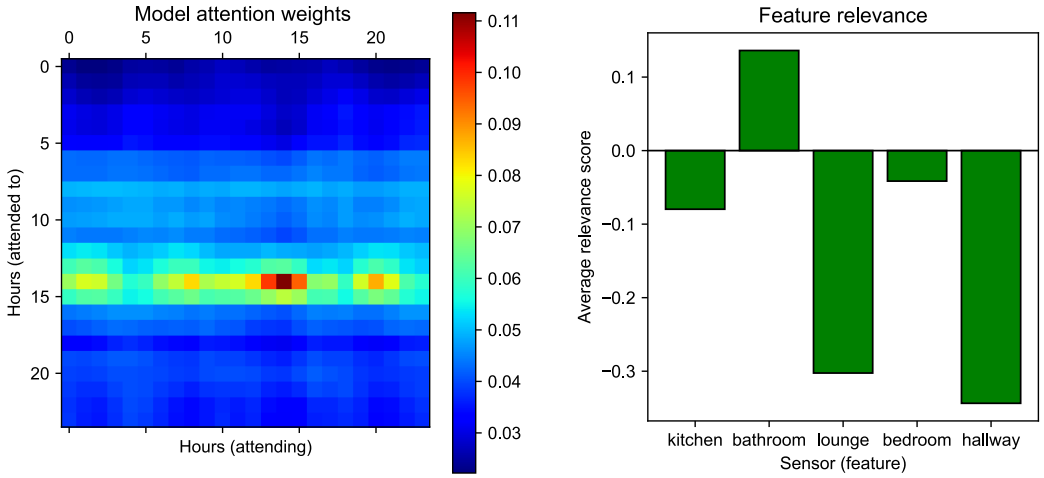

Figure 3: Temporal attention map (left) and feature importance (right) for a participant day flagged as anomalous. The peak attention scores are observed around 2-3 PM. The bathroom sensor had the most significant impact on the model outcome.

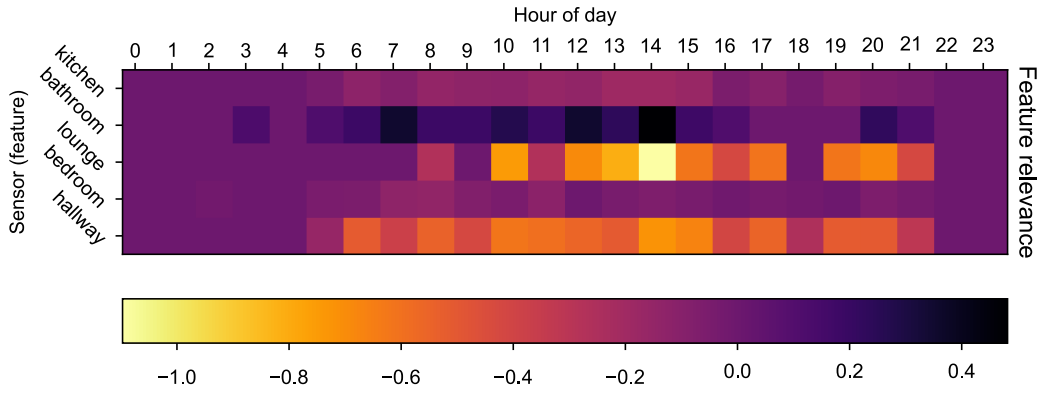

Figure 4: Visualization of the spatiotemporal "signature map" for a specific participant day, merging temporal attention with sensor importance. The bathroom sensor input had the most significant impact on the model outcome, with peak attention scores observed at 7 AM, 12 PM, 2 PM, and 8 PM. Lounge and hallway activity were contra-indicative of the model's decision.