Long Horizon Anomaly Prediction in Multivariate Time Series with Causal Autoencoders

Mulugeta Weldezgina Asres[1], Grace Cummings[2], Aleko Khukhunaishvili[3], Pavel Parygin[4], Seth I. Cooper[5], David Yu[6], Jay Dittmann[7], and Christian W. Omlin[8]

[1,8] *University of Agder, Norway*
mulugetawa@uia.no
christian.omlin@uia.no

[2] *University of Virginia, USA*
gec8mf@virginia.edu

[3] *University of Rochester, USA*
Aleko.Khukhunaishvili@cern.ch

[4] *National Research Nuclear Univ., Russia*
pavel.parygin@cern.ch

[5] *University of Alabama, USA*
seth.cooper@cern.ch

[6] *Brown University, USA*
david_yu@brown.edu

[7] *Baylor University, USA*
jay_dittmann@baylor.edu

ABSTRACT

Predictive maintenance is essential for complex industrial systems to foresee anomalies before major system faults or ultimate breakdown. However, the existing efforts on Industry 4.0 predictive monitoring are directed at semi-supervised anomaly detection with limited robustness for large systems, which are often accompanied by uncleaned and unlabeled data. We address the challenge of predicting anomalies through data-driven end-to-end deep learning models using early warning symptoms on multivariate time series sensor data. We introduce AnoP, a long multi-timestep anomaly prediction system based on unsupervised attention-based causal residual networks, to raise alerts for anomaly prevention. The experimental evaluation on large data sets from detector health monitoring of the Hadron Calorimeter of the CMS Experiment at LHC CERN demonstrates the promising efficacy of the proposed approach. AnoP predicted around 60% of the anomalies up to seven days ahead, and the majority of the missed anomalies are abnormalities with unpredictable noisy-like behavior. Moreover, it has discovered previously unknown anomalies in the calorimeter's sensors.

1. INTRODUCTION

Modern industrial systems utilize sensors to monitor physical quantities such as voltages, currents, flows, temperature, pressure etc. These measurements monitor system state by detecting deviations from normal operating conditions. As one of the pillars of Industry 4.0, Predictive Maintenance (PdM), which primarily depends on early anomaly detection, aims at predicting critical anomalies of a system to improve asset availability by actuating early maintenance before major system faults (Langone, Cuzzocrea, & Skantzos, 2020). Anomaly prediction is an extension of anomaly detection (AD) and focuses on predicting anomalies from early symptoms. It has cost saving potential for large complex systems through prevention of unforeseen system faults, unplanned

downtimes, and maintenance (Wagner & Hellingrath, 2021; Tang, Chen, Bao, & Li, 2019; Huang, Wu, & Wang, 2016; X. Li, Zhang, Ding, & Sun, 2020; Langone et al., 2020). However, most of the data-driven PdM models in the literature employ supervised approaches that require prior labeled anomalies and are limited to short-range predictions (Tang et al., 2019; Huang et al., 2016; X. Li et al., 2020; Langone et al., 2020; Wang, Liu, Zhu, Guo, & Hu, 2018; Hadj-Kacem, Jemaa, Allio, & Slimen, 2020).

In this study, we strive to predict anomalies through data-driven machine learning models from early warning patterns on unlabeled multivariate time series data sets. We propose AnoP, an end-to-end **Ano**maly **P**rediction system using unsupervised long sequence time series forecasting and anomaly detection mechanisms. The proposed system consists of a pipeline of multivariate time series autoencoder models, i.e., a long horizon sequence-to-sequence (S2S) time series forecasting (TSF) model and an AD model. The underlying concept employs a TSF model, trained on the interaction of multivariate sensor signals, to predict future temporal segments, and then uses an AD model to evaluate the predicted signals for potential anomalies. Furthermore, since additive outliers (transient and interpreted on short time scales) are generally unpredictable, our study aims at forecasting anomalous temporary changes that persevere for a certain period (multiple time steps) (T. Wen & Keyes, 2019).

As a use case study for anomaly prediction, we have discussed the Hadron Calorimeter (HCAL) of the CMS experiment at CERN. We have developed the AnoP system to predict anomalies from the multivariate diagnostics sensor data and leverage the health monitoring prognostics of the HCAL's Endcap. Capturing anomalies that persist for substantial periods, often manifested in decaying or growing trends, strange dips, or peaks, are the prime focus of the proposed system. We assessed the performance of the AnoP system in predicting temporal discords using various long sequence horizons on thirty-four Readout Boxes. Because of the lack of labeled anomaly data, we scrutinized the performance in forecasting anomalies with classification metrics as compared with the anomaly flags generated by the AD model when the true signals (non-forecasted) are supplied to it directly without the TSF model. Besides, we have incorporated an evaluation of the forecasting accuracy of the TSF model. Furthermore, we have demonstrated that the proposed system has revealed anomalies that have never been captured before in the HCAL.

The key contributions of our work are highlighted below:

- We present a data-driven unsupervised anomaly prediction mechanism, from heterogeneous multivariate time series sensor dataset.
- We introduce a time block-based S2S TSF model that captures temporal causal interactions for long sequence multivariate time series prediction.
- We present a first study on early prognostics through data-driven methods for the HCAL Endcap Readout Box (RBX) monitoring from diagnostic sensor data.

We discuss background on anomaly prediction and the HCAL system in Section 2, and highlight the data sets used in the study in Section 3. We present the methodology of the proposed AnoP system and modeling approach in Section 4. Section 5 provides performance evaluation in long sequence forecasting and anomaly prediction on the HCAL sensor data sets. Finally, Section 6 offers our conclusion.

2. BACKGROUND

This section discusses background on anomaly prediction, multi-timestep forecasting, and the HCAL system.

2.1. Time Series Anomaly Prediction

Inadequate maintenance techniques can reduce the overall productive capacity of equipment by up to 20%, and unplanned downtimes and reactive maintenance in industrial systems incur substantial costs each year (Kamat & Sugandhi, 2020). PdM applications often refer to performing anomaly detection, diagnostics, and prognostics taking into account the Prognostics and Health Management (PHM) algorithms (Wagner & Hellingrath, 2021).

Conventionally, industries carry out PdM using statistical tests, rule-based alerts, and preset threshold limits (Rezvanizaniani, Dempsey, & Lee, 2014). Owing to the current advancement in sensor and data processing technologies, recent PdM approaches emphasize on machine learning approaches to capture intricate hidden patterns (Wang et al., 2018; X. Li et al., 2020; Wagner & Hellingrath, 2021). However, the existing data-driven approaches for PdM revolve around the development of supervised models which aim at specific labeled data or/and rely on feature extraction signal processing tools such as variants of Fourier transform, Wavelet transform, statistical based and principal component analysis (PCA) (Tang et al., 2019; Huang et al., 2016; X. Li et al., 2020; Langone et al., 2020; Wang et al., 2018; Hadj-Kacem et al., 2020; Hamaide & Glineur, 2021). In (Hadj-Kacem et al., 2020), a machine learning-based anomaly prediction model was proposed using forecasting future time steps mechanism for mobile networks. However, the approach covers short sequences (forecast up to a 16-step horizon) and relies on linear regression, PCA, and supervised logistic regression. Moreover, the efforts on automated feature extraction, via end-to-end deep learning, for prognosis mainly focus on remaining useful time (RUL) estimation (Gugulothu et al., 2017). Generally, the adoption of the above methods for multivariate complex systems is constrained due to high-cost data labeling on heterogeneous sensors. Besides, early signs of anomalies are often not easily seen by experts and

are challenging to annotate in large data sets from numerous monitoring sensors. Furthermore, operational quality-altering anomalies, which do not lead to an ultimate breakdown, are often overlooked. Therefore, unsupervised end-to-end deep learning methods are essential for anomaly prediction system development. Our AnoP approach employs unsupervised models and provides much longer horizon forecasting by capturing non-linear temporal interactions among multidimensional sensors via deep learning models. It determines when a system anomaly will happen, the nature of the anomaly pattern, and the affected sensors.

2.2. Long Sequence Time Series Forecasting

Many real-world applications require long sequence time series predictions, such as price forecasting in the stock market (Y. Liu, Gong, Yang, & Chen, 2020), e-commerce sell prediction (R. Wen, Torkkola, Narayanaswamy, & Madeka, 2017), traffic forecasting (Y. Li, Yu, Shahabi, & Liu, 2017), electricity consumption projecting (Y. Liu et al., 2020; R. Wen et al., 2017; Cinar et al., 2017), weather forecasting (Y. Liu et al., 2020; Cinar et al., 2017) etc. To forecast long sequence time series signals, a model with a high prediction capability (the ability to capture long-range dependencies between predictor and target data effectively) is required (Zhou et al., 2021).

Generally, long sequence forecasting approaches employ S2S autoencoder paradigm using recurrent neural network (RNN) variants (Y. Li et al., 2017; Y. Liu et al., 2020; Qin et al., 2017; Cinar et al., 2017; R. Wen et al., 2017) and Transformer (Zhou et al., 2021). However, RNN-based models may have potential limitations in inference speed and accuracy when sequence length increase due to the recursive step-by-step inferencing (Zhou et al., 2021), and in performance because of deterioration when the length of the input sequence increases (Cho et al., 2014). To address these challenges, decoder models with parallel generation are proposed using attention mechanisms (Y. Liu et al., 2020; Qin et al., 2017; Cinar et al., 2017), multilayer-perceptron (MLP) (R. Wen et al., 2017) and Transformer (Zhou et al., 2021). Nevertheless, these approaches operate only with predefined short horizons (fewer than approximately 40 data points) that limits their scalability (Z. Liu, Loo, & Pasupa, 2021; Y. Li et al., 2017; Y. Liu et al., 2020; Qin et al., 2017; R. Wen et al., 2017) except in (Zhou et al., 2021). Zhou et al. (Zhou et al., 2021) demonstrated the efficacy of an Informer model, a Transformer autoencoder architecture, with various horizons in univariate and multivariate time series data sets. However, the Informer model still lacks S2S generation for longer horizons and requires training of separate models for each target horizon. Besides, sensor data in the real world scenarios are accompanied by missing or invalid values that often results in reading segments with variable length. Hence, incorporating RNN variant models remains relevant for dealing such variability in a time series data.

2.3. Readout Boxes of the HCAL Detector

The CMS Experiment is one of the two general purpose detectors operating at CERN's Large Hadron Collider (LHC) (Collaboration et al., 2008). The Hadron Calorimeter (HCAL) of CMS is responsible for measuring the energy of hadronic showers originating from the LHC collisions. The HCAL is divided into four subdetectors: the HCAL Barrel (HB), HCAL Endcap (HE), HCAL Outer (HO), and HCAL Forward (HF). This paper discusses only the monitoring data of the HE subdetector, a brass and scintillating plastic sampling calorimeter.

The HE is arranged into two hemispheres, HE Plus (HEP), and HE Minus (HEM). Each half is further divided into eighteen identical wedges. Signal from each wedge is read out by one "Readout Box" (RBX). Figure 1 showcases the RBX numbering and HE geometry. The RBX represents the smallest unit of front-end control and power, with each RBX consisting of a water-cooled aluminum shell housing the front-end data acquisition, control, and communication electronics. The electronics consist of low voltage distribution, high voltage distribution, four Readout Modules (RMs), a Calibration Unit (CU) and one next-generation Clock, Control, and Monitoring Module (ngCCM). The ngCCM provides backend-to-frontend communication, control, and clock distribution.

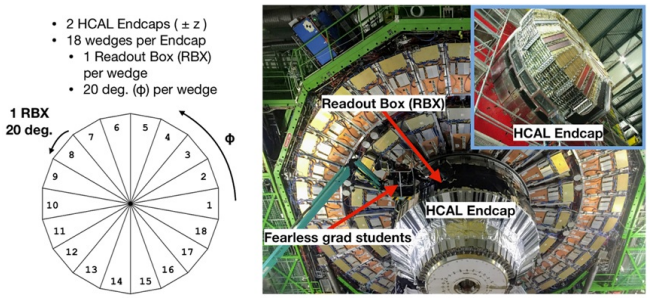


Figure 1. The HE subdetector of the CMS Experiment. Left: arrangement of eighteen RBXes. Right: installation position of the HE on the CMS detector.

To maintain physics data acquisition quality, predicting faults of the detector electronics is essential. Currently, the CMS HCAL only uses automated monitoring for general detector safety through established Detector Control and Detector Safety Systems (DCS and DSS, respectively). These systems use a small subset of the available monitored variables available to generate threshold-based alerts on quantities like temperature or bias voltage. Therefore, machine learning models have been explored for system monitoring automation of the CMS detectors through time series anomaly detection (Asres et al., 2021; Paltenghi, 2020; Azzolin et al., 2019; Wielgosz, Skoczen, & Wiatr, 2018; Wielgosz, Mertik, Skoczeń, & De Matteis, 2018). Anomalous behavior in additional variables can also indicate future detector performance issues, and escape the DCS and DSS monitoring. For

example, the gradual decrease in the monitored Received Signal Strength Indicator (RSSI) current, which is proportional to the received light at the front end from the back end optical communication links, preceded control communication loss during operation in 2018 and 2019 (Cummings & the CMS Collaboration, 2021). RSSI was not actively monitored, and trends such as depicted in Figure 2 could have been predicted. The proposed approach in this paper attempts to detect such anomalies from early signs before they affect data quality or result in loss of data.

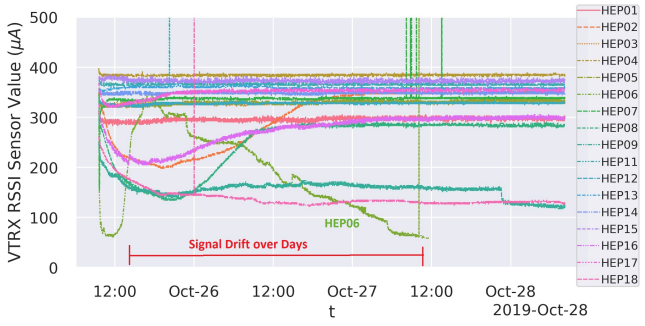


Figure 2. Gradual drifting anomalies on RSSI before ngCCM lost communication in 2019. A strong decay over three days is illustrated for the HEP06 RBX.

3. Data Set Description

In this study, we have utilized front-end electronics diagnostic sensor data from the HCAL. These data sets are recorded for detector health monitoring and diagnostic purposes, not for physics data analysis. We have used ngCCM monitoring data from the HE subdetector collected in 2018 using the HE monitoring service. The HE monitoring service communicates to the front-end electronics through the ngCCM server, a software that handles access to the ngCCM. The data set contains 86M readings of around 2600 monitored quantities, measured once per minute, from 34 active RBXes (HEP01–18 and HEM01–18, excluding HEM15 and HEM16) from September to December 2018. The signals are composed of current, voltage, and optical power measurements of various components of the ngCCM. Finally, we downsampled the data into hourly intervals by averaging to capture the relevant temporal information.

4. Methodology

This section provides the methodology of the proposed anomaly prediction approach and models.

The proposed AnoP system is composed of two multivariate time series autoencoder models combined in a pipeline, i.e., i) a multi-timestep TSF model, and ii) an AD model (see Figure 3). We have discussed below the mathematical formulation and model architectures for the TSF and AD of the AnoP in Section 4.1 and Section 4.2, respectively. Section 4.3 elaborates the data preprocessing, preparation of training data sets and model training.

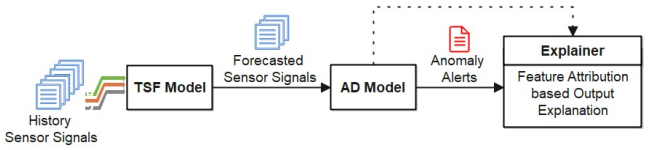


Figure 3. System design of the proposed AnoP system. The TSF model predicts a long sequence of signals, and the AD model produces anomaly status of the predicted signals based on reconstruction scores. The explainer yields explanation for the detected anomalies using post-hoc feature attribution estimation.

4.1. Multivariate Multi-timestep Forecasting Model

For long sequence forecasting, we propose a robust attention-based S2S dynamic conditional decoding mechanism. In essence, a TSF model needs to cope with two challenges in anomaly prediction. First, it should predict the deviating signals belonging to anomalies from their early fluctuation patterns. Second, it should also quickly adjust its prediction after intervention or maintenance, when normal system behavior is resumed. To achieve these capabilities, we integrate a conditional decoder for the TSF model where the latest time window of the sensors is used as conditional input. The conditional decoding enables the TSF model to respond faster when the sensor signals begin to evolve. Additionally, we employ dynamic decoding, a recursive conditional decoder, to allow dynamic long-horizon forecasting. Dynamic conditional decoding is a mechanism in which earlier slices or tokens from the model output are supplied into the decoder as conditional input to generate the subsequent output sequence. This approach has been successfully applied with S2S models in natural language processing domains such as language translation (Sutskever, Vinyals, & Le, 2014; Devlin, Chang, Lee, & Toutanova, 2018). Conditional decoding without the recursive dynamic decoding has also been extended into time series data sets in recent studies (Zhou et al., 2021).

4.1.1. Mathematical Formulation

Let the input time series data is $x^T \in \mathbb{R}^{N_x \times T}$ where N_x is the number of input sensors with a history sequence of $t_x \in [t' - T, t']$, with length of T. The TSF model F predicts the sequence $y^H \in \mathbb{R}^{N_y \times H}$ with a horizon time window of $t_y \in [t' + 1, t' + H]$ for N_y target sensors. Since, the TSF model employs S2S autoencoder, the encoder F_e maps the input x^T into context z_e and state vectors h_e Eq. (1):

$$z_e, h_e = F_e(x^T) \tag{1}$$

The decoder F_d utilizes dynamic conditional decoding that uses the context vectors z_e and conditional input sequences from the target sensors y_d from the last time steps $t_d \in [t -$

$T_d, t]$ with a size of T_d to predict the multi-timestep signals y^H and generate decoding state h_d Eq. (2):

$$y^H, h_d = F_d(y_d, z_e, h_d) \tag{2}$$

When inferencing long sequence horizon $H_l > H$ with size of l, the decoder uses dynamic decoding that behaves in an autoregressive manner employing a time block-based S2S approach (see Algorithm 1). The decoder initializes its states h_d from the encoder states, $h_d = h_e$, and then recursively predicts multi-timestep signal segments of the size H (from line 7 to 11 in Algorithm 1). The latest predicted horizon y^H is combined with the y_d to form a new conditional input to the decoder for the subsequent forecasts (line 9).

Algorithm 1 Multistep Forecasting Inference

1: **procedure** TIMEBLOCKS2SMULTISTEPFORECASTING(F, x, y_d, H_l)
▷ F : forecasting S2S encoder-decoder model
▷ x : multivariate input times series signals with size of $N_x \times T$
▷ y_d : initial decoder input from past time-window of the target signals
▷ H_l : time length of the target horizon
2: $H \leftarrow getModelHorizonSize(F)$
3: $N_i \leftarrow H_l/H$ ▷ number of forecasting iterations with basic block of H
4: $z_e, h_e \leftarrow F_e(x)$ ▷ get the learned context vectors and states from the encoder
5: $h_d \leftarrow h_e$ ▷ initial state of decoder
6: $y \leftarrow []$
7: **for** i *in* $[1, .., N_i]$: **do**
8: $y^H, h_d \leftarrow F_d(y_d, z_e, h_d)$
9: $y \leftarrow join(y, y^H)$ ▷ concatenate on the time dimension
10: $y_d \leftarrow getCondInput(y^H, y_d)$ ▷ update conditional input
11: **return** y
12: **procedure** GETCONDINPUT(y^H, y_d) ▷ returns decoder conditional input segment
13: $H \leftarrow length(y^H)$
14: $T_d \leftarrow length(y_d)$
15: **if** $H \leq T_d$ **then**
16: $y_d \leftarrow join(y_d\{t \in [H, T_d]\}, y^H)$ ▷ update the latest H steps of y_d from y^H
17: **else**
18: $y_d \leftarrow y^H\{t \in [H - T_d, H]\}$ ▷ get the latest T_d steps from the y^H
19: **return** y_d

Furthermore, to improve attentiveness of the conditional inputs and leverage the multi-timestep forecasting accuracy, the decoder employs a multi-attention mechanism (see Figure 4). The model is composed of three parallel attention layers; one for the encoded latent or context vectors z_e, and two blocks for the conditional multivariate sensor signals y_d on the feature (sensor quantity) and time dimensions, respectively Eq. (3):

$$\begin{aligned} \psi_{z_e} &= softmax(z_e) \\ \psi_{y_d^f} &= softmax(y_d^t) \\ \psi_{y_d^f} &= softmax(y_d^f) \end{aligned} \tag{3}$$

where ψ_{z_e} is attention on the learned encoder context vector z_e, and $\psi_{y_d^t}$ and $\psi_{y_d^f}$ are attention scores of the decoder conditional input y_d on its temporal and feature dimensions, respectively. Finally, attention scores are concatenated to form predictor features for the multi-timestep forecasting Eq. (4):

$$\psi = [\psi_{z_e} || \psi_{y_d^T} || \psi_{y_d^F}] \tag{4}$$

4.1.2. Model Architecture

The proposed TSF S2S autoencoder model is composed of residual dilated convolutional and GRU networks with attention (see Figure 4).

To achieve temporal causation learning, multiple convolutional layers are stacked in the network with increasing dilation size. The increasing dilation along subsequent layers expands the receptive field of the convolution operation in the time data (Bai, Kolter, & Koltun, 2018; He & Zhao, 2019). Furthermore, to mitigate the performance degradation for long input sequences, we have ameliorated the model with time dimension reduction through multilayer pooling. Moreover, residual skip connections are added in the convolutional network to enhance training with deep layers.

Unlike the encoder, the decoder utilizes an attention-based network that takes decoding inputs from the encoded latent features and conditional signals. Nevertheless, the remaining sections of the decoder consists of similar blocks as the encoder but in reverse order and in deconvolution configuration. It also employs a final deconvolution layer with unit kernel size for output stabilization. Generally, the number of convolutional blocks on the encoder and decoder may differ since the encoder attempts to learn relevant context from the history time window, whereas the decoder's purpose is to predict the signals in the horizon time window. Furthermore, the conditional input signals to the decoder pass through a convolutional embedding block, to extract relevant temporal features, before the attention network. Unlike previous studies (Qin et al., 2017; Zhou et al., 2021), the attention network in our model is not followed with a fully-connected layer to reduce model complexity. It is directly connected to the GRU network, and the input weights of the first GRU layer can provide a similar functionality as fully-connected layer.

4.2. Multivariate Time Series Anomaly Detection Model

The AD model employs variational autoencoder G that attempts to reconstruct $\overline{x}^T$ from a multivariate input data $x^T \in \mathbb{R}^{N \times T}$ from N sensors on a time sequence $t \in [t' - T, t']$. The encoder of the model provides normally distributed low-dimensional representation latent signals z Eq. (5). The decoder generates the reconstructed signals $\overline{x}^T$ from encoded latent signals Eq. (6):

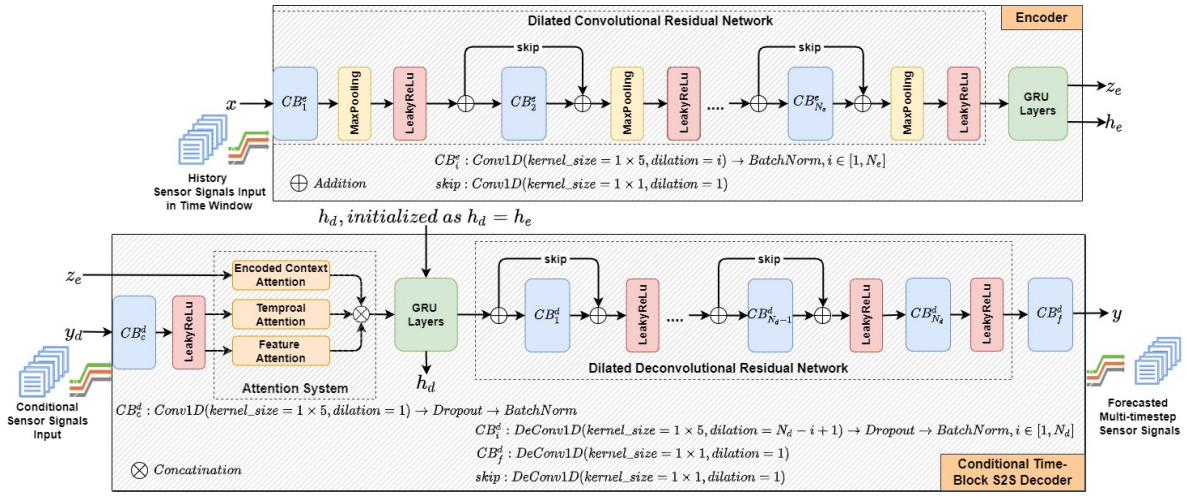


Figure 4. Architecture of the multi-timestep forecasting S2S autoencoder of the TSF model. The residual block: consists of a 1D *dilated convolutional* network while the recurrent neural network contains two *GRU (encoder hidden_size: 16 → 16, decoder: 16 → 256)* layers. The convolutional block: 1D *dilated convolutional (256 kernels*, except CB_c^d and CB_f^d with *16* and N_y *kernels*, respectively) for fast localized feature extraction, *BatchNorm* for network weight regularization and faster convergence, *LeakyReLU* for non-linear activation, and *MaxPooling* for prominent features retrieval that are insensitivity to time translation. *Softmax*: builds the attention in the decoder. Finally, *Dropout=0.20* for further training regularization. Temporal causal learning via the convolutional layers with varying size of dilation and the GRU layers.

$$z = G_e(x^T) \tag{5}$$

$$\overline{x}^T = G_d(z) \tag{6}$$

Finally, the model estimates anomaly scores from the signal reconstruction errors. For each univariate sensor, reconstruction anomaly scores at time t' are calculated based on Mean Absolute Error (MAE) Eq. (7):

$$a_i(t') = \frac{1}{T} \sum_{t=t'-T}^{t'} |x_i(t) - \overline{x}_i(t)| \tag{7}$$

where x_i and $\overline{x}_i$ are the input and reconstructed signals of the i^{th} sensor. The multidimensional reconstruction score is finally converted into system anomaly score using Mahalanobis distance (D_{md}) estimation, multidimensional distance between a point (vector) and a distribution (De Maesschalck, Jouan-Rimbaud, & Massart, 2000) Eq. (8).

$$D_{md} = \sqrt{(A_i - \mu)^T \cdot C^{-1} \cdot (A_i - \mu)} \tag{8}$$

where D_{md} is the Mahalanobis distance. The vector $\mathrm{A_i}$ is the multivariate anomaly score of the i^{th} observation, the vector μ contains the mean values of the univariate scores (across all observations), and C^{-1} is the inverse covariance matrix of A. Finally, a threshold $K_{md} = \alpha_{md}\mu_{md}$ is applied on the D_{md} to generate anomaly flags. The $\mu_{md} = \mathbb{E}[D_{md}]$ is the mean distance, and α_{md} contains the adjustable parameters to tune detection sensitively.

Finally, the unsupervised autoencoder is built on 1D convolutional and GRU networks, accompanied by a post-hoc anomaly explainer based on feature attribution algorithms such as *Integrated Gradient* and *SHAP* (see Figure 5). The model is adopted from our previous work on multivariate AD for the HCAL sensor diagnostics, and further description and performance evaluation on the model can be found in (Asres et al., 2021).

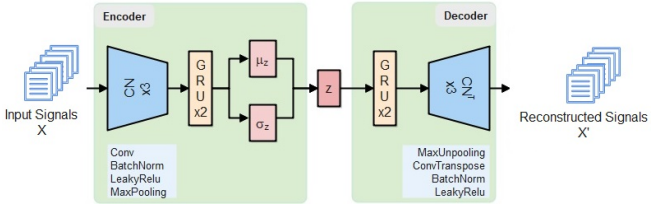


Figure 5. Architecture of the multivariate reconstruction autoencoder of the CGVAE AD model. The convolutional network is consists of three blocks; each consists of 1D *convolutional layer (64 kernels, kernel_size*=1×3*)*. The recurrent network consists of two *GRU (encoder hidden_size: 16 → 4, decoder: 4 → 16)* layers. μ_z and σ_z are fully-connected linear layers implementing the *variation layer* and $z = \mu_z + \sigma_z \odot \epsilon$, where $\epsilon \sim N(0, I)$ and $\odot$ signify an element-wise product.

4.3. Dataset Preparation and Model Training

Since the TSF and AD models of the proposed AnoP system require different training data sets, the models were

trained separately. The AD needs a training dataset with healthy instances or low anomaly contamination, while the TSF requires substantial predictable anomalies in its training dataset. However, obtaining clean data of healthy instances in the training data is one of the main challenges of semi-supervised learning of AD models (Munir, Siddiqui, Dengel, & Ahmed, 2018). We cleaned the potential outliers from each univariate sensor data in the training set using state-of-the-art time series outlier detection algorithm, Saliency Residual (SR) (Asres et al., 2021; Zhao et al., 2020). On the other hand, the TSF autoencoder was trained on the dataset contaminated with anomalous patterns to leverage its capability to forecast anomaly signals from early signs. The modeling approach is fully unsupervised and does not require any labeling. However, since anomalies are rare instances, the model may struggle to learn the anomaly signals due to the class imbalance. We attempted to mitigate the challenge with support of the AD model. We selected the data sources, the RBXes, that have a significant number of outliers (potential anomalies) spanning substantial periods on the sensor data.

Finally, we trained the autoencoder models with *Adam* optimizer using a *super-convergence cyclic* learning rate scheduling mechanism (Smith & Topin, 2019). To mitigate Kullback-Leibler (KL) divergence vanishing or latent squashing for the variational autoencoder of the AD model, we have applied a *cyclic annealing* method (Fu et al., 2019) when KL divergence loss regularizes the training cost function.

5. Experimental Results and Discussion

In this section, we present the performance of the proposed long sequence time series forecasting model and the AnoP system, and finally, share ideas for future research directions.

As discussed in Section 4, we trained the TSF and AD models of the AnoP on different data sets with twenty-six sensors per RBX. In our experiment, we have used the same sensors for the input and target, $N_x = N_y$. The TSF autoencoder was trained on two-month data, 10–11/2018, from six RBXes (HEM01, HEM04, HEM17, HEP14, HEP15, and HEP18), while one-month data, 10/2018, from four stable RBXes (HEM01, HEM07, HEM17, and HEP11) were used to train the AD autoencoder. The models were developed with PyTorch and trained up to 5000 iterations. Finally, we have evaluated performance of the proposed models on the date range of 25/09–03/12/2018 for thirty-four RBXes.

The TSF uses a $T = 120$ hours (5 days) sliding history time-window with prediction horizon sizes of $H = [24, 168]$ hours (1 to 7 days). The conditional decoder of the model uses the last $T_d = 24$ hours from the history time window for the target sensors. The AD model predicts anomalies on the 24 hours sliding window. We have set $\alpha_{md} = 10$, determined heuristically, to estimate the anomaly detection decision thresholds for the reconstruction anomaly detection.

Finally, we compared the anomaly prediction performance of the AnoP with the benchmark CGVAE AD model. The benchmark model is the same as the AD model of the AnoP except it detects the anomaly from the raw sensor signals in contrast to the AnoP, where the AD model detects anomalies from the forecasted signals.

5.1. Multi-timestep Forecasting Model Evaluation

In this section, we present the results on performance evaluation of the TSF model in forecasting long horizon sequences.

The model employs $N^e_{cd} = 2$ and $N^d_{cd} = 4$ casual residual convolution blocks for the encoder and decoder networks, respectively, and basic forecasting horizon $H = 24$ hours. We assessed the efficacy on multiple long horizon sizes, i.e., 24 to 168 samples (see Table 1). The results demonstrate that the model forecasted long horizons with slight performance degradation through time block S2S mechanism.

Table 1. Multivariate time series forecasting performance, averaged from all RBXes, on different horizons.

Horizon (H)	24h	48h	72h	96h	120h	144h	168h
MAE	0.418	0.430	0.444	0.464	0.473	0.503	0.529
MSE	1.392	1.416	1.465	1.515	1.558	1.635	1.705

MAE - Mean Absolute Error, MSE - Mean Square Error

Figure 6 illustrates the forecasting capability of the proposed attention mechanism with the conditional decoding as compared with conditional decoder without attention. The mean absolute error (MAE) and mean square error (MSE) performance improve substantially by 10–15% and 22–28%, respectively. Furthermore, Figure 7 portrays an ablation study on the TSF model demonstrating the promising contribution of the major building blocks of the model, i.e. the attention, conditional decoding, and convolution layers.

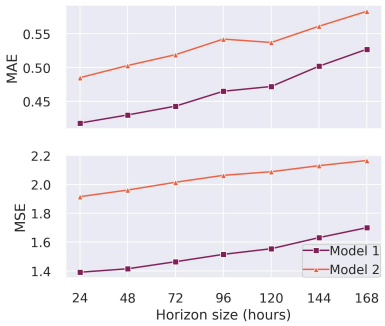
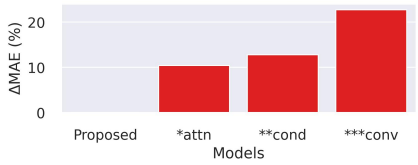


Figure 6. Multivariate time series forecasting performance comparison between different model configurations. *Model 1: the proposed attention-based conditional decoder*, and *Model 2: conditional decoder without attention.*

5.2. Anomaly Prediction Performance

In the absence of annotated data, we define an anomaly as an outlier that deviates from the expected nominal charac-



* is number of excluded blocks from the proposed TSF model

Figure 7. Ablation performance evaluation of the TSF model at $H = 24$ hours. The MAE score difference in percentage is give relative to the proposed model. **attn – w/o attention, **cond – w/o conditional decoding*, and ****conv – w/o convolution layers*.

teristics. Thus, not all anomalies indicate failure in the detector. The efficacy of the AD model was assessed as compared with benchmark error-counter variables of the HCAL in (Asres et al., 2021). However, the counters are less convenient to be used for anomaly prediction evaluation as they are ineffective in capturing most of the gradual system deterioration anomalies (Asres et al., 2021). Hence, we generated reference anomaly labels from the AD model, i.e., AD on the raw data (not forecasted) to assess the performance of the proposed anomaly prediction system.

Generally, on average, the AD model flagged around 160 anomalous reading points per RBX on the raw data, monitored from twenty-six sensors over a period of 10 weeks (see Figure 8). Exceptionally, higher number of flags were generated from a few RBXes due to higher variability on the readings from 1V2_CURRENT sensor on the slave control card of ngCCM (see discussion below at the end of this section).

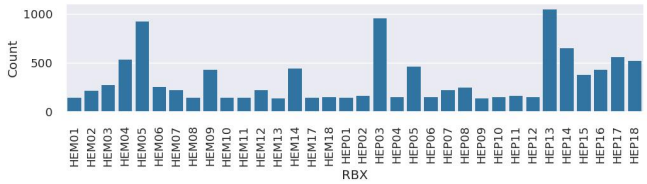
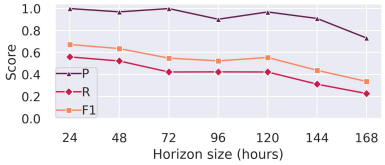


Figure 8. Number of anomaly data points, detected by the CGVAE AD model, that are used as reference flags for the evaluation of the anomaly prediction system. High number of anomalies in some RBXes such as HEM05, HEP03, HEP05, HEP13, and HEP14 due to noisy behavior of the 1V2_CURRENT sensor of the ngCCM slave control card.

Figure 9 and 10 portray the classification performance on prediction accuracy of the proposed AnoP system. The AnoP has predicted long horizon anomalies with high precision, demonstrating the robustness of the proposed system in avoiding false flags (see Figure 9). Despite this good performance, the recall is just below 0.60. This limitation is due to missed anomalies arising from unpredictable transient behavior. Additive noise is a prime cause of transient anomalies.

Our models revealed a noisy behavior of the 1V2_CURRENT sensor of the ngCCM slave control card of some RBXes. Figure 11 illustrates an example of the sensor's behavior and our



* P - Precision, R - Recall, F1 - F1-score

Figure 9. Anomaly prediction performance of AnoP as compared to the CGVAE AD model across different horizons.

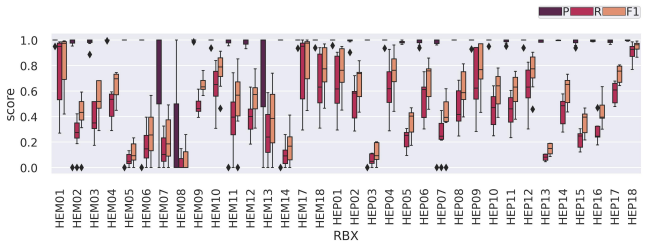


Figure 10. Distribution of anomaly prediction performance of AnoP on multiple horizons across RBXes. The lower performance in some RBXes is generally due to additive transient anomalies and noisy slave control card sensors. HEM08 has missing sensor which impacts the prediction (data was imputed with nominal value).

AnoP model's response. The AD model generated substantial anomaly flags for those particular RBXes (see Figure 8), but the AnoP struggled to achieve good anomaly forecast (low recall) due to lack of learnable causal patterns (see Figure 10). While this was the first observation of this phenomenon in the HCAL, the behavior is not entirely unexpected. The slave control card can be noisier than the master due to the mounted FPGA's attempt to lock onto a non-existent incoming datastream, since the slave card does not maintain the backend communication link. This behavior does not impact operation, but monitoring its status would provide relevant information when the decision of switching the master ngCCM control card is made.

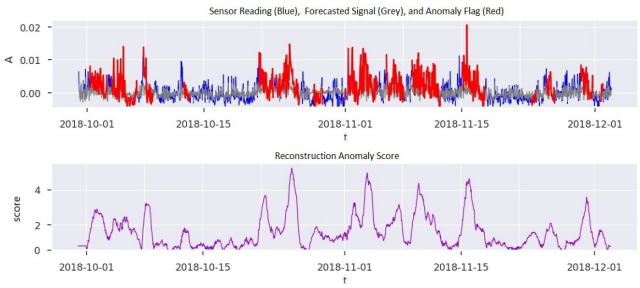


Figure 11. Anomaly prediction on 1V2_CURRENT sensor of the ngCCM slave control card of RBX HEP05. The sensor is found to be noisier in some RBXes. For instance, five times stronger noise-like fluctuation (around 0.01A) was observed as compared to its corresponding master card (0.002A) and slave card of the other RBXes. The sensor contributed a large number of anomaly flags for the RBXes. The value of the y-axis is normalized reading after subtracting the mean value across the period.

Persistent anomalies are often indicators of severe problems in the monitored system, and Figure 12 portrays a captivating anomaly captured from the successful forecast of persistent outliers, i.e., in the current and voltage sensors of the RBXes from October 28 to November 03, 2018. We found that during that time there had been Machine Development (MD) and Technical Stop (TS) tasks on the LHC. The MD weeks are planned in the LHC operation schedule to optimize and study the performance of the machine and to allow the operators to improve the long-term performance of the LHC. Following our finding, investigations revealed that the MD and TS task had unexpectedly affected the low-voltage supply of the RBX. The changes were within tolerance, and did not negatively impact HCAL's performance, but this knowledge allows the HCAL team to better prepare for LHC interventions in the future.

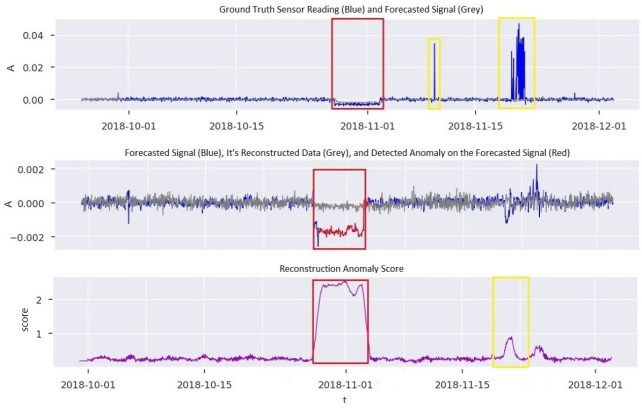


Figure 12. Forecasting capability on persistent and transient anomalies on the 1V2_CURRENT sensor of the master control card of the HEP03. (Top) forecasted signal from the TSF autoencoder using 24 hours horizon as compared to the ground truth signal. (Middle) signal reconstruction via the AD autoencoder from the forecasted signal. (Bottom) the estimation of the reconstruction-based anomaly score on the forecasted signal. Red boxes highlight the persistent outliers (successfully forecasted), whereas the yellow boxes enclose the transient or spike outliers (challenging to forecast).

5.3. Directions for Future Research

The robustness of AnoP relies on the accuracy of the employed TSF and AD models. In general applicability to sensor data with limited anomaly samples, two suggestions can be rendered generally to mitigate the class-imbalance during training of the TSF model, i.e., (i) weighted training loss functions, and (ii) data augmentation through synthetic data generation. Having an AD model beforehand, the data sets can be annotated with ease and higher weights can be assigned to the sections with anomalous patterns during training loss estimation. The other alternative is to generate and incorporate synthetic data into the training dataset (Ducoffe, Haloui, & Gupta, 2019). The recent progress on deep generative adversarial network (GAN) models has demonstrated good capability on multivariate time series signals (Ducoffe et al., 2019; Yoon, Jarrett, & Van der Schaar, 2019).

6. Conclusion

Predictive Maintenance, owing to its versatile leverages in significantly cutting maintenance costs and downtimes, has become a pillar application of Industry 4.0. In this study, we have demonstrated the efficacy of the anomaly prediction approach (AnoP) through unsupervised end-to-end long time series forecasting and anomaly detection mechanisms on multivariate time series data. The experimental evaluation on the CMS HCAL diagnostic monitoring sensor data sets has unveiled that anomalies that persevere for a certain period can be forecasted from early indications. The developed anomaly prediction system is expected to enable prognostics and predictive maintenance in the HCAL during LHC RUN III. Currently, the AnoP is under pre-production testing phase for the HCAL monitoring. Finally, the proposed approaches of the AnoP are generic enough to be applied with less effort for predictive maintenance applications in other domains with time series data.

References

Asres, M. W., Cummings, G., Parygin, P., Khukhunaishvili, A., Toms, M., Campbell, A., ... Omlin, C. W. (2021). Unsupervised deep variational model for multivariate sensor anomaly detection. In *Ieee pic* (pp. 364–371).

Azzolin, V., Andrews, M., Cerminara, G., Dev, N., Jessop, C., Marinelli, N., ... Vlimant, J.-R. (2019). Improving data quality monitoring via a partnership of technologies and resources between the cms experiment at cern and industry. In *Epj web of conference* (Vol. 214, p. 01007).

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*.

Cinar, Y. G., Mirisaee, H., Goswami, P., Gaussier, E., Aït-Bachir, A., & Strijov, V. (2017). Position-based content attention for time series forecasting with sequence-to-sequence rnns. In *Iconip* (pp. 533–544).

Collaboration, C., Chatrchyan, S., Hmayakyan, G., Khachatryan, V., Sirunyan, A., Adam, W., ... others (2008). The cms experiment at the cern lhc. *JInst*, *3*, S08004.

Cummings, G., & the CMS Collaboration. (2021). *Cms hcal vtrx-induced communication loss and mitigation.* private communications.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometr Intell.*

Lab. Syst., *50*(1), 1–18.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Ducoffe, M., Haloui, I., & Gupta, J. S. (2019). Anomaly detection on time series with wasserstein gan applied to phm. *IJPHM*, *10*(4).

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv:1903.10145*.

Gugulothu, N., Tv, V., Malhotra, P., Vig, L., Agarwal, P., & Shroff, G. (2017). Predicting remaining useful life using time series embeddings based on recurrent neural networks. *arXiv:1709.01073*.

Hadj-Kacem, I., Jemaa, S. B., Allio, S., & Slimen, Y. B. (2020). Anomaly prediction in mobile networks: A data driven approach for machine learning algorithm selection. In *Ieee/ifip noms* (pp. 1–7).

Hamaide, V., & Glineur, F. (2021). Unsupervised minimum redundancy maximum relevance feature selection for predictive maintenance: Application to a rotating machine. *IJPHM*, *12*(2).

He, Y., & Zhao, J. (2019). Temporal convolutional networks for anomaly detection in time series. In *Journal of physics: Conference series* (Vol. 1213, p. 042050).

Huang, C., Wu, X., & Wang, D. (2016). Crowdsourcing-based urban anomaly prediction system for smart cities. In *Proceedings of acm cikm* (pp. 1969–1972).

Kamat, P., & Sugandhi, R. (2020). Anomaly detection for predictive maintenance in industry 4.0-a survey. In *E3s web of conferences* (Vol. 170, p. 02007).

Langone, R., Cuzzocrea, A., & Skantzos, N. (2020). Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *DKE*, *130*, 101850.

Li, X., Zhang, W., Ding, Q., & Sun, J.-Q. (2020). Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. *Journal of Intelligent Manufacturing*, *31*(2), 433–452.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv:1707.01926*.

Liu, Y., Gong, C., Yang, L., & Chen, Y. (2020). Dstp-rnn: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Systems with App.*, *143*, 113082.

Liu, Z., Loo, C. K., & Pasupa, K. (2021). A novel error-output recurrent two-layer extreme learning machine for multi-step time series prediction. *Sustainable Cities and Society*, *66*, 102613.

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, *7*, 1991–2005.

Paltenghi, M. (2020). *Time series anomaly detection for cern large-scale computing infrastructure* (Unpublished doctoral dissertation). Politecnico di Milano (IT).

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *arXiv:1704.02971*.

Rezvanizaniani, S. M., Dempsey, J., & Lee, J. (2014). An effective predictive maintenance approach based on historical maintenance data using a probabilistic risk assessment: Phm14 data challenge. *IJPHM*, *5*(2).

Smith, L. N., & Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In T. Pham (Ed.), (Vol. 11006, pp. 369 – 386). SPIE. doi: 10.1117/12.2520589

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Neurips* (pp. 3104–3112).

Tang, Z., Chen, Z., Bao, Y., & Li, H. (2019). Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. *Struct. Cont. and Health Mon.*, *26*(1), e2296.

Wagner, C., & Hellingrath, B. (2021). Supporting the implementation of predictive maintenance: a process reference model. *IJPHM*, *12*(1).

Wang, J., Liu, C., Zhu, M., Guo, P., & Hu, Y. (2018). Sensor data based system-level anomaly prediction for smart manufacturing. In *Ieee bigdata* (pp. 158–165).

Wen, R., Torkkola, K., Narayanaswamy, B., & Madeka, D. (2017). A multi-horizon quantile recurrent forecaster. *arXiv:1711.11053*.

Wen, T., & Keyes, R. (2019). Time series anomaly detection using convolutional neural networks and transfer learning. *arXiv:1905.13628*.

Wielgosz, M., Mertik, M., Skoczeń, A., & De Matteis, E. (2018). The model of an anomaly detector for hilumi lhc magnets based on recurrent neural networks and adaptive quantization. *IFAC EAAI*, *74*, 166–185.

Wielgosz, M., Skoczen, A., & Wiatr, K. (2018). Looking for a correct solution of anomaly detection in the lhc machine protection system. In *Icses* (pp. 257–262).

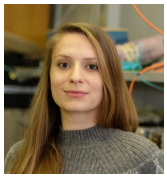
Yoon, J., Jarrett, D., & Van der Schaar, M. (2019). Time-series generative adversarial networks. *NeurIPS*, *32*.

Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., ... Zhang, Q. (2020). Multivariate time-series anomaly detection via graph attention network. *arXiv:2009.02040*.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of aaai.*

BIOGRAPHIES

Mulugeta Weldezgina Asres is a Ph.D. candidate in Artificial Intelligence at the University of Agder, Norway. His current research is on AI models for the Hadron Calorimeter monitoring and diagnostics at the CMS Experiment at CERN. He received his B.Sc. and M.Sc. in Computer Engineering and Gold Medal Award for the highest CGPA from EiT-M, Mekelle University. He conducted post-graduate research on machine learning for non-intrusive energy monitoring, power substation systems, and telecom networks at the Polytechnic University of Turin. His research interests focus on system monitoring, time series modeling, deep learning, and Industry 4.0.

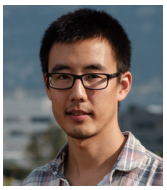


Grace Cummings is a Ph.D. candidate in Experimental High Energy Particle Physics at the University of Virginia. She received her Bachelor of Science in Physics at Virginia Commonwealth University. Her current research interest is calorimetry for particle physics with an emphasis on detector development, systems testing, and front-end instrumentation.

Aleko Khukhunaishvili is a research associate at the University of Rochester. He received his Ph.D. in Physics at Cornell University in 2014. His current research interests focus on precision Standard Model measurements at the LHC.

Pavel Parygin is a Ph.D. candidate in experimental high energy physics at the National Research Nuclear University MEPhI. He received his honors diploma in semiconductor electronics and physics of semiconductors at the National University of Science and Technology MISiS. He is a member of the CMS collaboration at the CERN and currently leading the operations group of the Hadron Calorimeter of the CMS experiment.

Seth I. Cooper is a research scientist at the University of Alabama, where he also did postdoctoral research. He received his B.A. from Carleton College in Physics and Computer Science and his Ph.D. from the University of Minnesota. Based at CERN, his current research focuses on data acquisition and online monitoring of detector systems, in addition to searches for physics beyond the standard model. He received the CMS Achievement Award in 2014.



David Yu completed his Ph.D. in Physics at the University of California, Berkeley in 2015. He is currently a senior research associate at Brown University and a distinguished researcher at the Large Hadron Collider Physics Center of the Fermi National Accelerator Laboratory. He conducts experimental high energy physics research at the CMS experiment at the Large Hadron Collider at CERN.

Jay Dittmann received his Ph.D. in Physics from Duke University, North Carolina, USA in 1998. He is currently a professor of Physics at Baylor University, engaged in experimental high energy physics research using data collected by the CMS experiment at the Large Hadron Collider at CERN in Geneva, Switzerland. He is a member of the American Physical Society.

Christian W. Omlin has been a professor of Artificial Intelligence at the University of Agder since 2018. He has previously taught at the University of South Africa, University of the Witwatersrand, Middle East Technical University, University of the South Pacific, University of the Western Cape, and Stellenbosch University. His expertise is in deep learning with a focus on applications ranging from safety to security, industrial monitoring, renewable energy, banking, sign language translation, healthcare, bioconservation, and astronomy. He is particularly interested in the balance between the desire for autonomy using AI technologies and the necessity for accountability through AI imperatives such as explainability, privacy, security, ethics, and artificial morality for society's ultimate trust in and acceptance of AI. He received his Ph.D. from Rensselaer Polytechnic Institute and his MEng from the Swiss Federal Institute of Technology, Zurich, in 1995 and 1987, respectively.