

Received 2 August 2023, accepted 14 September 2023, date of publication 18 September 2023,  
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3316600



## RESEARCH ARTICLE

# TransSiamese: A Transformer-Based Siamese Network for Anomaly Detection in Time Series as Approach for Fault Location in Distribution Grids

JAVIER GRANADO FORNÁS<sup>ID</sup><sup>1</sup>, ELÍAS HERRERO JARABA<sup>ID</sup><sup>2</sup>, (Member, IEEE),  
AND ANDRÉS LLOMBART ESTOPIÑAN<sup>ID</sup><sup>1</sup>, (Member, IEEE)

<sup>1</sup>CIRCE Technology Center Parque Empresarial Dinamiza, 50018 Zaragoza, Spain

<sup>2</sup>Department of Electronic Engineering and Communications, University of Zaragoza, 50018 Zaragoza, Spain

Corresponding author: Javier Granado (jgranado@fcirce.es)

This work was supported by the FLEXIGRID Project from the European Union's Horizon 2020 Research and Innovation Programme under Grant 864579.

**ABSTRACT** Fault localization in distribution grids represents a crucial aspect in achieving the concept of smart grids within electrical networks. Despite the existence of various approaches to address this issue, it remains an open and challenging topic in real situations and, therefore, complex grids. One of the main challenges stems from the scarcity of labelled real-world examples is due to the inherent chaotic nature of faults (short circuits between a phase and ground). Obtaining sufficient fault examples for neural network training purposes becomes extremely difficult. Efforts have been made to simulate fault signals and apply data augmentation techniques. However, for fault location specifically, classical augmentation techniques are not applicable due to the unique nature of the fault signals. In this paper, we propose a novel approach to address the problem of extracting fault location information from TDR (Time Domain Reflectometry) signals. This kind of signals, involve injecting pulses into the grid and record these pulses as a bounced signal due to the different impedance changes of the grid. Our approach relies on employing a Transformer for anomaly detection. This Transformer-based Siamese network architecture (abbreviated by TransSiamese) is inspired by TranAD model. This Transformer has been modified to be trained in a Siamese way with a few examples (pre-fault signals/normal state and fault signals/abnormal state). After training phase, we can run inference mode by feeding it signals representing faulty grid states (fault signals). The Transformer attempts to predict the evolution of the signal, however, from a certain point to the end, the predicted fault signal deviates from the True signal supplied. Thanks to the intrinsic property of the Siamese network, it dampens the differences between the learned and the current signal when it comes to pre-fault; on the contrary, it enhances these differences when we are immersed in a fault signal.

**INDEX TERMS** Artificial neural networks (ANNs), deep learning, transformers, siamese neural networks, anomaly detection, fault location, transmission lines.

## I. INTRODUCTION

The current challenges faced by modern society have led to an increased electrification of the energy system with a greater emphasis on distribution networks. Therefore, it is important to prepare these infrastructures for future challenges, such as the installation of multiple distributed generation resources,

or to achieve the goal of implementing the smart grid concept in electrical networks.

In this context, the automatic detection, classification, and location of faults, is an essential requirement and highly beneficial as it saves time and money compared to manual methods.

Self-repair mechanisms [1] require that this automatic methods are being incorporated to minimize the impact of power outages [2], [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Janmenjoy Nayak<sup>ID</sup>.

Faults are often due to short circuits between one or more phases to ground, which can affect the entire network. The use of time domain reflectometry (TDR) [5], [6], [7], [8] is a physical principle used to obtain signals by injecting pulses into the network. This periodic injection of pulses allows the electrical response of the network to be frequently updated [9], [10], [11]. When a fault occurs, the returned signal carries embedded information about the status of the distribution grid [12].

The detection, classification, and localization of faults in power grids remain prominent and essential research areas within the field of electrical engineering, garnering continued attention and significance.

On the other hand, classification and linear regression are two fundamental methodologies within the field of machine learning. Classification is employed when the objective involves assigning instances or examples to predefined classes, entailing the prediction of categorical or discrete label variables. Linear regression is utilized when the aim is to predict a continuous or numerical variable. The primary objective is to ascertain a linear relationship between the input features or variables and the target variable.

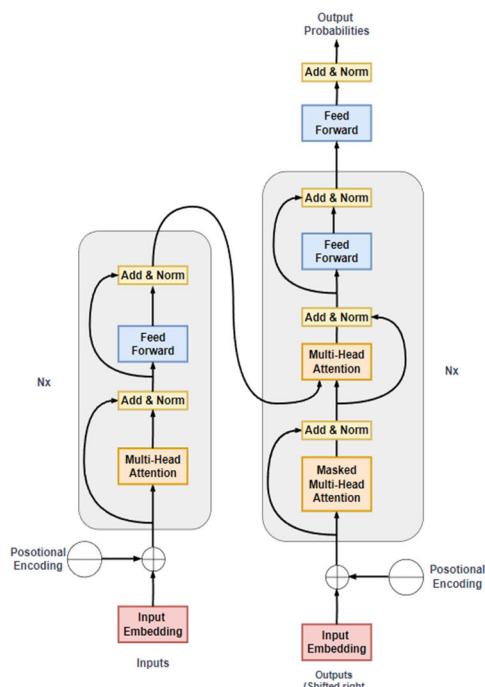
In summary, classification focuses on assigning instances to predefined classes, whereas linear regression is employed to predict numerical values. Both approaches employ specific techniques and algorithms tailored to their respective objectives and are applied in distinct contexts and problem domains within the realm of machine learning.

Detection and classification of faults falls in classification problems. Those problems involve determining the type of fault that occurred based on various types of signals, e.g., those generated by the grid response to pulse injection (TDR). One effective method for fault detection and classification is the use of neural networks, which requires a large database of labelled examples for training. However, due to the chaotic and catastrophic nature of faults, it is difficult to obtain such a database, and data augmentation techniques are often used.

In a previous study [13], the authors proposed a pulse injection solution based on the TDR technique, where a real network was modelled, and characteristic fault types were simulated. To avoid the lengthy simulation process, the authors generated a database of 200 signals examples [14] and synthesized the rest up to 10,000 using Generative Adversarial Networks (GANs). This synthetic example generation technique is useful for data augmentation in fault classification problems.

However, for fault location, which falls within the category of linear regression problems, the use of GANs cannot be applied as it is not possible to train them to generate different distances to the fault. Moreover, traditional data augmentation techniques involving transformations, or the addition of noise are also not applicable due to the temporal embedded information dependence of the signals.

Considering all these limitations, the authors propose the use of anomaly detection between pre-fault and fault signals. These signals have the characteristic that they are practically



**FIGURE 1. Structure of the first transformer (2017).**

identical until the point where the fault occurs. From that point onward until the end of the signal, the signals become different, resulting in a continuous anomaly from the fault occurrence point to the end. This type of anomaly is not present usually in the datasets used for anomaly detection in the state-of-the-art (SOTA). Therefore, this characteristic necessitates a review of the different models used for anomaly detection.

In this article we are faced with a dichotomy defined by a problem of anomaly detection to locate electrical faults in a grid, and on the other hand, the need to address a problem different from the one normally considered in this field of detection.

In order to address this dual requirement, a transformer-based solution has been chosen [15], [23]. A type of network that has demonstrated its high performance in applications with time series data, and whose inner encoder-decoder distribution is shown in Fig. 1.

Transformers learn to reconstruct signals they are trained with. The authors claim using the pre-fault and fault signals to train the Transformer. The pre-fault is the signal of the electric network's response to the injection of the pulse (TDR) when the grid is operating normally, and fault signal is the response when the grid is in fault state [7]. The propagation and reflection of TDR signals through an electrical network are very similar as long as there are no significant impedance changes in the network (faults). The transformer-type neural network is trained to model the pre-fault (normal state) signals and thus, can detect anomalies in the fault signals. After training phase, when a fault signal is presented to the trained network,

the Transformer attempts to predict its evolution. Since the fault signal has an anomaly caused by the fault effect, the Transformer will fail to predict the evolution of the signal at some point and from then on.

Using this concept, we can detect this anomaly and make an approximation to the temporal zone where the anomaly occurred. Overall, the proposed method shows promise for fault location in power lines.

Having said that, our main objective is to undertake a first approach to automatic fault location using anomaly detection techniques between the pre-fault and fault signals. Specifically, a study of time series (pre-fault and fault signals) based on Transformers is conducted.

## II. RELATED WORK

Having explained the context in which this work will be situated, it is worthwhile to expand on previous work done by other researchers. It will be divided into two separate sections:

First, we will present solutions that have addressed the problem of anomaly detection using techniques other than those proposed in this article. Since they do not fall into an unique category, we will refer to them as the General case.

On the other hand, we will focus on solutions based on Transformers. In this second section, we will justify the inspirations drawn from these articles that have led us to propose the TransSiamese model developed in this article.

### A. GENERAL APPROACHES

Within a generalist scope there are a multitude of solutions in the anomaly detection environment, and an adequate filter must be determined when presenting significant work in a single article that is not focused on being a literature review. This filter will be based on presenting different approaches prioritizing the date of publication and the variability of different approaches.

A first example of one of these approaches is the discord search, which aims to identify abnormal patterns within the data. LRRDS (Local Recurrence Rate based Discord Search) is a notable framework proposed to tackle this task. LRRDS leverages recurrence plots derived from the original time series data to accurately detect discords. To improve efficiency, innovative strategies such as optimizing pairwise distance comparisons between subsequences [16] or utilizing models like SES-AD [17], which project raw sequences into a lower-dimensional space, have been explored.

The challenge of detecting anomalies in large and complex systems with factors like the lack of ground truth labels, collective occurrence of anomalies, and high-dimensional data is addressed by the CSCAD model [18]. This model combines a graph convolutional network and a variational autoencoder to exploit feature space correlation and improve sample reconstruction for anomaly detection.

In the context of spacecraft anomaly detection, an unsupervised anomaly detection approach has been proposed. This method [19] employs a Gated Recurrent Unit (GRU)-based Recurrent Neural Network (RNN) with Extreme Value

Theory (EVT). A two-layer ensemble learning-based predictor framework is developed to learn the normal behaviour of multiple data channels, allowing the detection of anomalies in spacecraft telemetry.

Another category of anomaly detection models is based on Long Short-Term Memory (LSTM) methods. For example, in [20] LSTMAD learns structural features from normal training data and uses a statistical strategy based on prediction errors for anomaly detection. Experimental evaluations demonstrate the superior accuracy of LSTMAD in detecting anomalies compared to existing approaches, making it particularly effective in applications like ECG datasets and real-world data analysis.

MFAD (Meta-feature-based Anomaly Detection) is an alternative approach for time series anomaly detection [21]. It utilizes meta-features to describe the local dynamics of univariate or multivariate sequences. Unlike traditional sliding window approaches, MFAD defines six meta-features to statistically describe the local dynamics of a 1-D sequence of arbitrary length, reducing computational complexity and demonstrating superior results in identifying abnormal states in various applications.

Autoencoders have also been explored for anomaly detection purposes. The article [22] presents USAD (Unsupervised Anomaly Detection for Multivariate Time Series), a fast and stable method based on adversarial trained autoencoders. The unsupervised learning capability of the autoencoder architecture, combined with the adversarial training and model structure, enables the efficient identification of anomalies.

Some of them (DAGMM [23]), utilizes a deep autoencoder to generate low-dimensional representations and reconstruction errors for each data point, feeding them into a Gaussian Mixture Model. Unlike previous methods, DAGMM jointly optimizes the autoencoder and mixture model parameters, striking a balance between reconstruction, density estimation, and regularization. This approach avoids the need for pre-training and outperforms state-of-the-art techniques in anomaly detection.

In the State of the Art, we can find also identification of anomalies through reconstruction probabilities. Data reconstruction is based on learning robust representations using techniques like stochastic variable connection and planar normalizing flow (OmniAnomaly [24]).

The core idea of OmniAnomaly is to capture normal patterns in multivariate time series by these representations.

Furthermore, when an anomaly is detected in an entity, OmniAnomaly can provide interpretations based on the reconstruction probabilities of its constituent univariate time series.

In the Graph Neural Network-Based Anomaly Detection (GDN [25]), the authors combine structure learning with graph neural networks and utilize attention weights to provide explanations for the detected anomalies in real sensors. It is demonstrated that this method effectively captures correlations between sensors, enabling the deduction of the underlying cause of a detected anomaly.

In [26] it is shown a contrastive autoencoder (CAE) based approach for detecting anomalies in multivariate time-series data. The autoencoder is trained using a contrastive loss function that encourages the model to learn a representation that is similar for normal patterns and dissimilar for abnormal patterns. The authors also propose a new evaluation metric, the anomaly score, to measure the degree of anomalousness of a time series. The method outperforms several state-of-the-art methods for anomaly detection in multivariate time-series data, and the authors conduct extensive ablation studies to demonstrate the importance of the contrastive loss function and the anomaly score evaluation metric.

Some of the models presented here (DAGMM, Omni-Anomaly, GDN, LSTM\_AD) are later used in this work to perform comparisons against our TransSiamese model.

As can be seen, the field of anomaly detection in time series analysis is continuously evolving, and various approaches and models are being explored to address the challenges presented by different types of data and applications.

### B. TRANSFORMERS-BASED APPROACHES

Once the state-of-the-art in anomaly detection has been presented in a general manner, in this section, we will focus our attention on those models that use Transformers and have served as inspiration for developing our TransSiamese model.

Everything started with a paper called “Attention Is All You Need,” published in 2017 [27], where was introduced an encoder-decoder architecture based on attention layers, which the authors called the Transformer (Fig.1). Time-series anomaly detection has been successfully implemented using transformer-based models due to their ability to model long-term dependencies [15], [28]. Long-term dependencies occur in time-series data when past events have an influence on future events, such as in the case of financial data.

Transformers model long-term dependencies using the self-attention mechanism. That mechanism enables the model to focus on critical parts of the input sequence while generating the output sequence. The mechanism allows the model to learn which parts of the input sequence are relevant for predicting the output and thus capture long-term dependencies.

To detect anomalies in time-series data, transformers architecture is commonly used since a Transformer is a neural network that learns to reconstruct the input data. In such a way that the model is trained on normal data under standard conditions, and the reconstruction error is used to measure the anomaly. Data points or tiny temporal regions that deviate significantly from normal behavior are considered *anomalies*.

Several transformer-based models have shown promising results for anomaly detection in time-series data, and future research in this area is expected to yield even more promising results. Models such as Gated Recurrent Transformers (GRT) [29], Multi-Head Attention Networks (MHAN) [30], [31], [32], [33], Deep Transformers [34], [35], Graph Convolutional Transformers (AGRU) [36], [37], [38].

In addition, the literature includes publications that propose different transformer-based frameworks for detecting

anomalies in multivariate time-series data. For instance, in [39] authors present a novel framework utilizing the transformer encoder architecture for the purpose of learning representations of multivariate time series. (GCT) [40], [41], [42], and Attention-based Gated Recurrent Units. The model incorporates an unsupervised pre-training scheme, which demonstrates substantial performance benefits compared to fully supervised learning on downstream tasks. Notably, these advantages are observed even in the absence of additional unlabeled data, as it effectively leverages existing data samples. By evaluating on multiple publicly available datasets, good performance in regression and classification tasks is proven [26], [36], [37], [38], [39].

Of all the models that we have reviewed from the state of the art, a first source of inspiration for this article appeared in [28]. The authors propose the *Anomaly Transformer*, a method that incorporates an association discrepancy term into the transformer loss function to capture both temporal and contextual dependencies in the data. The authors also propose a new evaluation metric, the anomaly detection rate (ADR), to measure the performance of their method on different types of anomalies. The method outperforms several state-of-the-art methods for anomaly detection in time-series data, and the authors conduct extensive ablation studies to demonstrate the importance of the association discrepancy term and the ADR evaluation metric.

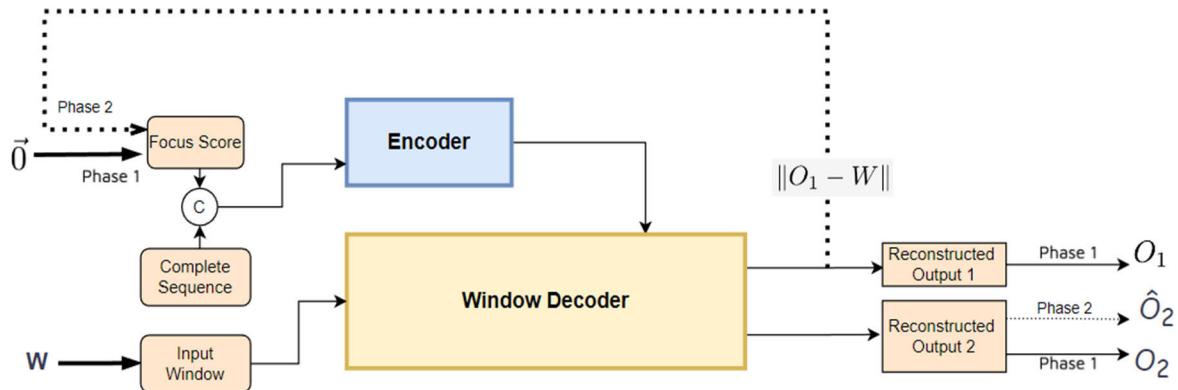
Here, our inspiration is based on the fact of using a new Anomaly-Attention mechanism that takes advantage of the short-term history of a few tens of samples of the signal. The encoder treats this history and passes it to the decoder to join it with the signal arriving at the current time, which improves the unsupervised detection of anomalies in time series.

Our second source of inspiration was the feedback loop the authors propose in [43] (TranAD) for focus scoring and adversarial training, specially the first one.

With this mechanism, we can highlight the small differences we have in our signals. The model, (Fig.2) uses a transformer-based encoder-decoder architecture that is designed to capture both temporal and cross-dimensional dependencies in the data. This architecture is trained to reconstruct the input data, and the reconstruction error is used as a measure of anomaly. Additionally, the authors propose a new unsupervised training algorithm that leverages the properties of the transformer architecture to learn a representation that is sensitive to anomalies in the data. They evaluate their method on several benchmark datasets and show that it outperforms several state-of-the-art methods for anomaly detection in multivariate time-series data.

The use of focus score-based auto conditioning, enable robust multimodal feature extraction and, adversarial training help gain stability. With this feature we can highlight small differences in our signals.

Furthermore, the adoption of model-agnostic meta learning (MAML) enables training the model with limited data availability.



**FIGURE 2.** Transformer (TranAD) model with feedback used to improve the encoder response inspired by anomaly transformer [28].

While simple transformer models may overlook anomalies characterized by subtle deviations, the introduction of the adversarial training procedure allows for the amplification of reconstruction errors. The adversarial training process consists of two distinct phases. In the initial phase, the model aims to generate an input, known as the focus score, serves as a crucial element in guiding the attention network within the Transformer Encoder. This enables the extraction of temporal trends. Subsequently, in the second phase, the output is conditioned based on the deviations derived from the first phase.

The focus score, generated during the first phase, serves as an indicator of the discrepancies between the reconstructed output and the provided input. It operates as a prior knowledge, influencing the modification of attention weights during the subsequent self-conditioning phase. approximate reconstruction of the input window. The deviation between the inferred reconstruction and the original.

Nevertheless, although the use of the focus score is maintained in our work, the adversarial training method was not implemented due to the poor results obtained. Instead, the origin of the third inspiration is the Siamese Networks [44], [45], [46], [47], [48], [49]. Thanks to the main characteristic, which reduce the differences between similar examples and increase the differences between different examples, we can train the Transformer to learn to model our signals better and thus improve the detection of anomalies.

In summary, all this path discussed in this section of the article has led us to a solution that works particularly well with the nature of the data handled in our context, that uses certain novel ideas from other authors, and that presents a novel solution in the anomaly detection environment. Thus, we can summarize our main contributions as follows:

- 1) We propose a novel Transformer-based approach, TransSiamese, which has demonstrated his ability to work as anomaly detector, but even as a viable and potential approach to the problem of fault localization.
- 2) A Siamese scheme has been used to enhance the powerful characteristics of this network to enhance the

minimum differences between two signals to be compared (thanks to the Contrastive-based learning used), damping the differences between the learned and the current signal when it comes to pre-fault; and enhancing these differences when we are immersed into a fault signal.

- 3) Together a methodology to overcome the inherent issue of scarcity of real-world examples in such linear regression problems, where traditional Data Augmentation techniques may not be feasible due to the signal nature. And finally, in inference mode, we use an *Energy based Anomaly Score function* and propose a Weighted-energy based Anomaly Score function, which enhances the differences between the two-time sections of the signal and boosts the Siamese performance: the normal conditions section and the one corresponding to the time just after the failure occurs.

### III. METHODOLOGY

This section outlines the proposed methodology which utilizes the injection of signals into the distribution line (TDR).

These signals (pre-fault and fault signals) are then used to detect anomalies through the application of the *TransSiamese* model.

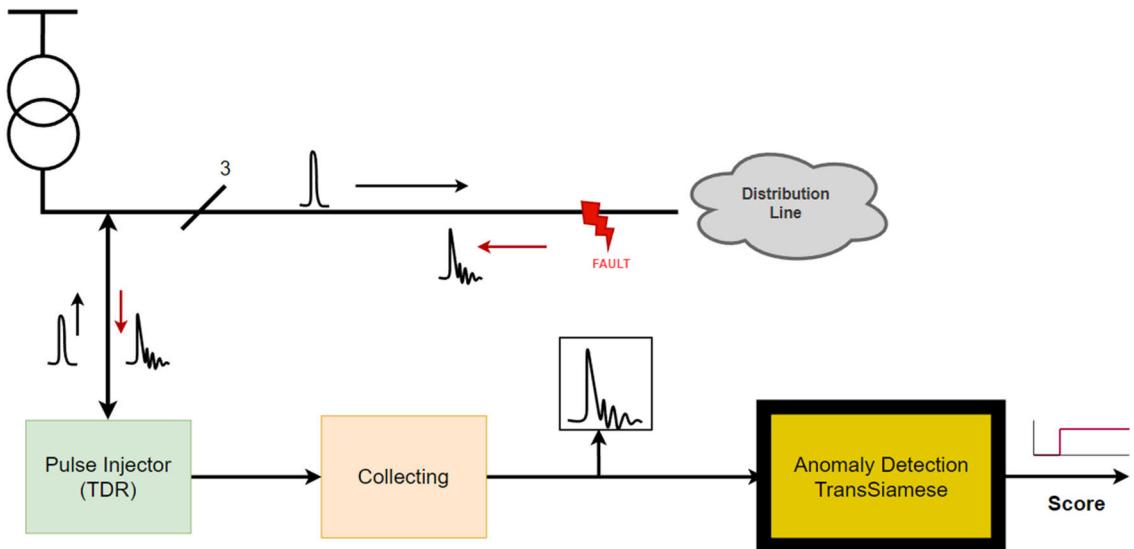
The following subsections will provide a detailed explanation of the methodology, as illustrated in Fig. 3.

#### A. COLLECTING PHASE: TDR SIGNALS

According to the physics of transmission lines, these pulses travel through the network and are reflected at every bifurcation. Therefore, some of them are bounced back [13].

The magnitude of these reflections depended on the impedance of the line at each bifurcation.

Therefore, the initial step involves generating a database [14] of signals using (PSCAD<sup>TM</sup>) software. In order to do so, a realistic electrical network was modelled, and five types of faults were simulated at various points along the distribution line. These fault types include short circuits between R, S, and T to Earth, as well as short circuits between R-S to Earth and R-S-T to Earth.



**FIGURE 3.** Diagram of connection to the grid on fault state together with the transmission of the significant injection pulses.

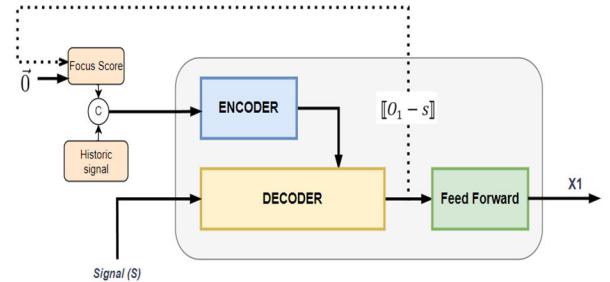
The TDR technique, which involves injecting short-period pulses ( $\sim 10$  ns) into each of the three phases of an electrical grid every few seconds (e.g., every 5 s), was utilized. As per the principles of transmission lines, these pulses travel through the network and are reflected at each bifurcation, with the magnitude of the reflections being influenced by the impedance of the line at each bifurcation.

#### B. ANOMALY DETECTION: TRANSFORMERS

Unsupervised anomaly detection in time series remains a challenging problem, particularly when dealing with time series that exhibit complex dynamics. While some previous methods have attempted to address this challenge through pointwise representation or pairwise association learning, they do not perform well in such contexts. Recently, Transformers have emerged as a powerful approach to model anomalies in time series, owing in part to the rarity of anomalies. That makes building associations between abnormal points and the rest of the series particularly difficult.

Real-world systems often work in a continuous manner, generating successive measurements monitored by sensors in our case, the measurements come from the response of the network to injection pulses (TDR). Those signals are successive reflections caused by changes in impedance throughout the network. Detecting any anomaly between two consecutive signals (pre-fault and fault) is reduced to detecting abnormal points within the two time-series. However, anomalies are usually very rare processes and are hidden among a large number of normal points within the time series.

In the case of fault location, labelling the distance to the fault (the time at which the first anomaly occurs between both time series) is very difficult to obtain, as real examples are generally chaotic events that cannot be easily reproduced or measured in the field due to their nature. In the other side,



**FIGURE 4.** Transformer with focus score based- self conditioning.

simulating these phenomena is also complex, as simulation times make it unfeasible to have a large enough number of examples to train a supervised system. And similarly, we have problems applying data augmentation techniques due to the nature of the signals. For all the above reasons, we will focus on detecting anomalies in time series under the premise of unsupervised training.

Having said that, we use a Transformer model with focus score-based self-conditioning (Fig. 4) in this work. The Transformer requires two inputs, the first input used by the Encoder and the second by the Decoder. As used in recent work based on the Anomaly Transformer model, the Encoder oversees the short-term temporal information (Historic signal,  $HS$ ), and the Decoder is in charge of combining the historic information with the current signal being processed. It should be noted that  $HS$  is constructed by joining the previous  $D$  signals to the current one (in our tests, the  $D$  parameter was set to the integer value of 10, since a variation of it did not cause significant changes in the performance of the whole system).

Therefore, in an initial state the  $HS$  signal history is concatenated with a zero vector and presented to the Encoder

in order to obtain a first approximation of the input signal. Then, the deviation observed from this first reconstruction, termed as the focus score, serves to facilitate the attention network within the Transformer Encoder in a second iteration. In this sense, this attention network is responsible for extracting temporal trends by focusing on sequences with high deviations. Subsequently, in this new iteration known as *self-conditioning* (dashed line) in [43], the focus score derived from the first iteration indicates the disparities between the reconstructed output and the provided input.

This modification proposed in [43] aims to extract short-term temporal trends, by enabling higher neural network activation for specific input sub-sequences.

Finally, the observation  $O_1$  returned by the Decoder in the second phase is processed by a single Feed Forward network from which the representation of the input signal  $S$  is finally obtained.

### C. TransSiamese MODEL

The main characteristic of *TransSiamese* architecture is their ability to efficiently capture and model complex relationships between pairs of input sequences thanks to the use of Siamese structure during training. That is why two identical copies of the Transformer architecture, known as “Siamese branches” are employed. Each branch takes a different couple of inputs and processes them independently. The obtained representations from both branches are then compared to perform tasks such as similarity comparison, pair classification, or feature extraction. Furthermore, the Siamese architecture allows for joint learning and knowledge sharing between the branches, which can enhance model efficiency and generalization. This is indeed our chosen alternative, and we think it is key to the performance shown in the results section.

Fig. 5a shows the operating diagram of the Siamese model used in this article, in which the two transformers are shown, one being a copy of the other. The calculation of the loss in (1) is also shown and will be explained later.

At the end, anomalies in the tested signals do not occur in a punctual way in time. Both types of signals (pre-fault and fault) are very similar up to the point where the fault occurs, after which the signals continue to be different. Therefore, we need to think of a solution that maximises the differences in both types of signals as a whole, rather than at specific points in time.

This characteristic raise from the use of the contrastive loss function (1) during the training phase. This adaptation improves the training process of the model and contributes to its overall performance. The contrastive loss aims to minimize the Euclidean distance between pairs of similar instances (Pre-fault signals) and, at the same time, maximize the distance between pairs of dissimilar instances (pre-fault vs fault signals).

In this way, it favors the model to discriminate effectively between different classes or categories. This promotes better separation and clustering of the data and trains the Transformer more effectively to predict similar time series such

as faults and pre-faults and achieve better anomaly detection. In short, we believe that the transformer is able to use the loss provided by the contrastive method to enhance the differences after each fault. In fact, it would be the fastest and most intuitive way of wanting to move away a pattern of one signal from another of a different type.

Finally, equation (1) is shown for a better understanding about all commented above in this section.

$$L = (1 - y) \cdot \|PF - F'\|^2 + y \cdot (M - \|PF' - F'\|^2) \quad (1)$$

where, if  $y$  is 1 (when fault and non-fault signals are used), the left-hand term disappears, and we attempt to maximize the distance between examples until some threshold  $M$ . On the other hand, if  $y$  is 0 (when both signals correspond to faults), the right-hand term disappears, and we attempt to minimize the distance between examples.

### D. ANOMALY SCORE FUNCTIONS

On the other hand, since the results obtained from the TransSiamese network (inference mode) have to be exploited in the best possible way, we want to present a series of measures that aim to optimize and boost the output provided by the network (Fig. 5b). We have defined three Anomaly Score functions: the MAPE approximation, the Energy, and the Weighted-energy.

The simplest is the one which joins together a Euclidean distance with the MAPE function (here it is referred to as Standard). In the second one we make an estimation of the energy because it has been identified, as seen in (Fig. 6), that from the fault onwards the signals are seen with higher energy. And in the third point we also propose the Weighted-energy because we see that the difference in energy depends on the amplitude over the time domain of the signal.

-Standard function (Std) is defined as (2)

$$\text{Std} = \left\| \frac{(x_1 - x_2)}{x_1} + \sqrt{(x_1 - x_2)^2} \right\| \quad (2)$$

It is equivalent to the function MAPE (mean absolute percentage error) + RSME (root mean squared error). Where  $x_1$  and  $x_2$  are the outputs of the Transformer in Inference mode.

-The Energy function is defined as (3)-(5).

This function computes the energy over a window ( $w$ ) period of each signal ( $x$ ) generated by the Transformer and calculate the difference ( $D$ ) between them.

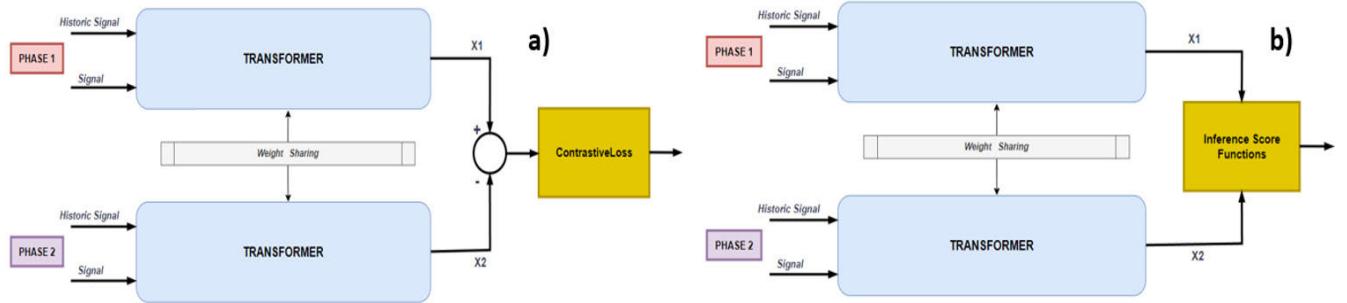
$$E_{x_1} = \int_0^w |x_1(t)|^2 dt \quad (3)$$

$$E_{x_2} = \int_0^w |x_2(t)|^2 dt \quad (4)$$

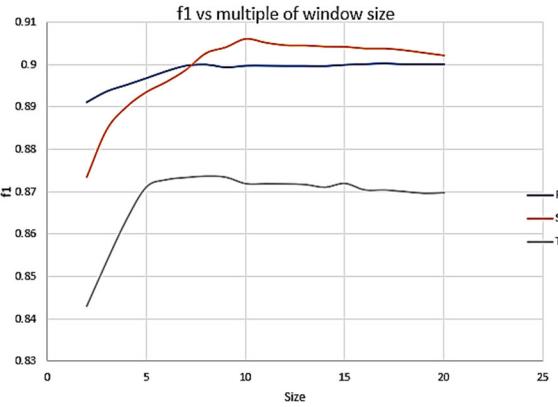
$$D(E_{x_1}, E_{x_2}) = \|E_{x_1} - E_{x_2}\| \quad (5)$$

-The Weighted-energy function is defined as (6)-(9).

This function computes the energy over a window ( $w$ ) period of each signal ( $x$ ) generated by the Transformer and



**FIGURE 5.** a) Anomaly detection TransSiamese model during the Training phase, and b) Integration with the Anomaly score function for Inference phase.



**FIGURE 6.** F1 vs multiple window size.

calculate the difference (D) between them.

$$E_{x_1} = \int_0^w |x_1(t)|^2 dt \quad (6)$$

$$E_{x_2} = \int_0^w |x_2(t)|^2 dt \quad (7)$$

$$Dw(E_{x_1}, E_{x_2}, X) = \frac{\|E_{x_1} - E_{x_2}\|}{X} \quad (8)$$

$$X = \{x_{max} : \max_{i \in W_m} x_1(i)\} \quad (9)$$

$max$  is computed in windows of size =  $s \times \text{multiple}$ . For its calculation we have chosen to use the pre-fault signal ( $x_1$ ).

Finally, we want to show a couple of figures illustrating a couple of important characteristics to be understood. Fig. 6 shows the variation of f1 as a function of a multiple (sigma), corresponding to the size of the window for calculating the maximum in relation to the window used for the energy calculation. It shows that around  $\text{multiple}=10$  the f1 output stabilizes. We use this multiple to calculate the *Weighted-energy Anomaly Score*.

Fig. 7 shows the anomaly score obtained in Inference Phase. The automatic Threshold (AT) is marked as a red line. The output of the *Weighted-energy* function is computed by means of the Threshold to decide if this point is an anomaly.

In training mode, we use the ContrastiveLoss (1) as loss function. (Algorithm 1):

### Algorithm 1 The TransSiamese Train Algorithm

#### Require:

Encoder  $E$ , Decoder  $D$

Signals  $W_1$  and  $W_2$

Short-end historic signals  $W_{h_1}$  and  $W_{h_2}$

Number of Epochs  $N$

Evolutionary hyperparameter  $\epsilon$

Number of pairs for Siamese  $T$

1: Initialize weights  $E, D$

2:  $n \leftarrow 0$

3: **do**

4: **for**  $t = 1$  to  $T$  **do**

5:  $x_1 \leftarrow D\left(E\left(W_{h_1}, W_1, \vec{0}\right)\right)$

6:  $x_1 \leftarrow D\left(E\left(W_{h_1}, W_1, \|x_1 - W_1\|^2\right)\right)$

7:  $x_2 \leftarrow D\left(E\left(W_{h_2}, W_2, \vec{0}\right)\right)$

8:  $x_2 \leftarrow D\left(E\left(W_{h_2}, W_2, \|x_2 - W_2\|^2\right)\right)$

9:  $\text{Loss} \leftarrow L_{\text{cnt}}(x_i, x_j, \theta) = 1[y_i - y_j] \|f_\theta(x_i) - f_\theta(x_j)\|_2^2 +$

$1[y_i \neq y_j] \max(0, \epsilon - \|f_\theta(x_i) - f_\theta(x_j)\|_2)^2$

10: Update weights of  $E, D$  using Loss

11:  $n \leftarrow n + 1$

12: **while**  $n < N$  **do**

In inference mode, we use the *Anomalies Scores functions* (ASF) and *Calculate Automatic Threshold* (AT)(Algorithm 2):

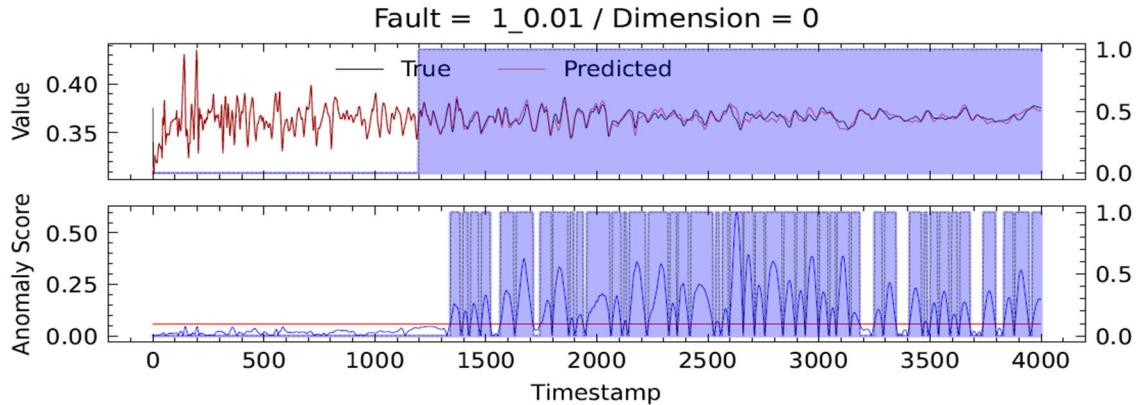
The ASF (Anomaly Score Function) could be one of the three defined in this work. *Standard*, *Energy* or *Weighted-energy*.

## IV. EXPERIMENTS & RESULTS

This section describes the methodology used in our experiments. An introduction is made to comment the configuration of the examples in the database, specifically how they have been selected for the different phases of the training (training, validation, and testing). Then, the results obtained for each case are presented.

### A. STRUCTURE OF THE EXPERIMENTS

In this study, we have used an existing database [14] containing 200 signals, which represent the grid's response to injected pulses using Time Domain Reflectometry (TDR) [7].



**FIGURE 7.** Example de anomaly visualization between the score of the true and predicted signal. Red line is the calculated umbral (AT).

## Algorithm 2 The TransSiamese Test Algorithm

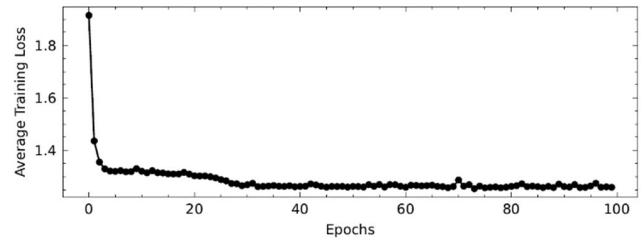
**Require:**

Encoder  $E$ , Decoder  $D$   
Test Dataset Signals  $W_1$  and  $W_2$   
Short-end historic signals  $W_{h1}$  and  $W_{h2}$   
Number of pairs for Siamese T  
1: **for** ( $t = 1$  to  $T$ )  
2:  $x_1 \leftarrow D(E(W_{h1}, W_1, \vec{0}))$   
3:  $x_1 \leftarrow D(E(W_{h1}, W_1, \|x_1 - W_1\|^2))$   
4:  $x_2 \leftarrow D(E(W_{h2}, W_2, \vec{0}))$   
5:  $x_2 \leftarrow D(E(W_{h2}, W_2, \|x_2 - W_2\|^2))$   
6:  $s \leftarrow ASF(x_1, x_2)$   
7:  $AT \leftarrow$  Calculate Automatic Threshold(s)  
8:  $y_i = 1 (s_i \geq AT)$

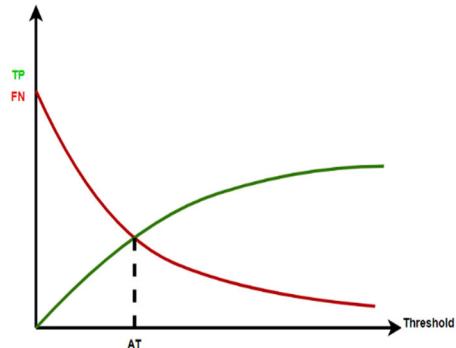
These signals were acquired using a simulator at a sampling rate of 100 Msps (Mega samples per second). Each of the three injected signals (R-S-T) was digitized for a duration of  $40 \mu s$ , after which the signal essentially vanished. Consequently, the combined sampling time for all three phases amounted to  $120 \mu s$ . With the given sampling rate, we obtained 12,000 digitized values (parameters).

As a result, the database consists of 200 signals, with each signal representing the grid's response to an injected pulse under a specific fault condition (i.e., a fault between Phase R and Ground occurring 500 m away from the device used for injection/measurement). Those 200 signals are subdivided. Into five groups related to the type of fault, so there are over 40 examples of each type.

Each fault group is defined according to the distance at which it is defined and the impedances specific to the corresponding fault. Thus, for each type of fault there are impedances of 0.01, 80.0, 150.0 and 1000.0 ohms. And each impedance is located at different distances from the injector: 12, 40, 283, 933, 757, 1833, 1274, 594, 968 and 649m.



**FIGURE 8.** Training graph after 100 epochs.



**FIGURE 9.** Automatic threshold calculated.

On the other hand, we trained the TransSiamese model in four different experiments (B-C-D-E). All of the training process had a good training result in terms of convergency (Loss), similar to the one depicted in Fig. 8, in which the number of 100 epochs is used in all experiments commented in this paper.

Following the explanation of the experiments carried out and in operational terms, the processing is based on three necessary steps: the execution of the neural system based on the Siamese network (III-C), the calculation of the anomaly score (III-D), and finally an automatic thresholding.

Specifically, and delving deeper into the third step, the automatic threshold corresponds to the point AT (Fig. 9), and

it is calculated from (10)

$$\begin{aligned} & AT \\ & = \operatorname{argmax} \left( f(\theta) := \left\{ \frac{TP(\theta)}{TP_{max}(\theta)} + \left( 1 - \frac{FP(\theta)}{FP_{max}(\theta)} \right) \right\} \right) \end{aligned} \quad (10)$$

being  $\vartheta$  a considered threshold over a set of attempts evaluated over the anomaly score values. The idea behind the equation (10) is to find an optimum point in which we can minimize the False Negatives (FN) and maximize the True Positives (TP). TP are those points where the results indicate an anomaly that really exists, and FN are those points where the result indicates an anomaly, but it does not really exist. After the AT threshold has been calculated, the results above the threshold are labeled as anomalies and the results below the threshold are labeled as “normal” point.

In all experiments we will show 3 different summaries of results and provide relevant conclusions for each experiment. On the one hand, we will show a summary statistics table where we will collect the f1, precision, recall and overall ROC/AU data obtained over the total number of tests performed. On the other hand, we will show a comparison of the f1 results for the 3 scoring functions discussed in section III-D. And finally, we will show an example of processing and thresholding of a complete signal from the time the pulse appears, the fault occurs, and a sufficient time where it is diluted over time.

In general terms, we notice that different experiments achieve similar performance on all the four-evaluation metrics. The *Weighted-energy* Anomaly score function shows the best performance in all the experiments. As we said, in inference mode, the difference between the *True* and *Predicted* signal must be weighted because the difference of the two signals (absolute value) is independent of the amplitude of them, so, the *Weighted-energy* is a best way to measure the Anomaly Score more fairly.

Finally, we have conducted experiments to compare the performance of our database with other state-of-the-art (SOTA) models. It should be noted that the three processes discussed in this section, III-C, III-D and automatic thresholding have been equally well executed regardless of the approach used in the comparison. In this way the results shown will be playing on equal footing with each other.

As it can be seen later in these experiments, it is evident how our model achieves significantly higher scores due to its well-adapted behavior to the nature of our signals. Similarly, we have also conducted experiments to compare the processing time in inference and training between our model and the other selected SOTA models.

Therefore, it will be able to conclude that we have managed to train a model with a single Pre-fault signal and 200 faults and we have found that training only with that Pre-fault signal and different combinations of faults the result (f1 score) has a good performance ( $>0.7$ ), testing with any ever seen example of any type of fault.

**TABLE 1. Experiment B statistics.**

	f1	Precision	Recall	ROC/AU
Standard	0.64158503	0.9387835	0.52802129	0.67264408
Energy	0.68725497	0.94739297	0.5571459	0.70070008
Energy Pond	0.77026476	0.9692443	0.6494186	0.77197252

But, before starting with the explanation of the experiments, we would like to proceed to discuss the statistics used in the experiments. On the one hand, the f1 metric is a widely used performance measure in binary classification tasks, which combines precision and recall into a single value. It provides a balanced assessment of a model’s ability to correctly classify positive and negative instances.

The f1 score is the mean of precision and recall. It provides a single metric that considers both precision and recall, giving equal importance to false positives and false negatives. It ranges from 0 to 1, where a value of 1 indicates perfect precision and recall, while a value of 0 indicates poor performance in correctly identifying positive instances. The f1 score is commonly used when there is an imbalance between the number of positive and negative instances in the dataset.

Precision is the ratio of true positive predictions to the sum of true positive and false positive predictions. It measures the accuracy of positive predictions, indicating how well the model identifies true positive instances.

Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the sum of true positive and false negative predictions. It quantifies the model’s ability to identify all positive instances correctly.

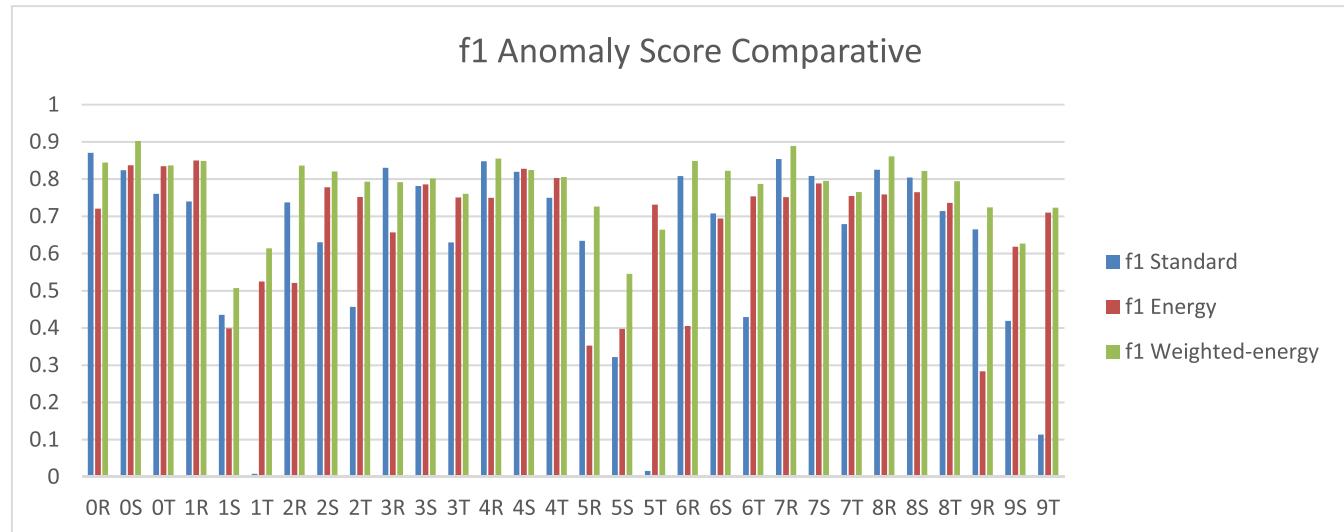
On the other hand, ROC/AUC will give information on how well the classifier distinguishes between pre-fault and fault sections. Values close to 1 define a perfect classification model, and values close to 0.5 correspond to models with a random and meaningless response.

## B. ANOMALY SCORE FUNCTIONS COMPARATIVE

In this first experiment, the idea is to test with signals from the same group (different examples for training and testing), in order to select the simplest experiment where we can clearly compare the different Anomaly Scores functions. Type 1 signals, which consist of a ground fault in phase R, have been chosen for both training and testing.

Our intention is to compare the performance of the three Anomaly Scores functions (Standard, Energy and Weighted-energy) that best fit our dataset when Pre-fault signals such as faults are known to the neural network.

Table 1 shows a significantly better performance of the Weighted-energy score in all four calculated statistics. On the other hand, it is worth noting in Fig. 10 the performance of the two energy-based scores on the 1T, 5T, and even on the 9T signal, corresponding to phase T signals in 3 different three-phase signals. And finally, the result of a fault corresponding to phase R, impedance 0.01 ohms, and at a distance of 1833m (12.23 sec) is shown.

**FIGURE 10.** Anomaly Score Graph for the same group of signals for training and testing.**TABLE 2.** Experiment C-1 statistics.

	f1	Precision	Recall	ROC/AU
Standard	0.689699343	0.945218886	0.563306952	0.6965848
Energy	0.685920359	0.946301277	0.557836452	0.697872624
Energy Pond	0.749123135	0.966526213	0.623625054	0.757293567

**TABLE 3.** Experiment C-2 statistics.

	f1	Precision	Recall	ROC/AU
Standard	0.657564425	0.928790453	0.529382265	0.669702676
Energy	0.68785823	0.92747917	0.56211502	0.67521833
Energy Pond	0.750869733	0.954086755	0.626748132	0.742819733

### C. TRAINING WITH FAULTS TYPE #1/TEST WITH ALL TYPES

For this experiment we perform two different tests. First, we compare the performance of the TransSiamese model training with Faults type #1 and testing with faults type #1 (different examples from training and test). (Table 2 / Fig. 12 / Fig. 13). In the second test, we carried out a comparison training with examples of faults type # 1 and testing with the rest of the different examples of the Database (Table 3 / Fig. 14 / Fig. 15).

Performance continues to be better for the Energy weighted score. But from this data we can draw the following conclusions:

- Fig. 12 shows the first 3 signals corresponding to an impedance of 0.01 ohms, the next 3 signals correspond with an impedance of 80.0 ohms, and the final third group of three signals corresponds to an impedance of 150.0 ohms.

- Focusing on the value of f1, the first group (0.01 ohms) has a value of 68.13%, 68.36% and 77.65% respectively with each score used. The second group (150 ohms) has a value

**TABLE 4.** Experiment D-1 statistics.

	f1	Precision	Recall	ROC/AU
Standard	0.691373549	0.921645093	0.563082093	0.664378462
Energy	0.69564127	0.925664542	0.569371442	0.670698514
Energy Pond	0.773586772	0.951525975	0.657702626	0.750835581

**TABLE 5.** Experiment D-2 statistics.

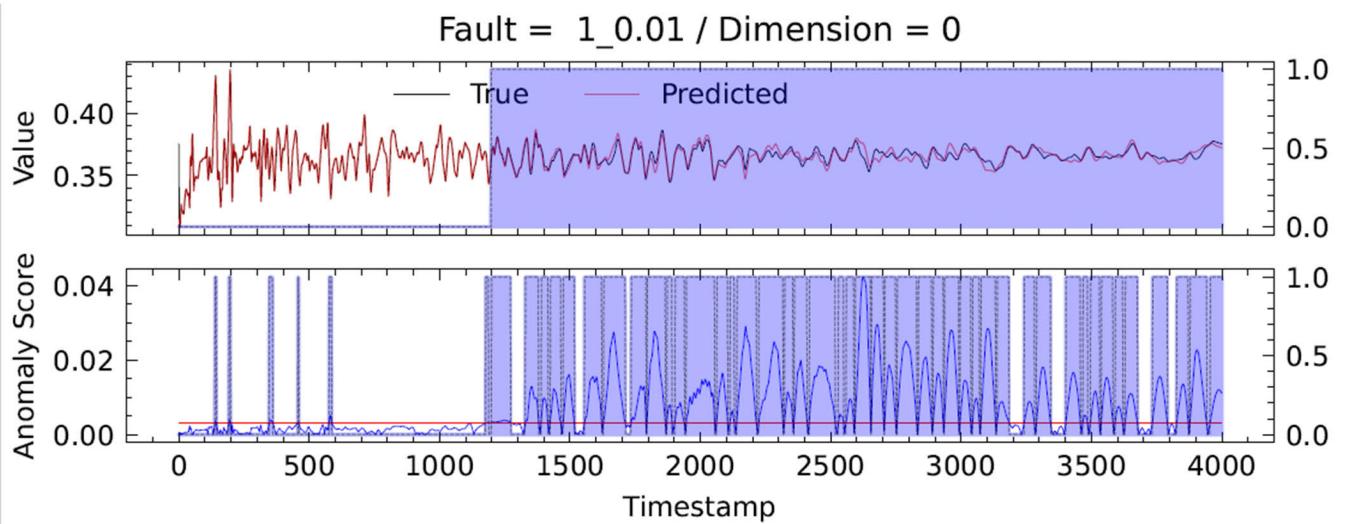
	f1	precision	recall	ROC/AU
Standard	0.659910107	0.930519634	0.533532216	0.673398363
Energy	0.683029015	0.932477107	0.556738539	0.678588514
Energy Pond	0.747073327	0.956830318	0.623118222	0.745009676

of 64.72%, 63.69%, and 69.9%. And finally, the third group (80 ohms) f1 scores are 68.21%, 65.35%, and 71.73%.

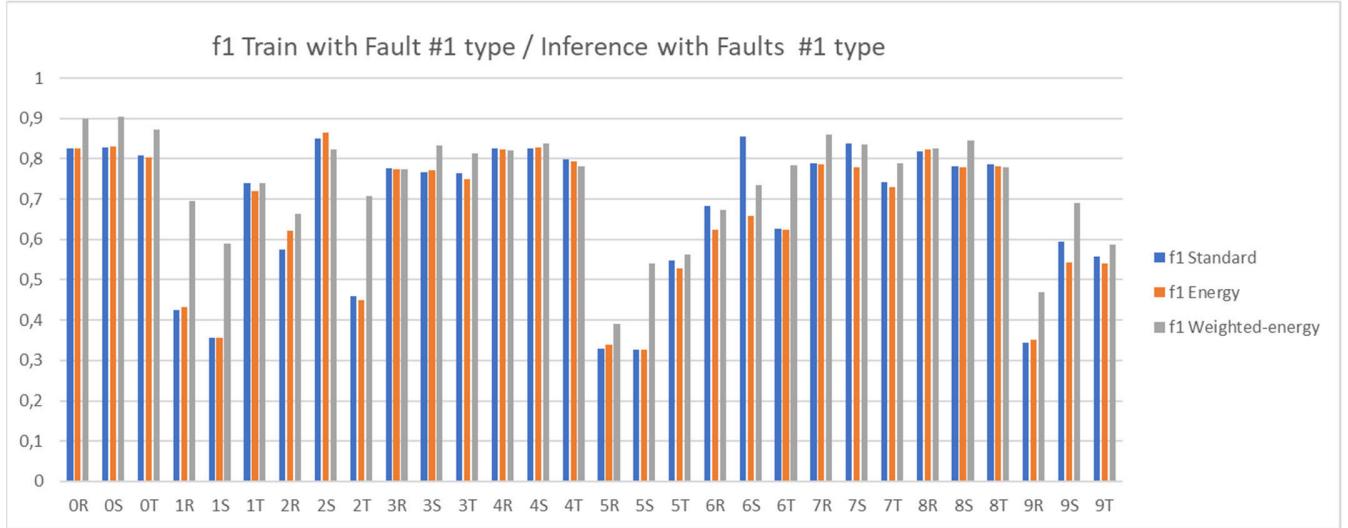
- Better performance is seen when the impedance of the fault is lower, as was logical in principle. And a worsening of the energy-based score with respect to the so-called standard. Nevertheless, the performance of the Weighted-energy remains above the other two.

### D. TRAINING WITH FAULTS TYPE #8/TEST WITH ALL TYPES

Now we carried out an experiment like the previous one C, but for the type of fault #8. For this experiment we also perform two different tests. First, we compare the performance of the TransSiamese model training with faults type #8 and testing with faults type #8 (different examples from Training and Test). (Table 4 / Fig. 16 / Fig. 17). In the second test, we carried out a comparison training with examples of Faults type # 8 and testing with the rest of the different examples of the database (Table 5 / Fig. 18 / Fig. 19).



**FIGURE 11.** Inference visualization of Anomaly Score with Weighted-energy from a Fault type #1.



**FIGURE 12.** Anomaly Score Graph for Training with Faults type #1 and Testing with Faults type #1 (different samples for training and testing).

However, unlike the previous experiment, in this case we will focus on the larger impedances to analyze their behavior. The conclusions obtained for the first experiment are:

- The value of f1 for each of the scores for an impedance of 150 ohms is 71.54%, 72.7%, and 77.31%. For the 1000 ohms case, the value of f1 would be 69.25%, 68.08%, and 77.31%.-

The impedance value plays an important role in the performance of the proposed system; and the type of fault simulated in this case (grounding of the three phases) is easier to deal with than in the case of the previous experiment.

In a second experiment, with inference test data other than the trained faults (#8), the results are very similar, except for a couple of faults, 22 and 30, in phases R and S respectively, where the energy-based score is above the other two scores. As a detail, this is a fault with an impedance of 20 ohms and 100 ohms respectively, and the spurious result is due to an

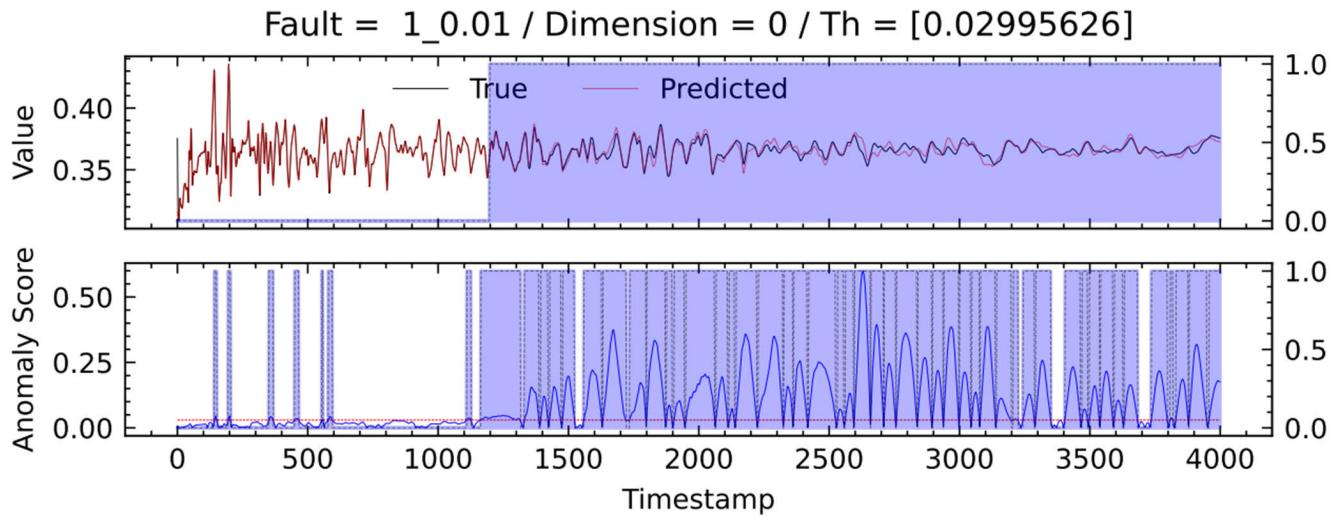
anomalous sporadic threshold behavior. Note that the colors of the graph have been changed to highlight these significant details.

#### E. TRAINING WITH ALL TYPE OF FAULTS/TEST WITH ALL TYPES

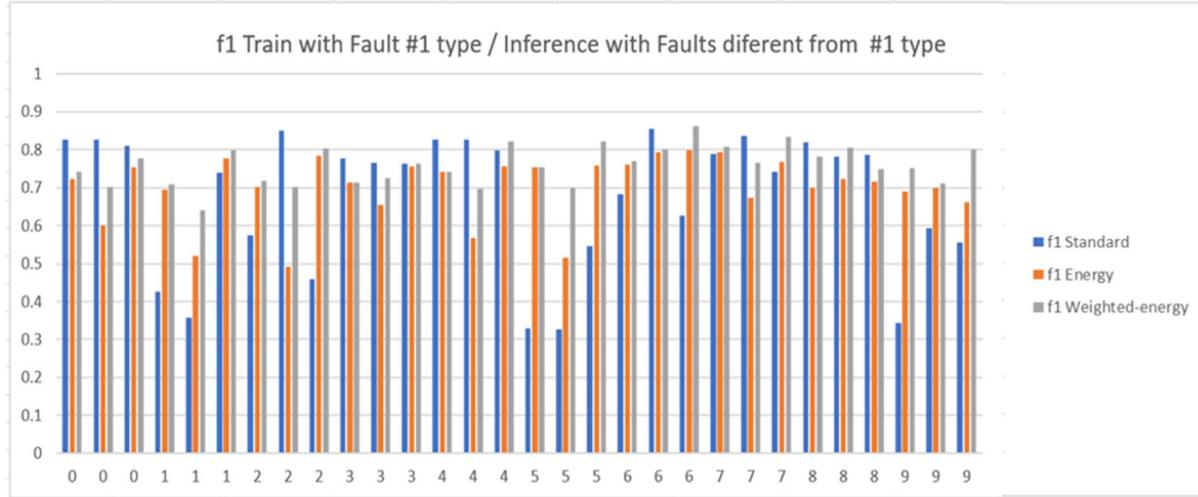
Now we carried out an experiment similar to the previous ones but with all the types of faults to train and all the types of faults to test. (Different examples from Training and Test). (Table 6 / Fig. 20 / Fig. 21).

In this last experiment, 3 faults of each impedance (0.01, 80, 150 and 1000 ohms) are used for each type of fault considered, with a total of 60 samples in the graph in Fig. 20.

It shows a greater stability in the results, and a growth in the performance of the standard and energy-based scores. But the best results for the Weighted-energy score are maintained



**FIGURE 13.** Inference visualization of Anomaly Score with Weighted-energy from a Fault type #1.



**FIGURE 14.** Inference visualization of anomaly score with weighted-energy from a fault type #1.

**TABLE 6.** Experiment E statistics.

	f1	precision	recall	ROC/AU
Standard	0,73399097	0,93714714	0,60816678	0,75515462
Energy	0,73356683	0,92937304	0,61009398	0,74988561
Weighted-energy	0,78584826	0,95209383	0,67382822	0,79679168

over the first two, highlighting the best results for this score in this last experiment. It should also be noted that there were no problems related to the computation of the threshold.

#### F. SOTA MODELS COMPARISON ANALYSIS

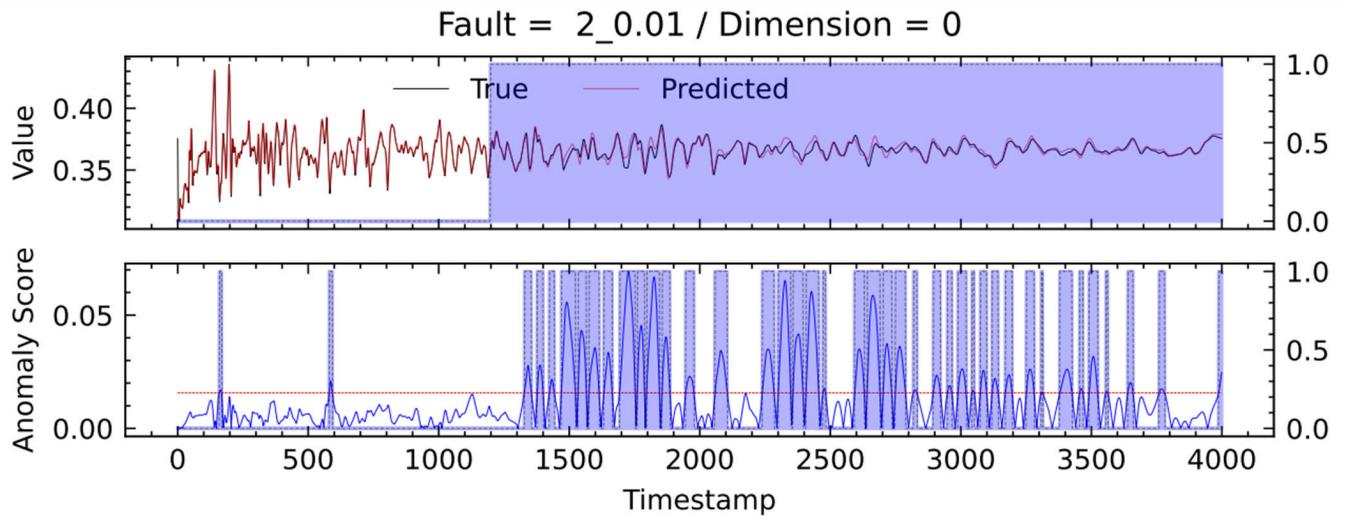
In this experiment, we compare the performance of TransSiamese model with SOTA models for multivariate time-series anomaly detection, including TranAD [43], LSTM\_AD [50],

DAGMM [23], OmniAnomaly [24], and GDN [25], training with our Dataset of 200 signals [14]. (Table 7 / Fig. 22 / Fig. 23 / Fig. 24 / Fig. 25 / Fig. 26). This comparison has been done thanks to the great work done by the authors of [40] on their public GitHub (<https://github.com/imperial-qore/TranAD>).

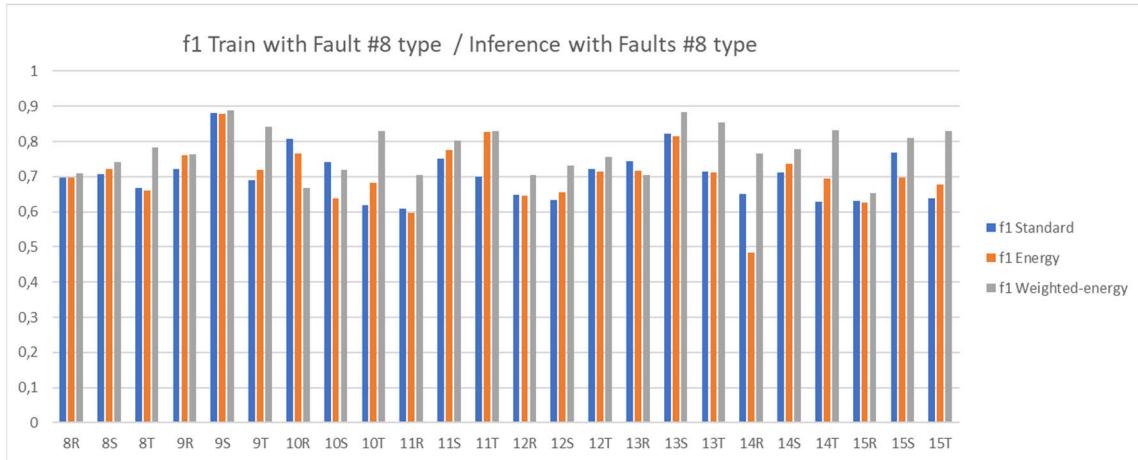
We also tested some simplified versions of TranAD (with some of the features annulled), but due to its low f1 scores, do not include the results in the comparative.

Finally, we have performed tests of the average training and inference times for all models (Table 8 / Table 9).

As previously mentioned, our database consists of temporal signals that reflect the response of an electrical network to the injection of a high-frequency pulse. These temporal series have the characteristic of continuous anomaly starting from a certain point until the end of the signal.



**FIGURE 15.** Inference visualization of anomaly score with weighted-energy from a fault type #1 example.



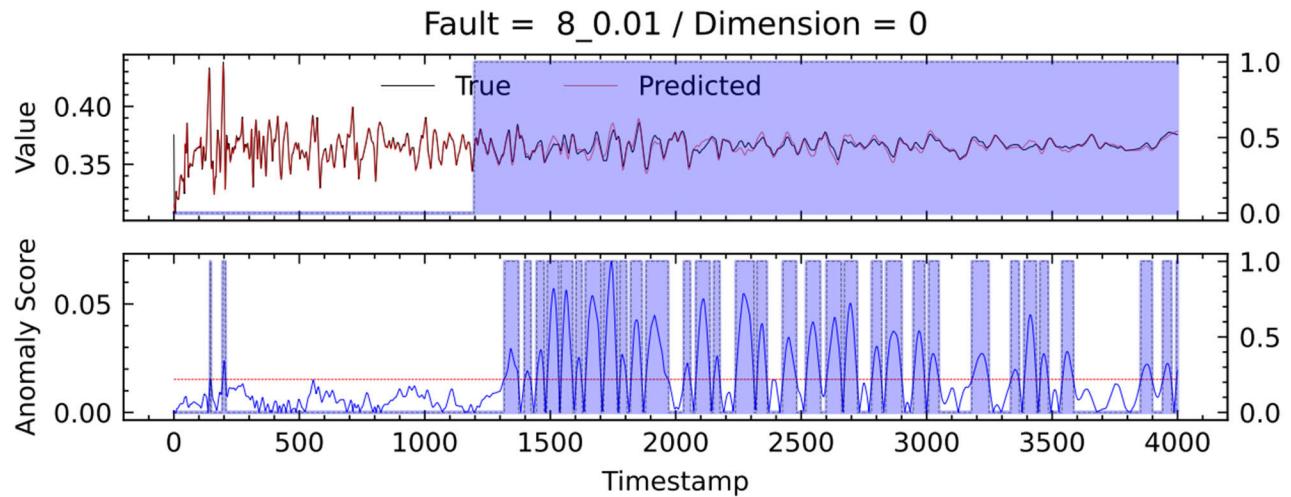
**FIGURE 16.** Anomaly score graph for training with faults type #8 and testing with faults type #8 (similar samples for training and testing).

**TABLE 7.** Performance comparison of TransSiamese with SOTA methods on the paper dataset [14]. The best scores are highlighted in bold.

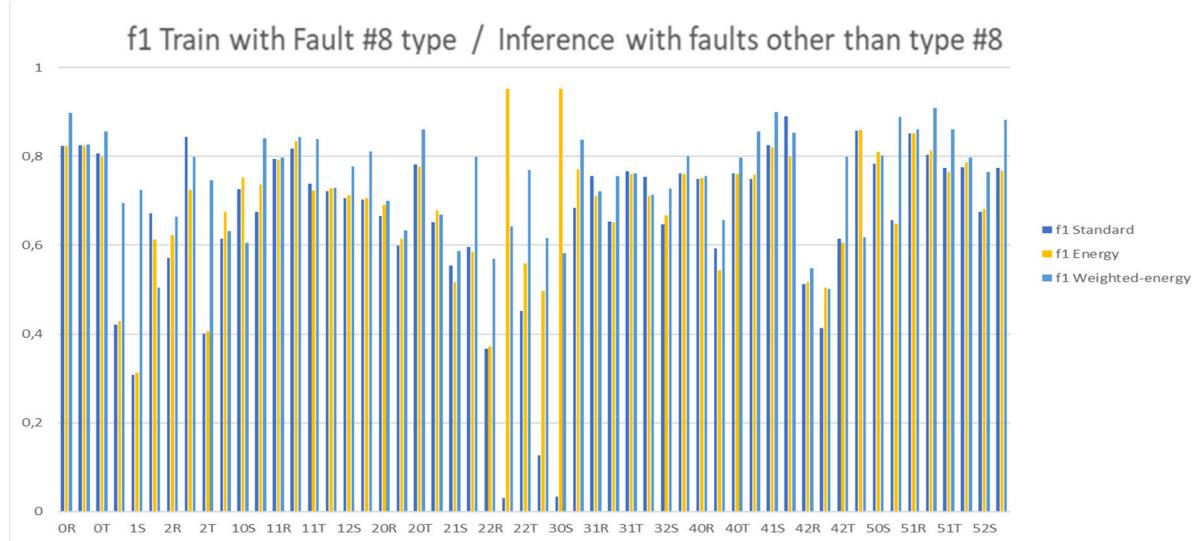
<i>Algorithm</i>	<i>Performance comparison</i>			
	<b>f1</b>	<b>precision</b>	<b>recall</b>	<b>ROC/AUC</b>
DAGMM	0.0439	0.2516	0.0289	0.4296
TranAD	0.2676	0.5339	0.2918	0.4991
OmniAnomaly	0.3520	0.7076	0.2836	0.4232
GDN	0.3893	0.6508	0.3366	0.4948
LSTM_AD	0.5656	0.7957	0.5216	0.5414
<b>TransSiamese</b>	<b>0.7858</b>	<b>0.9521</b>	<b>0.6738</b>	<b>0.7968</b>

In contrast to typical benchmarking datasets for anomaly detection, where signals have normally punctual anomalies throughout the entire signal, our signals exhibit this continuous anomaly characteristic. This peculiarity is due to the fact

that signal rebounds are almost identical until the point of the fault occurrence. From that point onward until the end of the signal, discrepancies are present throughout the signal's duration.



**FIGURE 17.** Inference visualization of anomaly score with weighted-energy from a fault type #8.



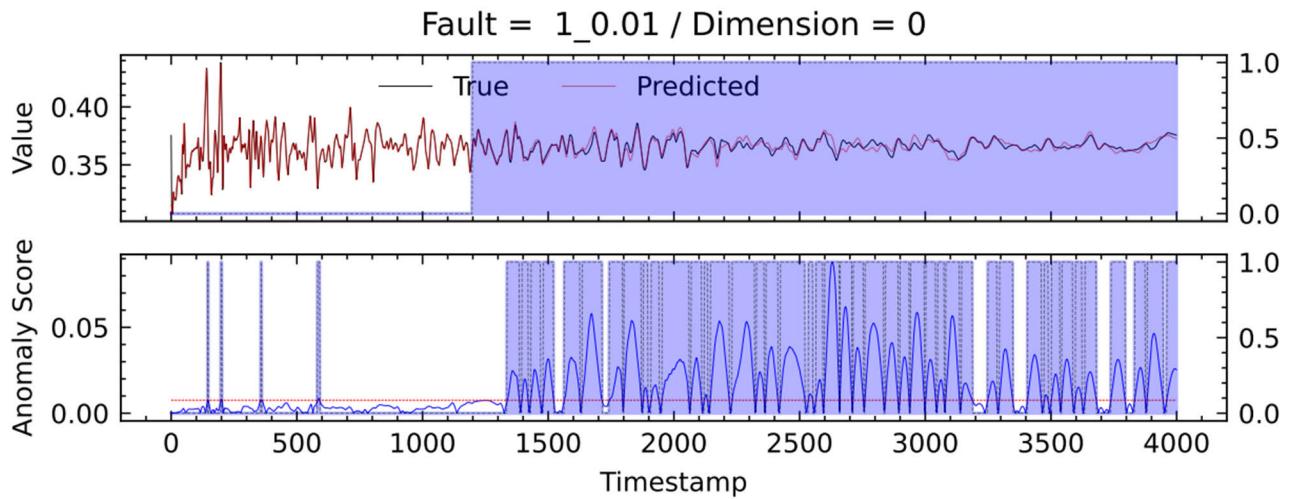
**FIGURE 18.** Anomaly score graph for training with faults type #8 and testing with faults different than type #8.

**TABLE 8.** Comparison of training times.

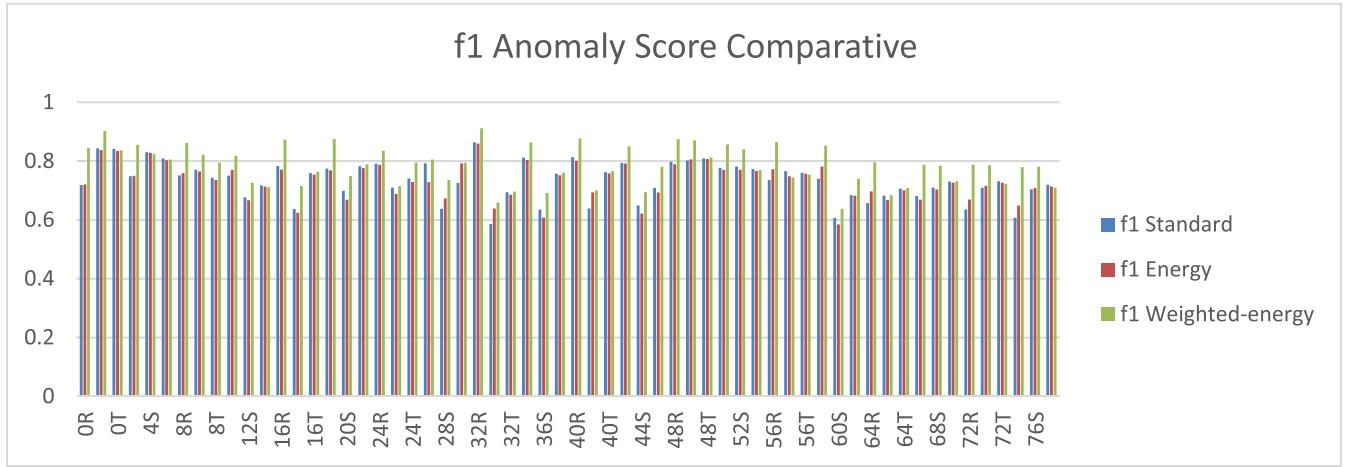
Algorithm	Training phase			
	Elapsed Time [seg]	Epochs	Time/epoch [seg/epoch]	% time comparative
GDN	46062.031	100	460.620	8833%
LSTM_AD	1415.840	100	14.158	272%
OmniAnomaly	1115.000	100	11.150	214%
DAGMM	798.188	100	7.982	153%
TranAD	302.844	100	3.028	58%
TransSiamese	521.465	100	5.215	100%

Training different SOTA models, as shown in Table 7, does not yield satisfactory results. While some models achieve f1 scores close to 0.5, this score does not truly reflect the essence

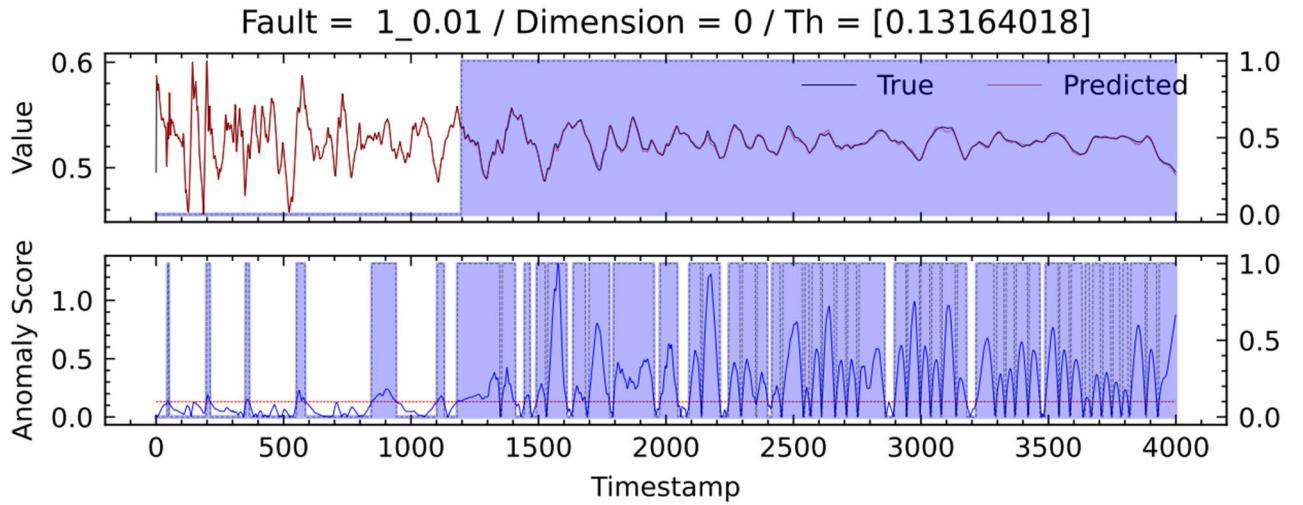
of the f1 measure (a metric that computes how many times a model makes correct predictions across the entire dataset). The reason is the nature of the mentioned anomalies in our



**FIGURE 19.** Inference visualization of anomaly score with weighted-energy from a fault type #1.



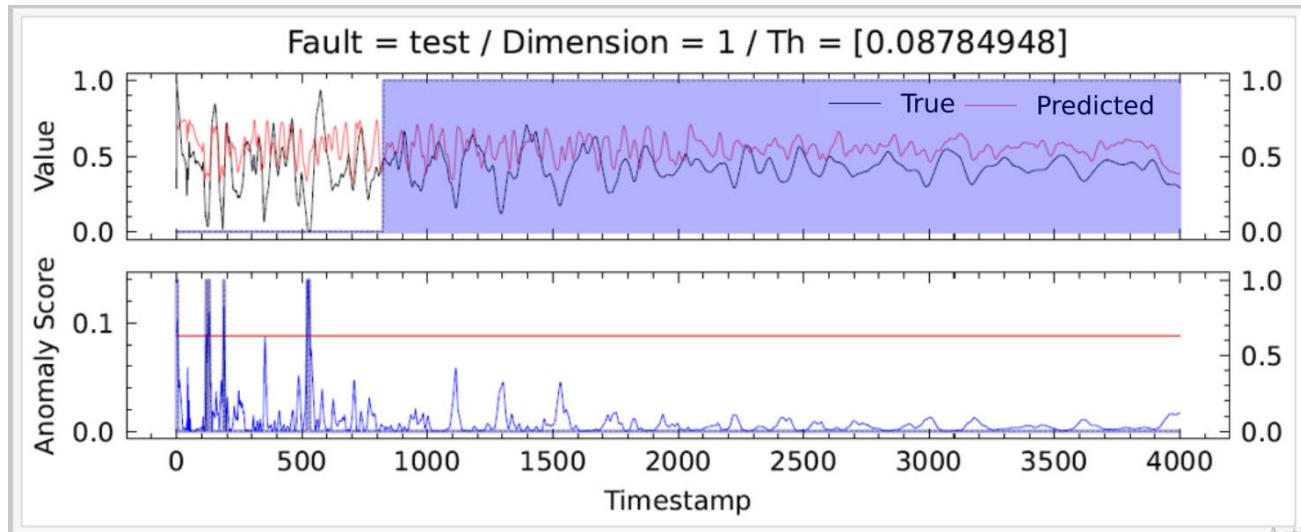
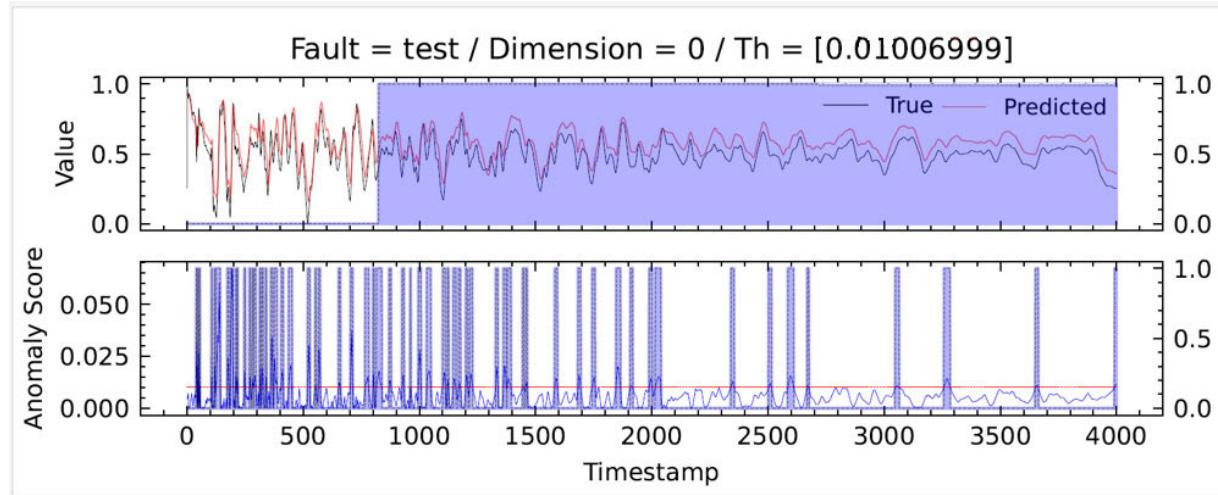
**FIGURE 20.** Anomaly score graph for training with all types of faults and testing with all types of faults (different samples for training and testing).



**FIGURE 21.** Inference visualization of anomaly score with weighted-energy from a fault type #1.

dataset. Since the signals have an initial zone with virtually no anomalies and from a certain time onwards, the anomaly persists until the end, a poorly-performing model will still

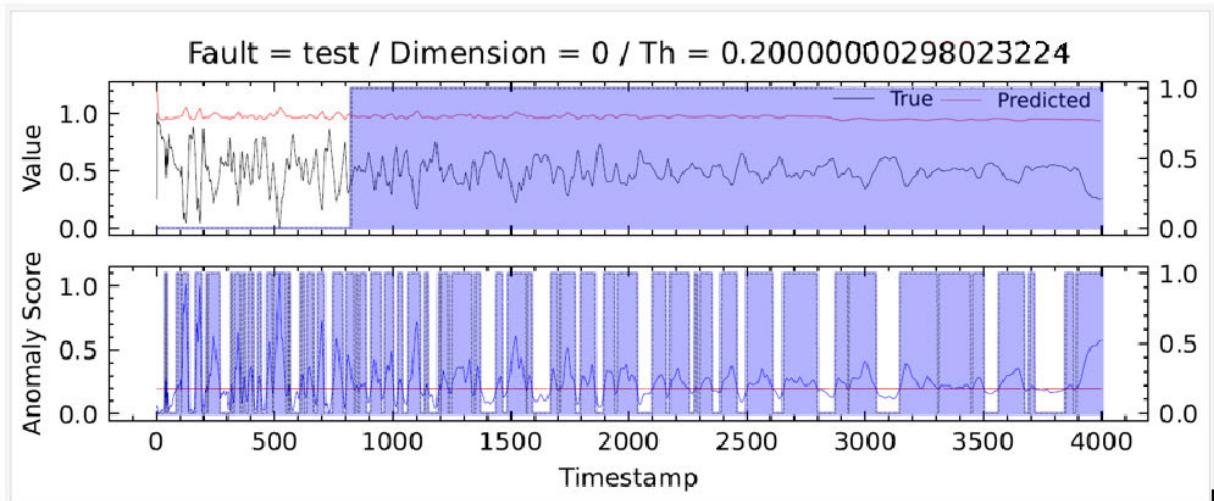
have reasonably high correct predictions by chance alone, as they will coincidentally align with the model's erroneous predictions.

**FIGURE 22.** Inference visualization of anomaly score from LSTM-AD model.**FIGURE 23.** Inference visualization of anomaly score from GDN model.**TABLE 9.** Comparison of inference times.

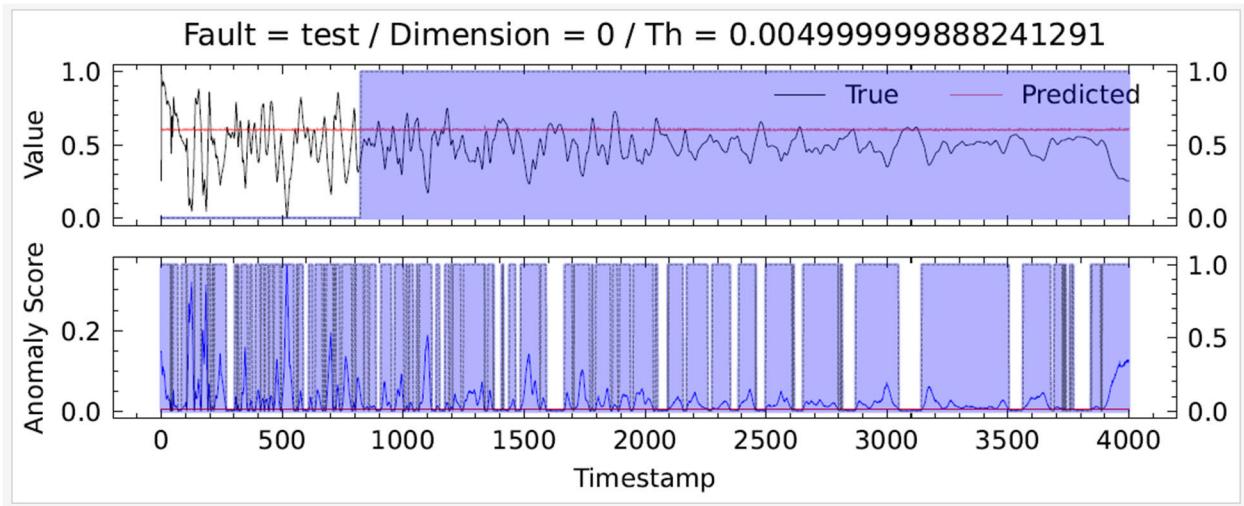
Algorithm	Inference phase			
	Threshold computing [seg]	Inference process [seg]	Total time consumed [seg]	% time comparative
GDN	0.07813	20.3005	20.379	10333%
OmniAnomaly	0.07813	2.4189	2.497	1266%
DAGMM	0.07813	1.7799	1.858	942%
LSTM_AD	0.07813	1.5991	1.677	850%
TranAD	0.07813	0.0791	0.157	80%
TransSiamese	0.07813	0.1191	0.197	100%

Figures (Fig 22, Fig 23, Fig 24, Fig 25, Fig 26) illustrate the response of each of the tested models, demonstrating that anomaly detection is weak in all of them due to the nature of our signals continuous anomalies. As previously mentioned, they all have a random “hit rate” within the anomaly or non-anomaly window.

This fact becomes evident when examining the ROC/AUC parameter, which reflects the relationship between the true positive rate (as a percentage of all occurring events) and the false positive rate (as a percentage of all non-occurring events). It can take values between 0 and 1, with a perfect classifier having an ROC/AUC of 1 and a random classifier



**FIGURE 24.** Inference visualization of anomaly score from LSTM-AD model.



**FIGURE 25.** Inference visualization of anomaly score from Omnipanomaly model.

having an ROC/AUC of 0.5. In our case, the models tested with our dataset have an ROC/AUC value around 0.5, revealing their poor performance (close to randomness).

The model presented in this paper combines the Siamese scheme with Contrastive learning and a score based on weighted energy computation during inference. These features result in an f1 score around 0.78 and ROC/AUC values of 0.85 when tested on our electrical signal fault database, indicating the superior performance of this model compared to other SOTA models for our specific database case.

In conclusion, we can say that our model doesn't necessarily claim superiority over other state-of-the-art (SOTA) models. Instead, we could say that it presents a new approach to anomaly detection, leveraging Transformers as a foundation.

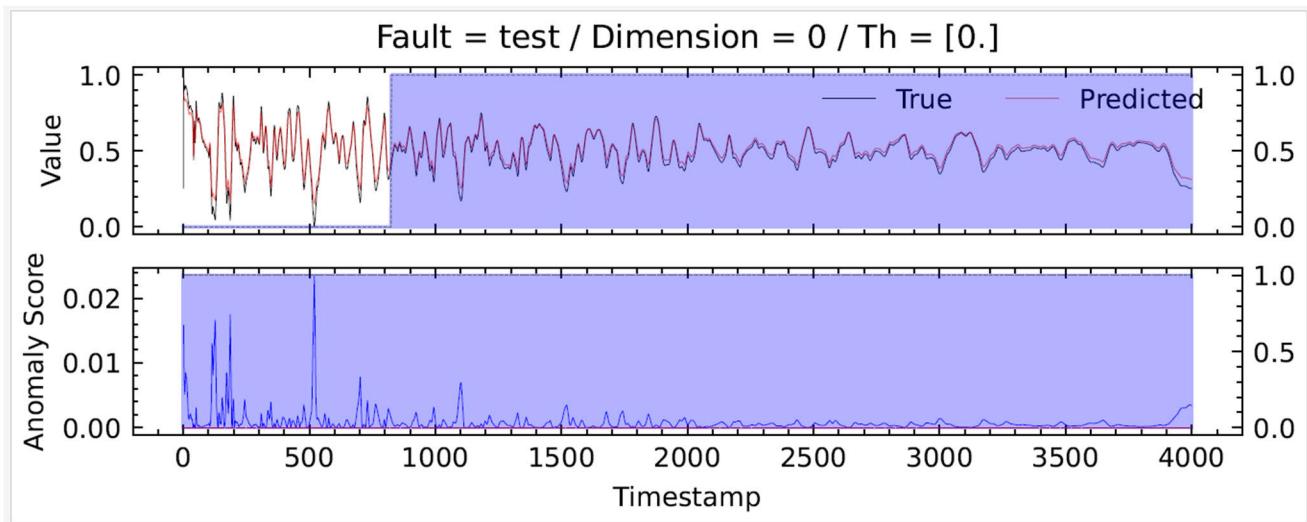
Most SOTA models usually focus on detecting anomalies that occur sporadically within a short time frame, but our

model is designed to identify anomalies that persist from a specific moment until the end of the data sequence.

Due to this reason the SOTA models are not well-suited to handle such data patterns. This approach allowed us to treat the entire signal as a whole and helped the network learn to distinguish specific temporal sections of the signal that highlight differences in the learned signal model.

Consequently, our model successfully learned to discern between the pre-fault section and the rest of the signal following the anomaly. So, it's important to note that our model's strength lies in its adaptability to a particular type of data where SOTA models might not perform well due to the inherent nature of the data.

In terms of computational cost comparison tests, the presented model "TransSiamese" (Table 8 and Table 9) ranks below only "TranAD" during training, while having lower computational cost than the rest of the models. In inference



**FIGURE 26.** Inference visualization of anomaly score from TranAD model.

mode, it is also surpassed only by “TranAD,” while having lower computational cost than the other models.

## V. CONCLUSION

The present study becomes a new approach to the problem of anomaly detection in temporal signals using Transformers as the main component of the proposed system. The difference with respect to other authors who have used similar techniques lies in the very nature of the data that needs to be processed.

These diverge from the data normally used in anomaly detection because of their marked character of long-term changes over an oscillating signal typical of electrical systems. This means that the algorithms present so far do not provide the results offered by our proposal. Mainly because the algorithms prior to ours focus on detecting anomalies that may occur in a short period of time and sporadically.

The use of the Siamese network scheme and the corresponding training algorithm, Contrastive learning, has helped to treat the signal as a whole. Causing the network to learn to discern specific temporal sections of the signal that would enhance the differences on the learned signal model. In this way, the network has successfully learned to discern between the pre-fault section, and the rest of the signal after the fault. In short, the Siamese-Contrastive pair clearly dampens the differences in the pre-fault section and amplifies these differences during the fault period. Therefore, we dare to conclude that this is one of the differentiating features with previous anomaly detection algorithms.

On the other hand, and unlike previous proposals, during training we have used the information from the fault signals, and not only learnt the representation of the pre-fault signals (in normal conditions) to highlight the differences when a fault unfortunately appears. Also, the database used is only 200 fault signals with a single pre-fault signal. Demonstrating

that the use of approximations such as the Transformer is able to obtain results using schemes such as the one presented based on the Siamese network. Finally, we have used 3 different scores to obtain the best possible thresholding. These three scores, and especially the Weighted-energy score, have given stability to the results of the Siamese network.

## VI. FUTURE WORK

Future work should focus primarily on improving the results shown in this paper. Although the results presented here are good and represent an f1 value close to 80%, the expectations of the working group are to improve them in order to achieve a reliable system that will allow us to move towards the goal of fault localization.

It is this challenge, localization, which is therefore our focus of interest. And it is around this epicenter that our research will pivot from now on. In this respect, we are interested in applying other models, such as diffusion models, to the creation of a more extensive signal database.

On the other hand, the objectives of the research team in the field of electrical protection include the detection of faults in power transformers, the adjustment of protection functions based on network parameters, the development of protection algorithms for the detection of power oscillations and loss of stability, the detection of faults in DC networks, and the automatic generation of diagnostic reports.

Finally, our great challenge is to take this study to a real installation, which we are already working on and are currently carrying out field tests to obtain real faults. Here we will have the opportunity to validate our studies with simulated data and make any corrections that may be considered.

## GRAPHICS OF EXPERIMENT B: ANOMALY SCORE FUNCTIONS COMPARATIVE

See Table 1, Figures 10 and 11.

**GRAPHICS OF EXPERIMENT C-1: TRAINING WITH FAULTS TYPE #1/TEST WITH FAULTS TYPE #1**

See Table 2, Figures 12 and 13.

**GRAPHICS OF EXPERIMENT C-2: TRAINING WITH FAULTS TYPE #1/TEST WITH FAULTS DIFFERENT OF TYPE #1**

See Table 3, Figures 14 and 15.

**GRAPHICS OF EXPERIMENT D-1: TRAINING WITH FAULTS TYPE #8/TEST WITH FAULTS TYPE #8**

See Table 4, Figures 16 and 17.

**GRAPHICS OF EXPERIMENT D-2: TRAINING WITH FAULTS TYPE #8/TEST WITH FAULTS OTHER THAN TYPE #8**

See Table 5, Figures 18 and 19.

**F. TRAINING WITH ALL TYPE OF FAULTS/TEST WITH ALL TYPES**

See Table 6, Figures 20 and 21.

**G. PERFORMANCE COMPARISON OF SEVERAL SOTA MODELS**

See Tables 7–9, Figures 22–26.

**REFERENCES**

- [1] Q. Wang, “Research on fault prediction and self-healing control technology of distribution network,” in *Proc. Int. Conf. Elect., Control Inf. Technol.*, Mar. 2022, pp. 1–5.
- [2] L. Peretto, R. Tinarelli, A. Bauer, and S. Pugliese, “Fault location in underground power networks: A case study,” in *Proc. ISGT*, Jan. 2011, pp. 1–6, doi: [10.1109/ISGT.2011.5759198](https://doi.org/10.1109/ISGT.2011.5759198).
- [3] A. Bahmanyar, S. Jamali, A. Estebsari, and E. Bompard, “A comparison framework for distribution system outage and fault location methods,” *Electr. Power Syst. Res.*, vol. 145, pp. 19–34, Apr. 2017, doi: [10.1016/j.epsr.2016.12.018](https://doi.org/10.1016/j.epsr.2016.12.018).
- [4] J.-H. Teng, W.-H. Huang, and S.-W. Luan, “Automatic and fast faulted line-section location method for distribution systems based on fault indicators,” *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1653–1662, Jul. 2014. [Online]. Available: <https://ieeexplore.ieee.org/cuarczo.unizar.es:9443/document/6695790/>
- [5] J. Oasa, M. Yamanaka, S. Higashiyama, Y. Inaoka, T. Hisakado, O. Wada, T. Matsushima, T. Hirayama, and K. Yamaoka, “Verification of fault location by TDR measurement on an actual line including multiple ground-mounted equipment,” in *Proc. CIRED 26th Int. Conf. Exhib. Electr. Distrib.*, 2021, pp. 1274–1278.
- [6] R. Lacoste, “Time domain reflectometry,” in *Robert Lacoste’s The Darker Side*, 2010, pp. 33–47, doi: [10.1016/B978-1-85617-762-7.00003-4](https://doi.org/10.1016/B978-1-85617-762-7.00003-4).
- [7] J. G. Fornás, E. H. Jaraba, H. Bludszuweit, D. C. García, and A. L. Estopíñan, “Modeling and simulation of time domain reflectometry signals on a real network for use in fault classification and location,” *IEEE Access*, vol. 11, pp. 23596–23619, 2023, doi: [10.1109/ACCESS.2023.3253772](https://doi.org/10.1109/ACCESS.2023.3253772).
- [8] Y. Miyoshi and S. Saba, “Some features and performances of type c transmission-line fault locators,” *Trans. Amer. Inst. Electr. Eng. III, Power App. Syst.*, vol. 76, no. 3, pp. 445–451, Apr. 1957, doi: [10.1109/AIEEPAS.1957.4499585](https://doi.org/10.1109/AIEEPAS.1957.4499585).
- [9] C.-C. Zhou, Q. Shu, and X.-Y. Han, “A single-phase Earth fault location scheme for distribution feeder on the basis of the difference of zero mode traveling waves,” *Int. Trans. Electr. Energy Syst.*, vol. 27, no. 5, May 2017, Art. no. e2298, doi: [10.1002/ETEP.2298](https://doi.org/10.1002/ETEP.2298).
- [10] W. Chonglin, W. Yangyang, L. Rui, and S. Gang, “Fault location for single-phase-to-earth faults based on transient traveling wave method and artificial pulse signal injection method,” in *Proc. Int. Conf. Electr. Control Eng.*, Jun. 2010, pp. 3737–3741. [Online]. Available: <https://ieeexplore.ieee.org/cuarczo.unizar.es:9443/document/5629889/>
- [11] M. Abad, U. Zaragoza, N. El, H. Saudi, A.-S. Arabia, and M. G.-G. Circe-Spain, “New fault location method for up-to-date and upcoming distribution networks,” Tech. Rep., 2015, pp. 15–18.
- [12] F. J. L. Padua and R. de Oliveira, “Allocation of PLC devices in a low-voltage grid,” in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Latin Amer. (ISGT Latin Amer.)*, Sep. 2019, pp. 1–4, doi: [10.1109/ISGT-LA.2019.8895457](https://doi.org/10.1109/ISGT-LA.2019.8895457).
- [13] J. G. Fornás, E. H. Jaraba, A. L. Estopíñan, and J. Saldana, “Detection and classification of fault types in distribution lines by applying contrastive learning to GAN encoded time-series of pulse reflectometry signals,” *IEEE Access*, vol. 10, pp. 110521–110536, 2022, doi: [10.1109/ACCESS.2022.3214994](https://doi.org/10.1109/ACCESS.2022.3214994).
- [14] E. Herrero, J. Granado, and A. Llombart, “CVLAB-UNIZAR/CIRCE\_fault\_database: Initial release,” Tech. Rep., Nov. 2022, doi: [10.5281/ZENODO.7316010](https://doi.org/10.5281/ZENODO.7316010).
- [15] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, G. Rasool, and R. P. Ramachandran, “Transformers in time-series analysis: A tutorial,” 2022, *arXiv:2205.01138*.
- [16] M. Hu, X. Feng, Z. Ji, K. Yan, and S. Zhou, “A novel computational approach for discord search with local recurrence rates in multivariate time series,” *Inf. Sci.*, vol. 477, pp. 220–233, Mar. 2019, doi: [10.1016/j.ins.2018.10.047](https://doi.org/10.1016/j.ins.2018.10.047).
- [17] Z. Ji, Y. Wang, K. Yan, X. Xie, Y. Xiang, and J. Huang, “A space-embedding strategy for anomaly detection in multivariate time series,” *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117892, doi: [10.1016/J.ESWA.2022.117892](https://doi.org/10.1016/J.ESWA.2022.117892).
- [18] H. Qin, X. Zhan, and Y. Zheng, “CSCAD: Correlation structure-based collective anomaly detection in complex system,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4634–4645, May 2023, doi: [10.1109/TKDE.2022.3154166](https://doi.org/10.1109/TKDE.2022.3154166).
- [19] G. Xiang and R. Lin, “Robust anomaly detection for multivariate data of spacecraft through recurrent neural networks and extreme value theory,” *IEEE Access*, vol. 9, pp. 167447–167457, 2021, doi: [10.1109/ACCESS.2021.3136505](https://doi.org/10.1109/ACCESS.2021.3136505).
- [20] Z. Ji, J. Gong, and J. Feng, “A novel deep learning approach for anomaly detection of time series data,” *Sci. Program.*, vol. 2021, pp. 1–11, Jul. 2021, doi: [10.1155/2021/6636270](https://doi.org/10.1155/2021/6636270).
- [21] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, X. Zhao, and L. Dong, “Detecting anomalies in time series data via a meta-feature based approach,” *IEEE Access*, vol. 6, pp. 27760–27776, 2018, doi: [10.1109/ACCESS.2018.2840086](https://doi.org/10.1109/ACCESS.2018.2840086).
- [22] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, “USAD: UnSupervised anomaly detection on multivariate time series,” Tech. Rep., 2020, doi: [10.1145/3394486.3403392](https://doi.org/10.1145/3394486.3403392).
- [23] B. Zong, “Deep autoencoding Gaussian mixture model for unsupervised anomaly detection,” in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, vol. 5, Feb. 2018, pp. 3800–3809.
- [24] 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 *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2828–2837, doi: [10.1145/3292500.3330672](https://doi.org/10.1145/3292500.3330672).
- [25] A. Deng and B. Hooi, “Graph neural network-based anomaly detection in multivariate time series,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 5, May 2021, pp. 4027–4035, doi: [10.1609/AAAI.V35I5.16523](https://doi.org/10.1609/AAAI.V35I5.16523).
- [26] H. Zhou, K. Yu, X. Zhang, G. Wu, and A. Yazidi, “Contrastive autoencoder for anomaly detection in multivariate time series,” *Inf. Sci.*, vol. 610, pp. 266–280, Sep. 2022, doi: [10.1016/j.ins.2022.07.179](https://doi.org/10.1016/j.ins.2022.07.179).
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [28] J. Xu, H. Wu, J. Wang, and M. Long, “Anomaly transformer: Time series anomaly detection with association discrepancy,” 2021, *arXiv:2110.02642*.
- [29] J. Abulizi, Z. Chen, P. Liu, H. Sun, C. Ji, and Z. Li, “Research on voiceprint recognition of power transformer anomalies using gated recurrent unit,” in *Proc. Power Syst. Green Energy Conf. (PSGEC)*, Aug. 2021, pp. 743–747, doi: [10.1109/PSGEC51302.2021.9542338](https://doi.org/10.1109/PSGEC51302.2021.9542338).
- [30] H. Wang, X. Shen, M. Tu, Y. Zhuang, and Z. Liu, “Improved transformer with multi-head dense collaboration,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2754–2767, 2022, doi: [10.1109/TASLP.2022.3199648](https://doi.org/10.1109/TASLP.2022.3199648).

- [31] M. Yu, D. Wu, W. Rao, L. Cheng, R. Li, and Y. Li, "Automated road crack detection method based on visual transformer with multi-head cross-attention," in *Proc. IEEE Int. Conf. Sens., Diag., Prognostics, Control (SDPC)*, Aug. 2022, pp. 328–332, doi: [10.1109/SDPC55702.2022.9915808](https://doi.org/10.1109/SDPC55702.2022.9915808).
- [32] M. Garg, D. Ghosh, and P. M. Pradhan, "Multiscaled multi-head attention-based video transformer network for hand gesture recognition," *IEEE Signal Process. Lett.*, vol. 30, pp. 80–84, 2023, doi: [10.1109/LSP.2023.3241857](https://doi.org/10.1109/LSP.2023.3241857).
- [33] J. Zhang, Y. Chen, and J. Chen, "Join-chain network: A logical reasoning view of the multi-head attention in transformer," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2022, pp. 1–11, doi: [10.1109/ICDMW58026.2022.00123](https://doi.org/10.1109/ICDMW58026.2022.00123).
- [34] Q. Ma, M. Zhang, Y. Xu, J. Song, and T. Zhang, "Remaining useful life estimation for turbofan engine with transformer-based deep architecture," in *Proc. 26th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2021, pp. 1–6, doi: [10.23919/ICAC50006.2021.9594150](https://doi.org/10.23919/ICAC50006.2021.9594150).
- [35] R. Fan, "Transformer-based deep learning method for the prediction of ventilator pressure," in *Proc. IEEE 2nd Int. Conf. Inf. Commun. Softw. Eng. (ICICSE)*, Mar. 2022, pp. 25–28, doi: [10.1109/ICI-CSE55337.2022.9828926](https://doi.org/10.1109/ICI-CSE55337.2022.9828926).
- [36] B. Padi, A. Mohan, and S. Ganapathy, "End-to-end language recognition using attention based hierarchical gated recurrent unit models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5966–5970, doi: [10.1109/ICASSP.2019.8683895](https://doi.org/10.1109/ICASSP.2019.8683895).
- [37] P. Awwal and S. Naval, "Optimized attention-based Long-short-term memory and gated recurrent unit for malware detection in windows," in *Proc. Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl. (CENTCON)*, Dec. 2022, pp. 217–222, doi: [10.1109/CENTCON56610.2022.10051287](https://doi.org/10.1109/CENTCON56610.2022.10051287).
- [38] S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6294–6298, doi: [10.1109/ICASSP39728.2021.9414489](https://doi.org/10.1109/ICASSP39728.2021.9414489).
- [39] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2114–2124, doi: [10.1145/3447548.3467401](https://doi.org/10.1145/3447548.3467401).
- [40] R. Bai, M. Li, B. Meng, F. Li, M. Jiang, J. Ren, and D. Sun, "Hierarchical graph convolutional skeleton transformer for action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6, doi: [10.1109/ICME52920.2022.9859781](https://doi.org/10.1109/ICME52920.2022.9859781).
- [41] T. L. Hoang, T. D. Pham, and V. C. Ta, "Improving graph convolutional networks with transformer layer in social-based items recommendation," in *Proc. 13th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2021, pp. 1–6, doi: [10.1109/KSE53942.2021.9648823](https://doi.org/10.1109/KSE53942.2021.9648823).
- [42] R. Slama, W. Rabah, and H. Wannous, "STr-GCN: Dual spatial graph convolutional network and transformer graph encoder for 3D hand gesture recognition," in *Proc. IEEE 17th Int. Conf. Autom. Face Gesture Recognit. (FG)*, Jan. 2023, pp. 1–6, doi: [10.1109/FG57933.2023.10042643](https://doi.org/10.1109/FG57933.2023.10042643).
- [43] S. Tuli, G. Casale, and N. R. Jennings, "TranAD," *Proc. VLDB Endowment*, vol. 15, no. 6, pp. 1201–1214, Feb. 2022, doi: [10.14778/3514061.3514067](https://doi.org/10.14778/3514061.3514067).
- [44] H. Takimoto, J. Seki, S. F. Situju, and A. Kanagawa, "Applied artificial intelligence anomaly detection using Siamese network with attention mechanism for few-shot learning," *Tech. Rep.*, 2022, doi: [10.1080/08839514.2022.2094885](https://doi.org/10.1080/08839514.2022.2094885).
- [45] *Siamese Attention Networks*.
- [46] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210, doi: [10.1109/IGARSS46834.2022.9883686](https://doi.org/10.1109/IGARSS46834.2022.9883686).
- [47] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "Learning a distance function with a Siamese network to localize anomalies in videos," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 2598–2607.
- [48] C. Ye and Q. Ma, "TS2 V: A transformer-based Siamese network for representation learning of univariate time-series data," *Tech. Rep.*, 2022, doi: [10.1109/CSCWD54268.2022.9776300](https://doi.org/10.1109/CSCWD54268.2022.9776300).
- [49] *A Transformer-Based Siamese Network for Change Detection | IEEE Conference Publication | IEEE Xplore*. Accessed: May 14, 2023. [Online]. Available: <https://ieeexplore.ieee.org.cuarzo.unizar.es:9443/document/9883686>
- [50] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," 2016, *arXiv:1607.00148*.



**JAVIER GRANADO FORNÁS** received the B.Sc. degree in industrial engineering (specialized in industrial electronics) and the M.Sc. degree in electronics engineering (intelligent environments specialization) from the University of Zaragoza, Spain, in 1994 and 2014, respectively. He is currently pursuing the Ph.D. degree in deep learning around the classification and localization of faults in distribution lines. Since 2009, he has been a Senior Researcher with the Electronics Systems Group, CIRCE Technology Center. His main research interests include electronic design control projects around deep learning and algorithms for fault detection.



**ELÍAS HERRERO JARABA** (Member, IEEE) received the Ph.D. degree in engineering from the University of Zaragoza. He has been a Professor with the University of Zaragoza, since 2002. In 1999, and for four years, he worked in the automotive field with the Production Department, Opel Spain. Within the university, he has developed his research in the area of computer vision, and since 2012, he has focused his interest in the field of neural networks, and more recently, in deep learning.

In the meantime, he held the position of Coordinator of the Smart Vehicle Initiative with the Aragon Institute for Engineering Research (I3A). He currently holds a European patent and has led to more than six projects funded in public competitions, more than 20 projects with companies, 12 indexed publications, and more than 50 contributions to international conferences. It is worth mentioning the 3M award for his research in the CvLAB Research Group. His teaching work has focused on electronics, from its basics to power electronics, including embedded systems.



**ANDRÉS LLOMBART ESTOPIÑÁN** (Member, IEEE) received the Industrial Engineering degree and the Ph.D. degree in electrical engineering from the University of Zaragoza, in 1994 and 2000, respectively. He has been the General Director of CIRCE, since April 2016 and the former Executive Director, since January 2011. He was a Lecturer with the Electrical Department, University of Zaragoza, from June 2003 to May 2018. In November 2011, he was designated by the Science and Innovation Ministry as the expert in the Energy Area Committee of the 7th Funding Program of the European Union, being in charge of the coordination of electricity grid topics. In CIRCE, he is in charge of the Innovation and Promotion Unit, created by himself in May 2009. From March 2007 to March 2009, he was the Subdirector of the Institutional Relations of the Superior Polytechnic Centre with the University of Zaragoza. Later, he researched the impact reduction of power electronic source grids using passive filters in the kilowatt range. From December 1994 to May 2001, he was an Associate Professor with the Electrical Engineering Department, University of Zaragoza, where he performed the following teaching activities: electric circuit theory, industrial actuation, wind energy, and renewable energy integration. He participated in more than 35 research and development +i projects; in 16 of them, he was the primary researcher. Authors of ten articles in indexed journals and more than 50 contributions to international congresses. Active participation in forums, associations, and platforms linked to activity lines. Eight patents, of which seven are being exploited.