Statistics and Deep Learning-based Hybrid Model for Interpretable Anomaly Detection

Thabang Mathonsi
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
thabang.mathonsi@wits.ac.za

Terence L van Zyl
Institute for Intelligent Systems
University of Johannesburg, South Africa
tvanzyl@uj.ac.za

Abstract—**Hybrid methods have been shown to outperform pure statistical and pure deep learning methods at both forecasting tasks, and at quantifying the uncertainty associated with those forecasts (prediction intervals). One example is Multivariate Exponential Smoothing Long Short-Term Memory (MES-LSTM), a hybrid between a multivariate statistical forecasting model and a Recurrent Neural Network variant, Long Short-Term Memory. It has also been shown that a model that (i) produces accurate forecasts and (ii) is able to quantify the associated predictive uncertainty satisfactorily, can be successfully adapted to a model suitable for anomaly detection tasks. With the increasing ubiquity of multivariate data, and new application domains, there have been numerous anomaly detection methods proposed in recent years. The proposed methods have largely focused on deep learning techniques, which are prone to suffer from challenges such as (i) large sets of parameters that may be computationally intensive to tune, (ii) returning too many false positives rendering the techniques impractical for use, (iii) requiring labeled datasets for training which are often not prevalent in real life, and (iv) understanding of the root causes of anomaly occurrences inhibited by the predominantly black-box nature of deep learning methods. In this article, an extension of MES-LSTM is presented, an interpretable anomaly detection model that overcomes these challenges. With a focus on renewable energy generation as an application domain, the proposed approach is benchmarked against the state-of-the-art. The findings are that MES-LSTM anomaly detector is at least competitive to the benchmarks at anomaly detection tasks, and less prone to learning from spurious effects than the benchmarks, thus making it more reliable at root cause discovery and explanation.**

Index Terms—**anomaly detection, interpretability**

I. INTRODUCTION

Time series anomaly detection is useful in applications from a variety of industries [1]. In their analytical study comparing classical and deep learning-based anomaly detection models, Munir et al. [45] observe that deep learning outperforms the classical approaches. In another study, Mathonsi and van Zyl [41] conclude that deep learning is at least competitive to statistical-based models. However, there is still a relative gap that exists for hybrid approaches in anomaly detection within the multivariate setting. Furthermore, root cause discovery and explainability remains an open research problem in the context of multivariate anomaly detection [51, 27].

Mathonsi and van Zyl [42] present a statistics and deep learning-based hybrid model, Multivariate Exponential Smoothing Long Short-Term Memory (MES-LSTM). The method leverages multivariate exponential smoothing within the pre- and postprocessing modules. It also circumvents some problems associated with large sets of parameters requiring tuning, computational cost at training, and extensive inference time. These problems are circumvented by incorporating a small parsimonious recurrent deep neural network for learning, for instance, the cross-dependency structure within the covariates.

In this paper, MES-LSTM is extended to the task of anomaly detection. We also investigate the potential of explainability and interpretability for such a model. These goals are executed with an application to the renewable energy domain.

A. Literarure Review

The literature review is segmented into (i) work that has been presented in the research community relating to explainable anomaly detection, and (ii) explainer systems or techniques for interpreting anomaly detection models, explanation discovery, and root cause analysis.

Furthermore, a video can also be considered time series when the streaming images are taken as a matrix of pixel values with coordinates and a time dimension. As such, some techniques that were traditionally applied to streaming video have also been adapted and extended to fit time series anomaly detection problems. As such, techniques from computer vision have not been excluded in the review of review recent advances and the state-of-the-art.

1) Explainable Anomaly Detection: Some work has been done in unsupervised machine learning. Nguyen et al. [47] for instance, consider the problem of anomaly detection in networks data, with an application to an Internet Service Provider example. The authors show their approach, using variational autoencoders, is able to effectively detect malicious attacks on a network. They use the gradients from the autoencoders for model interpretability.

Rad et al. [51] consider the problem of explainable anomaly detection framed within a high dimensionality setting. The authors report competitive model performance without significant gains in computational cost.

Carletti et al. [14] offer a technique for interpreting Isolation Forests (IF) [36], a commonly used model for anomaly detec-

tion tasks. They limit their focus to the scenario of Industry 4.0., with a particular focus on root cause analysis. Root cause analysis, as an application domain for explainable anomaly detection, emphasizes the need for robust models that are deployable in realistic industrial settings [28, 39].

Westerski et al. [68] consider explainable anomaly detection within a framework for procurement fraud identification. To this end, they consider real data from a government department in Singapore. The authors report their techniques, in real-life deployment, resulted in cost and time reduction of up to 10% compared to previously applied compliance checks. There lack in ubiquity of such studies illustrates the need for models that do not suffer from high computational cost. Great computational expense is one of the biggest criticisms of deep learning as it hinders real-time deployment in real-world applications [10].

Liznerski et al. [37] use an explainable convolutional model for one class image classification, and detail some challenges from spurious effects such as watermarks. A similar approach is followed by Pang et al. [49]. A distinguishing factor is that the latter considers both training with no anomalies, and trainingwith anomalous observations as well. Considering the problem of streaming images, Wu et al. [69] apply denoising autoencoders to surveillance video. The authors report competitive results with reduced computational time.

In terms of image classification, Ruff et al. [54] offer a comprehensive review of other techniques that have been applied in the field, and systematically compare their performance on benchmark examples. Another notable review [48] looks at deep anomaly detection, and critically analyses the advantages and disadvantages of various techniques. Others, such as the works of Chalapathy and Chawla [15] and Kiran et al. [31], only consider different classes of models applied to specific application domains. The interested reader is also directed to a discussion of explainability in health data [65], financial data [23], and object detection and recognition in image data [57, 44]. For a more general discussion of concepts related to explainability such as interpretability, and understandability, there are targeted resources such as Barredo Arrieta et al. [7].

Applications for time series classification or anomaly detection can be discerned from the plurality of public dataset repositories, such as the UCR [17] or the UEA archive [4], for instance. Such applications, with a focus on explainable anomaly detection, include predictive maintenance at industrial sites [58], or rule extraction for unsupervised anomaly detection conducted [6].

Some models have been successful at classification and anomaly detection tasks, but due to their complex hierarchical architectures, incorporating explainability proves difficult. Some of these techniques are discussed here for completeness. One such advancement is an ensemble method, Collective of Transformation-based Ensembles (COTE) [2], where 35 models are ensembled over different time series representations based on transformations such as time warping [29] or shapelets [9], for instance.

COTE was further extended by Lines et al. [35], using Hierarchical Voting (HIVE-COTE). HIVE-COTE performed well over the UCR and UEA archives [3]. Using a weighted probabilistic ensemble [33], this ensemble approach combines Shapelet Transform Classifier (STC) [26], Contractable Bag of Symbolic-Fourier Approximation Symbols (CBOSS) [43], Time Series Forest (TSF) [20] and Random Interval Spectral Ensemble (RISE) [35].

2) Explainability Systems and Methods: When categorized based on scope of the explanation, explainer systems offer either local or global explanations. Local explanations explain a single prediction result over the entire model, i.e., it explains the conditional interaction between dependent and independent variables with respect to the single prediction. As mentioned by Ribeiro et al. [52], the explanation are required to make sense within a local setting. In the context of the current study, this means that one explanation should be valid in some region encompassing immediate *neighbors*. The immediate neighbors are understood to be anomalous observations occurring around the same time, and of the same type.

Explainability of anomalies can also be conducted for a (potentially large) *set* of anomalies, for example, in the form of rule lists. These are called global explanations. Finding a truly global explanation, one that applies to multiple anomalous observations of different types occurring at different times is a difficult task [27]. As such, global explanation are usually aggregates of different explanations, or the most representative explanations for the entire model.

Another distinction can be drawn between model-specific and model-agnostic explainer systems. The former are applicable to certain kinds of model(s) (say for instance, strictly convolutional or strictly recurrent models, but usually not applicable to both) while the latter can be applied to multiple models.

Yet another distinction can be drawn between feature attribution, path attribution, and association rule mining techniques. Feature attribution determines the contribution of each feature towards the model's prediction for a given input example. This attribution shows the relationship between a feature and the prediction. As a result, users are able to understand which features their network relies on.

Path attribution methods explain the output of the model that is based on gradients. That means the contribution of each feature is computed by aggregating the gradients from baseline values to the current input along the path. One such method is Path Integrated Gradients (PIG) [63].

In contrast, association rule mining finds correlations and co-occurrences between features in a large dataset. They are considered as most interpretable prediction models with their simple if-then rules. A rule is essentially an if-then statement with two components: an antecedent and a consequent. The input feature with a condition is an antecedent and a prediction is its consequent. The popular techniques to extract the rules from a large dataset are Scalable Bayesian Rule Lists (SBRL) [70], Gini Regularization (GiniReg) [12] and Rule Regularization (RuleReg) [11]. Such techniques have been applied successfully to classifiers in surveillance tasks [66].

Barredo Arrieta et al. [7] discuss transparent models that automatically incorporate explainability such as Logistic regression, decision trees, and nearest neighbour models, as well as post-hoc models, that are explainable with the aid of an additional technique.

Explainability techniques that have been developed for or are typically used for image-based models include Deep Learning Important FeaTures (DeepLIFT) [60], Local Explanation Method using Nonlinear Approximation (LEMNA) [24], and Gradient-weighted Class Activation Mapping (Grad-CAM) [57]. These explainer systems usually output heatmaps [32] or saliency visualisations [62] that rank the feature importance of input images input to the network. A good example of saliency maps is presented by Siddiqui et al. [61], for example, with application to convolutional layers.

For time series models, techniques often employed are Model Agnostic Supervised Local Explanations (MAPLE) [50], Local Interpretable Model-agnostic Explanations (LIME) [52], Local rule-based explanations (LORE) [22], Learning to explain (L2X) [16] and Shapley additive explanations (SHAP) [38]. As a cautionary note in particular for time series modeling, there is difficulty due to temporal dependence inherent in the data. As a consequence, surrogate solutions such as LIME or SHAP neglect the chronological sequence ordering of the model inputs.

LIME [52] explains model inferences by using a local interpretable sparse linear model as an approximation. Anchors [53] offers an incremental improvement on LIME by replacing the linear model used as proxy with a logical rule for explaining inferences. Anchors offers better coverage and explainability of anomalous neighbors, but is not readily applicable to time series data. Other local explainer systems that rank feature importance include responsibility scores (RESP) [8] and axiomatic attribution [63].

B. Motivation

It is straightforward to motivate for deep learning as a mechanism for solving time series classification problems and anomaly detection tasks. One reason is to leverage the feature learning abilities of deep learning algorithms [46]. Deep neural networks have also performed well at other tasks requiring temporal sequence modeling (similar to time series) such as natural language processing [5] and speech recognition [56]. Lastly, deep learning has been shown to scale well with increased dimensionality [30].

However, there exists some challenges with the deep learning approach. These include large sets of parameters that may be computationally expensive to tune, and long inference time that may be impractical in settings that require fast reliable information as feedback from models [42].

As evidenced from existing scholarly research, there is a great need for explanation discovery and interpretable anomaly detection with real-world applications such as root cause analysis [36, 28, 39]. There is a need to circumvent the computational cost and time complexity usually associated with deep learning that prevents them from being used outside of a laboratory setting, and enables deployment in real-world applications such as in the compliance study conducted by Westerski et al. [68] or the streaming video study by Wu et al. [69].

In addition, learning from spurious effects can contaminate the root cause and explanation discovery leading to stakeholders making bad decisions informed by incorrectly explained model inferences. It is important to minimize the effects of learning from, say, random noise in time series data or even watermarks in image data [37].

As a final point, this study may be motivated using another factor from real world applicability. If an anomaly is identified, it might be time consuming for a domain expert or human agent to inspect all the components that may have possibly contributed to the anomalous event in order to identify the root cause. It may be more useful to the inspector if, for instance, a model is able to narrow the search space down to a reasonable fraction of components that are most probable to have contributed to the anomaly.

C. Contribution

The novel contribution can be summarized as follows:

- a statistics and deep learning-based forecast machinery is extended to anomaly detection tasks;
- this new hybrid anomaly detection model (i) incorporates a dynamic threshold-setting approach, which learns and adapts the applicable threshold as new information becomes available, and (ii) functions within a semi-supervised framework, so no golden labels are required for training or detecting the thresholds for detecting anomalies; and
- the presented approach is augmented with explainability and interpretability, thus enabling root cause analysis, and how well the model avoids learning from spurious effects using a novel metric.

II. METHODOLOGY

Renewable energy resources, such as wind/solar farms, affect the grid in a different way compared to conventional fossil fuel generators due to their stochastic nature. In particular, the uncertain disturbances introduced by renewables may compromise operational grid safety. This scenario emphasizes the need for system operators to accurately identify disturbances in a timeous fashion. They are then able to perform corrective measures timely so as to ensure the safety of the grid.

System operators have access to streaming time-stamped measurements, from monitors such as phasor measurement units. These measurements enable system operators to answer critical questions including (i) When is an event happening? (ii) What type of event is happening? and (iii) Where is the source that caused the event? These are the research questions stated succinctly, and they would be phrased equivalently for other domains besides renewable energy regeneration.

Following the methodology of Zheng et al. [72], who first proposed the Power Systems Machine Learning dataset

(PSML) [73] for use within the machine learning for decarbonized energy grids domain, the problem can be formulated as follows.

A. Problem Statement

The streaming measurements can be denoted by $X \in \mathbb{R}^{N \times K}$, where N is the number of available observations and K is the number of measurements or covariates.

Event detection aims to answer the first question above, by identifying an oscillation occurrence when or if it happens. Answering this question involves using a model $\mathcal{H}$ to identify the oscillation occurrence given sequence X, i.e., $\mathcal{H} : X \to \{0, 1\}$. Suppose an event occurs at time t_*: an alarm should be raised when $t \overset{\geq}{m} t^*$, and as soon as possible (1 predicted).

Event classification answers the second question above based on streaming sensor measurements. Given the observations X, the model $\mathcal{H}$ must classify the underlying event type ξ, i.e., $\mathcal{H} : X \to \xi$. PSML presents a truly multivariate problem as ξ is more than just binary classification (i.e. normal or anomalous), but it constitutes a subset of disturbances $\mathcal{C}$ where $\mathcal{C} := \{$branch fault, branch tripping, bus fault, bus tripping, generator tripping, forced oscillation$\}$. This problem framing emphasizes the need to keep track of multiple streams of data with interdependent covariates that are autocorrelated interacting within the global grid. By observing variables such as voltage from each bus in the system, the aim is to determine based on thresholds, for example, if and what kind of event is occurring.

Event localization focuses on locating events from disturbances $\mathcal{C}$, or the root cause of events (for forced oscillations) by observing measurements. The model $\mathcal{H}$ must map measurements X to the bus(es) z nearest to the events detected or the root cause of the events, i.e., $\mathcal{H} : X \to z$, where z is a subset of buses $\mathcal{Z}$ in the entire system.

B. Algorithms

The following benchmark models are used: Inception-Time [21], multi-channels deep convolutional neural networks (MC-DCNN) [74], Residual Network (ResNet) [67], Time series attentional prototype network (TapNet) [71], Minimal random convolutional kernel transform (MiniRocket) [19], one-Nearest neighbour with Euclidean distance (NNEuclid) [34], independent dynamic time warping (iDTW) [59], and dependent dynamic time warping (dDTW) [59]. The architectures of the different benchmark models are briefly described next.

1) Classical Models:

a) Nearest Neighbour: The first of classical model is one-Nearest neighbour with Euclidean distance (NNEuclid). Nearest neighbour classifiers with a distance function have been among the most popular techniques for time series classification [34]. In one study, classifiers with dynamic time warping distance perform well as baselines [3]. In another, Lines and Bagnall [34] shows dynamic time warping is at least competitive to all the other distance measures considered. Interestingly, the best performers in the study are reported to be ensembling neural network classifiers combined with different distance measures.

b) Dynamic Time Warping: Dynamic Time Warping (DTW) can be applied to time series data composed of varying length, but for simplicity, the following description is limited to the case involving series of equal length, such as presented by Bagnall et al. [3]. The distance between two equal length series $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)$ is calculated as follows:

1) $\boldsymbol{\psi}$ is a matrix sized $m \times m$ where $\boldsymbol{\psi}_{i,r} = (a_i - b_r)^2$
2) A warping path $\rho = ((e_1, f_1), (e_2, f_2), \ldots, (e_s, f_s))$ is a contiguous set of matrix indices from $\boldsymbol{\psi}$, subject the constraints:
   - $(e_1, f_1) = (1, 1)$
   - $(e_s, f_s) = (m, m)$
   - $0 \leq e_{i+1} - e_i \leq 1 \, \forall \, i < m$
   - $0 \leq f_{i+1} - f_i \leq 1 \, \forall \, i < m$
3) Let $p_i = \boldsymbol{\psi}_{e_i, f_i}$, then the path distance is $\mathcal{D}_p = \sum_{i=1}^m p_i$
4) Multiple warping paths exists, but the aim is to find a path that minimizes the accumulative distance $\rho^* = \min_{p \in \rho} \mathcal{D}_p(\boldsymbol{a}, \boldsymbol{b})$
5) Solving the following relation yields the optimal distance:

$$\text{DTW}_{(i,r)} = \boldsymbol{\psi}_{i,r} + \min \begin{cases} \text{DTW}_{(i-1,r)} \\ \text{DTW}_{(i,r-1)} \\ \text{DTW}_{(i-1,r-1)} \end{cases}, \quad (1)$$

where the final distance is given by $\text{DTW}_{(m,m)}$.

Some improvements may be applied to DTW for increased efficiency, such as constraining deviations from the diagonal but this falls beyond the cope of this paper. Shokoohi-Yekta et al. [59] defines strategies for applying DTW to multivariate setting. These are independent and dependent approaches.

The independent method, iDTW, as the name suggests, has a separate treatment for each dimension. Using a separate distance matrix for each dimension, iDTW then sums the resulting time warping distances:

$$\text{iDTW}_{i,r}(\boldsymbol{x_a}, \boldsymbol{x_b}) = \sum_{k=1}^d \text{DTW}(x_{a,i,k} - x_{b,r,k})^2 \quad (2)$$

The main idea behind Dependent dynamic time warping (dDTW) is the assumption that the accurate warping is identical for all the dimensions. Given a single time series, the matrix $\boldsymbol{\psi}_{i,r}$ is no longer considered the distance between two points, but is redefined as the Euclidean distance between the two vectors that constitute a representation of the full dimensional space. The dependant strategy is more efficient as warping is simultaneous for all the dimensions, and the distance between steps i and r in terms of time resolution is given by:

$$\boldsymbol{\psi}_{i,r}(\boldsymbol{x_u}, \boldsymbol{x_v}) = \sum_{k=1}^d (\boldsymbol{x_{u,k}}, \boldsymbol{x_{v,k}})^2. \quad (3)$$

There also exists an adaptive strategy [59] for selecting between independent and dependent dynamic time warping. How the distance is chosen depends on an instance-by-instance

threshold deducible from the training data. Adaptive time warping falls beyond the scope of the current study.

2) Deep Learning-based Models:

a) MiniRocket: MiniRocket [19] is adapted from Rocket [18] which was ranked among the best performers on multiple datasets in a recent study [55]. The authors report MiniRocket is at most 75 times faster that Rocket, with comparative accuracy. Rocket combines convolution kernels that are randomly initialised, with a linear classifier, typically ridge regression or logistic regression. The method produces feature maps where the maximum value the proportion of positive values (ppv) are extracted.

Hyperparameter tuning is restricted to the following search spaces. The length $\varsigma \in \{7, 9, 11\}$; the kernel weights $w_i \sim \mathcal{N}(0, 1)$; the dilation, d, is sampled from the exponential distribution; and whether or not the series is padded is decided with equal probability.

In contrast, MiniRocket attempts to minimize the randomness characteristic of Rocket. It achieves this by pre-assigning values to a subset of the hyperparameters discussed above, or limiting the search space to a smaller grid, yet still reportedly achieving comparable accuracy. The changes are summarized in Table I [19], where $\mathcal{N}$ is the normal distribution and $\mathcal{U}$ is the uniform distribution.

TABLE I: Summary of changes from Rocket[18] to MiniRocket [19].

	Rocket	MiniRocket
length	$\{7, 9, 11\}$	9
weights	$\mathcal{N}(0, 1)$	$\{-1, 2\}$
bias	$\mathcal{U}(-1, 1)$	from convolution output
dilation	random	fixed
padding	random	fixed
features	ppv, max	ppv
number of features	20,000	10,000

b) MC-DCNN: Multi Channel Deep Convolutional Neural Network (MC-DCNN) [74] is a modification of conventional deep convolutional neural networks. The convolutions are applied independently per covariate in the multivariate input space.

Every dimension of the multivariate input data goes through two convolutional stages with eight filters each of length five and configured with ReLU activation functions. After each convolution there is a max-pooling operation, followed by a fully connected layer. Softmax is used for the final classification.

c) ResNet: ResNet [67] architecture has three convolutional layers within each of three residual blocks, followed by a Global Average Pooling (GAP) layer. The main idea behind ResNet is the use of residual shortcuts connecting consecutive convolutional layers. The key difference when compared with conventional convolutions from fully convolutional networks for instance, is the addition of these linear shortcuts. The shortcuts reduce the vanishing gradient effect [25], by enabling the gradient to flow directly through these connections. In a recent study [55], ResNet ranked among the best performers on multiple datasets.

d) InceptionTime: InceptionTime incorporates ResNet [67] and Inception modules [64]. An Inception module takes as input multivariate series of size $m \times k$. By using a bottleneck layer with length and stride one, it reduces the dimensionality to $m \times k$ where $k' < k$. InceptionTime assigns random initial weights to five instances of the artificial neural network and ensembles them for greater stability [21]. One out of the five networks replaces the three blocks of the aforementioned three classical convolutional layers from ResNet with two blocks of three Inception modules each. However, the new blocks also maintain residual connections, and they too are followed by GAP and softmax layers.

e) TapNet: The final benchmark model considered combines classical and deep learning-based traits. Zhang et al. [71] note that deep learning methods are good at learning low dimensional features and classical approaches such as dynamic time warping are competitive for applications involving small datasets. TapNet, combining these traits, has a network architecture composed of three modules: Random Dimension Permutation, Multivariate Time Series Encoding and Attentional Prototype Learning. These modules can further be broken down into fully connected layers, three sets of convolutional layers that are one dimension each, a global pooling layer, batch normalisation, and Leaky Rectified Linear Units (LReLU) [40].

C. Anomaly Detection

The benchmark models and their hyperparameters are kept constant from the original manuscripts that the techniques were initially introduced. For MES-LSTM, similarly, the model architecture as described by Mathonsi and van Zyl [42] is retained. This hybrid model is used in conjunction with the methodology presented by Mathonsi and van Zyl [41], in particular Algorithm 1 is employed in order to adapt the forecast machinery for the task of anomaly detection.

Algorithm 1 Anomaly Detection

---

if $U_t \leq y_t \leq L_t$ **then**
  y_t is normal
else
  if $\text{IS}_\alpha(y_t) \geq 1.33 \times \text{IS}_\alpha(y_*)$ **and** $y_t > 10 \times \text{std}\{\ldots, y_{t-3}, y_{t-2}, y_{t-1}\}$ **then**
    y_t is anomalous {where y_* is the last anomalous observation}
  end if
end if

Training time series are extracted from the millisecond transient phasor measurement units data. Training samples are randomly selected amounting to 439 time series, and the remaining 110 time series are used for testing (20%). Each sequence has metadata associated with the event type similar to the classification use case, i.e. *branch fault, branch trip, bus*

fault, bus trip, gen trip. Each time series has a sequence length of 960 observations, representing 4s in the system recorded at 240Hz. There are 91 dimensions for each time series, including voltage, current and power measurements across the transmission system. The experiment is repeated 35 times to mitigate the stochastic nature of the deep learning models.

D. Interpretability

The presented approach uses model-agnostic feature attribution techniques. LIME [52] is a local explainability technique suitable for local explanations. This method perturbs the input in the neighbourhood of an instance and examines the output of the model. LIME, thus, indicates the input features the model considers when making a prediction. LIME works by using a proxy based on a simple model, intrinsically interpretable, such as a linear regression model. The surrogate model is applied around each prediction between the input variable space and the corresponding outcome variables space. Explainability is discerned from the main anomaly detection model by perturbing the input variables of a multivariate observation and tracking how the predictions change.

SHAP [38] use Shapley values from cooperative game theory, which indicates what reward players can expect depending on a coalition function. To extend this approach to the explainability of artificial intelligence agents, players are considered as features and reward as the outcome of the model. Although SHAP also provides global interpretability (behaviour of the entire model), this paper focuses on its ability to shed light on local interpretability (behaviour of a single prediction). This local focus enables the use of SHAP in conjunction with LIME and facilitates comparison. The importance rank of feature i is deduced by taking all subsets of features except feature i, $D \setminus x_i$, and computing the effect of the output predictions after adding feature i back to all the subsets previously extracted. All the contributions are then combined to compute the marginal contribution of the each feature.

The labeled dataset used in this study stipulates which predictor contributed most to an anomaly. Ideally, for the interpretability component to be useful for stakeholders, the correct contributor should be identified and ranked high up in terms of feature importance. To ensure this, a novel metric is presented, that can be tuned to the appropriate task-specific sensitivity.

Definition 2.1 (Mean Discovery Score): Let β indicate the task-specific sensitivity, i.e. ideally, the principal contributing predictor κ should be ranked in the highest β features in terms of importance. Let the ith out-of-sample observation that is in fact anomalous be denoted by $y_i \in \{a\}$, where $\{a\}$ is the set of all anomalies. Then, by aggregating how many times this ranking occurs at the specific sensitivity, the mean discovery score (MDS_β) is given by

$$\text{MDS}_\beta = \frac{1}{\mathbf{card}\left(\{a\}_D^{A_i}\right)} \sum_{t=1}^{m} \mathbb{1}_{y_i \in \{a\} \cap \{\mathcal{R}(\kappa_i) \leq \beta\}}, \quad (4)$$

where the rank of an attribute's feature importance is denoted by $\mathcal{R}$, and $\mathbf{card}\left(\{a\}_D^{A_i}\right)$ is how many times an algorithm A_i is able to detect anomalies in dataset D.

MDS_β is useful for scoring the explainability of accurate anomaly detection models, and would not be suitable for use if the set of anomalies correctly detected $\{a\}$ by algorithm A_i is small in comparison to the set of overall anomalous events.

E. Analysis

InceptionTime, MC-DCNN and ResNet are implemented in Tensorflow, TapNet and MiniRocket in Pytorch, and the DTW techniques from scratch. The code implementations in this study retain the default settings detailed by the respective authors in their original manuscripts. Some algorithms have modules embedded that perform hyperparameter tuning. In cases where this is applicable, the hyperparameter optimization modules are kept as is, but no additional tuning is conducted. Below, the configurations for each algorithm is detailed.

For DTW the full warping window is used. MiniRocket is configured with a ridge regression classifier and 10,000 kernels. TapNet uses defaults set to 500 trees, 3,000 epochs, a learning rate of 10^{-4}, weight decay of 10^{-2}, stop threshold of 10^{-8}, number of filters given by 256, 256, and 128 respectively, kernels by 8, 5, and 3 respectively, while dilation is one with no dropout. ResNet has 1,500 epochs, a batch size of 16, learns at a rate of 10^{-2} which, if no improvement is observed for 50 epochs, is set to 5^{-2}. ResNet is configured with three residual blocks composed of three convolutional layers each, where the sizes of the kernels are 8, 5, and 3 respectively, and 64, 128, and 128 filters respectively per convolutional layer for each block. InceptionTime runs for 1,500 epochs with a batch size of 64. InceptionTime is configured with dual residual blocks, each composed of three Inception modules where the sizes of the kernels are sizes 10, 20, and 40 respectively per module, and learns at a rate of 10^{-2}, which, if no improvement is observed for 50 epochs, is set to 5^{-2}.

III. RESULTS AND DISCUSSION

This section analyzes the results of the performance of the proposed method compared to the state of the art. Both the anomaly detection and the interpretability tasks are analyzed and discussed.

A. Anomaly Detection

Table II details the aggregated results for the area under the ROC (auROC) curve. When observing the standard deviation, the deterministic models all have no variability in their results, i.e. 1NN, and the dynamic time warping models. InceptionTime shows the most variability and this property may be somewhat undesirable within the context of assisting domain experts. A good anomaly detection model should have results that are not only accurate, but are also consistent over multiple trials. In this regard of variability, ResNet has the most consistent results from the non-deterministic models.

Table II also indicates that MC-DCNN is the best performing anomaly detection model, followed closely by MES-LSTM, ResNet, InceptionTime and MiniRocket. TapNet is not much better than the deterministic distance-based models.

The box and whisker plot in Figure 1 indicate MES-LSTM achieves the highest performance score on a single trial (highest whisker end-point), although MC-DCNN has the highest overall mean aggregated over all trials. Inception Time and MiniRocket have the highest variability in terms of performance results, whilst ResNet is the most consistent model.

Fig. 2: Area Under the PR curve Distribution Boxplot for all Trials.
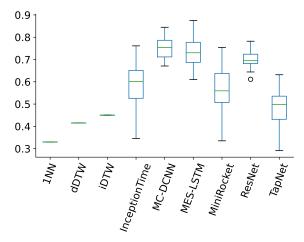

Fig. 1: Area Under the ROC curve Distribution Boxplot for all Trials.

Examining the area under the Precision-Recall (auPR) curve in Table III reaffirms the above discussion. The main difference is in the range of performance scores. The range is higher across all the models as the auPR metric takes into account the class imbalance inherent in the data. The box and whisker plot in Figure 2 further shows the maximum auPR is achieved by InceptionTime, at the cost of the aforementioned variability (which also adversely impacts the overall performance mean). With regards to highest overall performance mean, the top five models are, in order from best, MC-DCNN, MES-LSTM, ResNet, InceptionTime, and MiniRocket.

A Student's t-test for statistical significance is conducted at the $\alpha = 0.01$ level of significance, for the performance results in terms of anomaly detection. The null hypothesis is H_0: the benchmark models outperform MES-LSTM. From Tables IV and Table V, the only instance where one is unable to reject the null hypothesis at the $\alpha = 0.01$ level of significance is for MC-DCNN.

B. Interpretabilty and Explainability

Since there are 96 covariates in total, a starting point would be to consider what a domain expert might consider useful inference from a machine learning tool. In case of power trip, for instance, it would be time consuming to check all 96 possible contributors. However, checking a reasonable subset would be more feasible. Below, β is set to elements in the
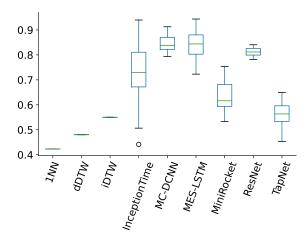
range $\{5, 10, 15\}$, although this range can be determined by requirements specific to a particular use-case scenario. The rationale behind the chosen range is ideally, a good explainer system should rank the chief contributor in the first five highest ranked features (saving the domain expert the most amount of time); an explainer with moderate skill would rank the chief contributor in the five to first ten highest ranked features, while the worst would rank the chief contributor in the first ten to 15 highest ranked features and beyond.

Below, only the top four performing models are considered. Anything that offers less than 50% in accuracy for anomaly detection can be argued to be no better than random guessing. TapNet, although above 50% in performance score, does not report anomaly detection performance skill much higher than the distance-based methods and is also not included in the discussion that follows.

Table VI shows MES-LSTM has the highest correct attribution at $\beta = 5$ features considered, followed by MC-DCNN, InceptionTime and ResNet. At $\beta = 10$ and at $\beta = 15$ features considered, MES-LSTM and InceptionTime are in the top two. ResNet peaks at around 80% correct attribution at $\beta = 15$ features considered, while InceptionTime has the highest overall score at 94.61%. Not one of the models reaches 100%, indicating that there are still some missing key contributing factors not accounted for even after 15 covariates are explored in terms of feature importance. From Table VII it is deducible that at $\beta = 5$ features considered, at most only 73% explainability is accounted for. The maximum score is achieved by MES-LSTM at $\beta = 15$ features considered.

The discovery scores are illustrated graphically in Figure 3. MES-LSTM has the highest correct attribution at all levels of features considered for both LIME and SHAP, except for LIME at $\beta = 10$ (where MES-LSTM is outperformed by InceptionTime).

TABLE II: Area Under Receiver Operating Characteristic curve for all Models over all Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0.3301	0.4146	0.4500	0.5822	0.7547	0.7376	0.5675	0.7012	0.4804
std	0.0000	0.0000	0.0000	0.1025	0.0500	0.0736	0.0925	0.0358	0.0812

TABLE III: Area Under Precision-Recall curve for all Models over all Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0.4225	0.4803	0.5500	0.7332	0.8470	0.8421	0.6359	0.8117	0.5604
std	0.0000	0.0000	0.0000	0.1087	0.0328	0.0549	0.0594	0.0170	0.0471

TABLE IV: Student's t-test for Testing Significance of auROC Performance Results (H_0: Benchmark Models Outperform MES-LSTM).

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MiniRocket	ResNet	TapNet
statistic	32.7568	25.9649	23.1195	7.2838	-1.1331	8.5118	2.6362	13.8820
p-value	0.0000	0.0000	0.0000	0.0000	0.8692	0.0000	0.0056	0.0000

TABLE V: Student's t-test for Testing Significance of auPR Performance Results (H_0: Benchmark Models Outperform MES-LSTM).

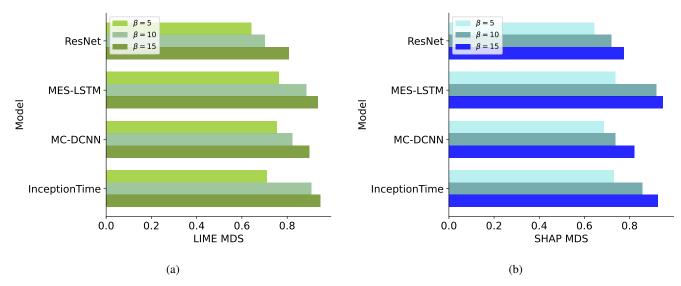|	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MiniRocket	ResNet	TapNet
statistic	45.2511	39.0184	31.5026	5.2917	-0.4495	15.0871	3.1367	23.0563
p-value	0.0000	0.0000	0.0000	0.0000	0.6726	0.0000	0.0016	0.0000

(a)　　　　　(b)

Fig. 3: Mean Discovery Score for Explainer Techniques applied to Top Four Anomaly Detectors. (**A**) LIME. (**B**) SHAP.

TABLE VI: Mean Discovery Score for LIME applied to Top Four Anomaly Detectors.

β	InceptionTime	MC-DCNN	MES-LSTM	ResNet
5	0.7113	0.7554	0.7634	0.6419
10	0.9059	0.8222	0.8856	0.7025
15	0.9461	0.8985	0.9346	0.8077

TABLE VII: Mean Discovery Score for SHAP applied to Top Four Anomaly Detectors.

β	InceptionTime	MC-DCNN	MES-LSTM	ResNet
5	0.7311	0.6871	0.7376	0.6437
10	0.8570	0.7380	0.9179	0.7205
15	0.9263	0.8219	0.9481	0.7754

IV. CONCLUSION

There are two main objectives in this paper. One is to build an anomaly detection machinery that is a hybrid composed of statistical and deep learning techniques. The second is to incorporate interpretability to such a technique. The desired outcomes related to these objectives are one, that the anomaly detection skill is at least competitive to the state-of-the-art; and two, that the explainability component has a good level of correct attribution.

The proposed method is outperformed in some instances with regards to anomaly detection by, for example, MC-DCNN. MC-DCNN is able to model the spatial correlations well, and this could be a contributing factor to the superior performance. MES-LSTM is outperformed marginally and although competitive with regards to anomaly detection, the overall performance is an area for improvement.

However, when it comes to correctly attributing important features for the anomalies detected by each of the top four performing models, MES-LSTM is the overall highest achiever. The high discovery scores are as a result of the architecture's good modeling of temporal dependence. Accurate attribution is important as it ensures the model is not learning from spurious effects. It also reinforces trust for manual inspectors of, say, mechanical systems when an anomaly within the system is detected and possible causes are reported.

The voltage measurements and current measurements are governed by both time evolution from external oscillation events and as well as spatial dependency from the inherent network connectivity over the entire grid. The problem framing is of multivariate spatio-temporal anomaly detection and interpretability. Future work may involve graph neural networks, which show significant promise in the tasks underpinned by similar settings, such as in climate modeling [13].

It is possible that through additional engineering of the benchmark algorithms and tuning of their hyperparameters, better overall performance could have been realized. However, the idea was to test the anomaly detectors based on the configuration recommendations suggested by the respective authors in their original manuscripts. The approach using original configurations mitigates biases that may result from optimising algorithms for particular datasets and particular tasks.

In this study, a novel metric is proposed for assessing usability of a model with regards to usefulness of the model's interpretability to a domain expert. There are many metrics for tasks such as forecasting and anomaly detection, but research centered around explainability is lacking in terms of metrics for ease of comparison among multiple models. Adding more metrics to measure the level of correct attribution is also an avenue for future research.

REFERENCES

[1] Adari S.K. Alla S. *Practical Use Cases of Anomaly Detection.* Apress, Berkeley, CA., 2019. doi: 10.1007/978-1-4842-5177-5_8.

[2] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with cote: The collective of transformation-based ensembles. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 1548–1549, 2016. doi: 10.1109/ICDE.2016.7498418.

[3] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017. doi: 10.1007/s10618-016-0483-9.

[4] Anthony J. Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn J. Keogh. The UEA multivariate time series classification archive, 2018. *CoRR*, abs/1811.00075, 2018. URL http://arxiv.org/abs/1811.00075.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1409.0473.

[6] Alberto Barbado, Oscar Corcho, and Richard Benjamins. Rule extraction in unsupervised anomaly detection for model explainability: Application to oneclass svm. *Expert Systems with Applications*, 189:116100, 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.116100.

[7] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

[8] Leopoldo Bertossi, Jordan Li, Maximilian Schleich, Dan Suciu, and Zografoula Vagena. Causality-based explanation of classification outcomes. In *Proceedings of the*

Fourth International Workshop on Data Management for End-to-End Machine Learning, DEEM'20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380232. doi: 10.1145/3399579.3399865.

[9] Aaron Bostrom and Anthony Bagnall. Binary shapelet transform for multiclass time series classification. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXII*, pages 24–46. Springer, 2017. doi: 10.1007/978-3-662-55608-5_2.

[10] Henrik Brink, Joseph Richards, and Mark Fetherolf. Scaling machine-learning workflows. In *Real-World Machine Learning*, chapter 9. Manning Publications Co., New York, NY, 2016. URL https://livebook.manning.com/book/real-world-machine-learning/chapter-9/8.

[11] Nadia Burkart, Marco Huber, and Phillip Faller. Forcing interpretability for deep neural networks through rule-based regularization. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 700–705, 2019. doi: 10.1109/ICMLA.2019.00126.

[12] Nadia Burkart, Philipp M. Faller, Elisabeth Peinsipp, and Marco F. Huber. Batch-wise regularization of deep neural networks for interpretability. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 216–222, 2020. doi: 10.1109/MFI49285.2020.9235209.

[13] Salva Rühling Cachay, Emma Erickson, Arthur Fender C. Bucker, Ernest Pokropek, Willa Potosnak, Suyash Bire, Salomey Osei, and Björn Lütjens. The world as a graph: Improving el Niño forecasts with graph neural networks, 2021.

[14] Mattia Carletti, Chiara Masiero, Alessandro Beghi, and Gian Antonio Susto. Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 21–26. IEEE, 2019. doi: 10.1109/SMC.2019.8913901.

[15] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. URL https://arxiv.org/pdf/1901.03407.pdf.

[16] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 883–892. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chen18j.html.

[17] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747.

[18] Angus Dempster, François Petitjean, and Geoffrey I Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020. doi: 10.1007/s10618-020-00701-z.

[19] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. *MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification*, page 248–257. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383325. doi: 10.1145/3447548.3467231.

[20] Houtao Deng, George Runger, Eugene Tuv, and Martyanov Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013. ISSN 0020-0255. doi: 10.1016/j.ins.2013.02.030.

[21] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

[22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018. URL https://arxiv.org/pdf/1805.10820.pdf.

[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018. ISSN 0360-0300. doi: 10.1145/3236009.

[24] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 364–379, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356930. doi: 10.1145/3243734.3243792.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. URL https://www.cs.princeton.edu/courses/archive/spring16/cos598F/msra-deepnet.pdf.

[26] Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall. Classification of time series by shapelet transformation. *Data mining and knowledge discovery*, 28(4):851–881, 2014. doi: 10.1007/s10618-013-0322-1.

[27] Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series. *Proceedings of the VLDB Endowment (PVLDB)*, July 2021. URL https://hal.archives-ouvertes.

fr/hal-03381732.

[28] Vimalkumar Jeyakumar, Omid Madani, Ali Parandeh, Ashutosh Kulshreshtha, Weifei Zeng, and Navindra Yadav. Explainit! – a declarative root-cause analysis engine for time series data. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 333–348, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450356435. doi: 10.1145/3299869.3314048.

[29] Rohit J Kate. Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2):283–312, 2016. doi: 10.1007/s10618-015-0418-x.

[30] Eamonn Keogh and Abdullah Mueen. *Curse of Dimensionality*, pages 314–315. Springer US, Boston, MA, 2017. ISBN 978-1-4899-7687-1. doi: 10.1007/978-1-4899-7687-1_192.

[31] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2), 2018. ISSN 2313-433X. doi: 10.3390/jimaging4020036. URL https://www.mdpi.com/2313-433X/4/2/36.

[32] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

[33] James Large, Jason Lines, and Anthony Bagnall. A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data mining and knowledge discovery*, 33(6):1674–1709, 2019. doi: 10.1007/s10618-019-00638-y.

[34] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015.

[35] Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 2018. doi: 10.1145/3182382.

[36] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

[37] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2021. URL https://openreview.net/pdf?id=A5VV3UyIQz.

[38] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017. URL https://proceedings.neurips.cc//paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[39] Minghua Ma, Zheng Yin, Shenglin Zhang, Sheng Wang, Christopher Zheng, Xinhao Jiang, Hanwen Hu, Cheng Luo, Yilin Li, Nengjun Qiu, et al. Diagnosing root causes of intermittent slow queries in cloud databases. *Proceedings of the VLDB Endowment*, 13(8):1176–1189, 2020. URL http://nkcs.iops.ai/wp-content/uploads/2020/04/paper-VLDB20-iSQUAD.pdf.

[40] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[41] Thabang Mathonsi and Terence L van Zyl. Multivariate anomaly detection based on prediction intervals constructed using deep learning. *Neural Computing and Applications*, pages 1–15, 2022. doi: 10.1007/s00521-021-06697-x.

[42] Thabang Mathonsi and Terence L. van Zyl. A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. *Forecasting*, 4(1):1–25, 2022. ISSN 2571-9394. doi: 10.3390/forecast4010001.

[43] Matthew Middlehurst, William Vickers, and Anthony Bagnall. Scalable dictionary classifiers for time series classification. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 11–19. Springer, 2019. doi: 10.1007/978-3-030-33607-3_2.

[44] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596.

[45] Mohsin Munir, Muhammad Ali Chattha, Andreas Dengel, and Sheraz Ahmed. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In M. Arif Wani, Taghi M. Khoshgoftaar, Dingding Wang, Huanjing Wang, and Naeem Seliya, editors, *ICMLA*, pages 561–566. IEEE, 2019. doi: 10.1109/ICMLA.2019.00105.

[46] Rodica Neamtu, Ramoza Ahsan, Elke A. Rundensteiner, Gabor Sarkozy, Eamonn Keogh, Hoang Anh Dau, Cuong Nguyen, and Charles Lovering. Generalized dynamic time warping: Unleashing the warping power hidden in point-wise distances. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 521–532, 2018. doi: 10.1109/ICDE.2018.00054.

[47] Quoc Phong Nguyen, Kar Wai Lim, Dinil Mon Divakaran, Kian Hsiang Low, and Mun Choon Chan. Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In *2019 IEEE Conference on Communications and Network Security (CNS)*, pages 91–99. IEEE, 2019. doi: 10.1109/CNS.2019.8802833.

[48] Guansong Pang, Chunhua Shen, Longbing Cao, and

Anton van den Hengel. Deep learning for anomaly detection: A review. *CoRR*, abs/2007.02500, 2020. URL https://arxiv.org/abs/2007.02500.

[49] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021. URL https://arxiv.org/pdf/2108.00462.pdf.

[50] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/b495ce63ede0f4efc9eec62cb947c162-Paper.pdf.

[51] Bijan Rad, Fei Song, Vincent Jacob, and Yanlei Diao. Explainable anomaly detection on high-dimensional time series data. In *Proceedings of the 15th ACM International Conference on Distributed and Event-Based Systems*, DEBS '21, page 2–14, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385558. doi: 10.1145/3465480.3468292.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

[54] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

[55] Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021. URL https://ueaeprints.uea.ac.uk/id/eprint/77815/7/Ruiz2020_Article_TheGreatMultivariateTimeSeries.pdf.

[56] Tara N. Sainath, Brian Kingsbury, Abdel-rahman Mohamed, George E. Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y. Aravkin, and Bhuvana Ramabhadran. Improvements to deep convolutional neural networks for lvcsr. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 315–320, 2013. doi: 10.1109/ASRU.2013.6707749.

[57] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf.

[58] Oscar Serradilla, Ekhi Zugasti, Julian Ramirez de Okariz, Jon Rodriguez, and Urko Zurutuza. Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data. *Applied Sciences*, 11(16), 2021. ISSN 2076-3417. doi: 10.3390/app11167376.

[59] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, 31(1):1–31, 2017. doi: 10.1007/s10618-016-0455-0.

[60] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017. URL http://proceedings.mlr.press/v70/shrikumar17a/shrikumar17a.pdf.

[61] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access*, 7:67027–67040, 2019. doi: 10.1109/ACCESS.2019.2912823.

[62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL https://arxiv.org/pdf/1312.6034.pdf.

[63] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[64] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

[65] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. doi: 10.1109/TNNLS.2020.3027314.

[66] Manjunatha Veerappa, Mathias Anneken, and Nadia Burkart. Evaluation of interpretable association rule mining methods on time-series in the maritime do-

main. In *International Conference on Pattern Recognition*, pages 204–218. Springer, 2021. doi: 10.1007/978-3-030-68796-0_15.

[67] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966039.

[68] Adam Westerski, Rajaraman Kanagasabai, Eran Shaham, Amudha Narayanan, Jiayu Wong, and Manjeet Singh. Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. *International Transactions in Operational Research*, 28(6): 3276–3302, 2021. doi: 10.1111/itor.12968.

[69] Chongke Wu, Sicong Shao, Cihan Tunc, Pratik Satam, and Salim Hariri. An explainable and efficient deep learning framework for video anomaly detection. *Cluster Computing*, 2021. doi: 10.1007/s10586-021-03439-5.

[70] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian rule lists. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3921–3930. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/yang17h.html.

[71] Xuchao Zhang, Yifeng Gao, Jessica Lin, and Chang-Tien Lu. TapNet: Multivariate time series classification with attentional prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6845–6852, 2020. doi: 10.1609/aaai.v34i04.6165.

[72] Xiangtian Zheng, Nan Xu, Loc Trinh, Dongqi Wu, Tong Huang, S Sivaranjani, Yan Liu, and Le Xie. Psml: A multi-scale time-series dataset for machine learning in decarbonized energy grids. *arXiv preprint arXiv:2110.06324*, 2021. URL https://arxiv.org/pdf/2110.06324.pdf.

[73] Xiangtian Zheng, Nan Xu, Dongqi Wu, Loc Trinh, Tong Huang, S Sivaranjani, Yan Liu, and Le Xie. PSML: A Multi-scale Time-series Dataset for Machine Learning in Decarbonized Energy Grids (Dataset), August 2021.

[74] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *International conference on web-age information management*, pages 298–310. Springer, 2014. doi: 10.1007/978-3-319-08010-9_33.