XES^{3}MaP: Explainable Risks Identified from Ensembled Stacked Self-Supervised Model to Augment Predictive Maintenance

Sarala M Naidu [1,2]
School of Innovation, Design and Engineering
[1]Mälardalen University, [2]Hitachi Energy
Västerås Sweden
sarala.mohan.naidu@mdu.se

Ning Xiong
School of Innovation, Design and Engineering
Mälardalen University
Västerås Sweden
ning.xiong@mdu.se

***Abstract*— Understanding the reasons behind a model's predictions is as important as achieving accurate results. The limited adoption of AI methods in fields like Energy and Industry is mainly attributed to the lack of trust, which is a crucial factor in user acceptance. Explainable AI is a recent approach to address this issue and enable the rapid deployment of AI in complex domains. This paper presents a framework for explainable anomaly detection and risk prognostics that utilizes an Ensembled Stacked Self-Supervised Model to augment predictive maintenance (XES^{3}MaP). The explanation produced is evaluated utilizing local accuracy metric. The proposed framework is tested on real-world industrial cooling system data, on which out of the ten anomalies identified, eight were successfully linked to maintenance actions, while two were attributed to random sensor measurement disturbances. This prognostic outcome approach establishes a new benchmark in the field.**

Keywords— anomaly detection, ensemble method, stacking, SHAP, shapely values, Explainable AI

I. INTRODUCTION

Predictive maintenance (PdM) serves as the backbone of industrial operations. It utilizes sensor data to safeguard asset health by identifying, analyzing, and tracking degradation, ultimately predicting failure progression. The three key activities leveraged in PdM are:

i. Anomaly Detection: Identifying deviations from normal system behavior through data analysis.

ii. Failure Prognosis: Estimating the remaining useful life (RUL) of an asset before failure occurs.

iii. Diagnostics: Classifying and pinpointing the root cause of equipment failure.

However, Explainable AI (XAI) within the prognostic and health management (PHM) domain is still a nascent field. Widespread adoption of XAI knowledge by PHM stakeholders, both in research and industry, is crucial for building trust in AI and facilitating its legal implementation within the industrial sector. In industrial settings, anomaly detection plays a crucial role in predictive maintenance (PdM), enabling the identification of potential failures or malfunctions in industrial equipment. However, traditional maintenance approaches often fail to provide insights into the data patterns that drive maintenance decisions. Anomaly detection algorithms powered by deep learning (DL) have been successfully identifying anomalies that might have gone unnoticed otherwise. However, a major drawback lies in their lack of explainability, posing challenges in convincing domain experts to trust and adopt these systems. Explanations of why an instance is flagged as anomalous can enhance trust and enable experts to effectively address potential issues identified. The need for instance-level explanations was recognized in the 1970s to model the inexact reasoning process of medical experts [1], but has resurfaced due to the increasing complexity of ML models.

XAI can address this challenge by illuminating anomalies and providing explanations for their occurrence. This enhanced understanding can guide proactive maintenance interventions, minimizing downtime and maximizing equipment lifespan. One such application for the explainability is for risk associated in an industrial cooling system employed in power transmission systems. Industrial cooling systems play a critical role in maintaining optimal operating temperatures by effectively dissipating the heat generated in thyristor valves within Converter Stations during power transmission. Regular maintenance of these cooling systems is essential to safeguard their optimal performance and prevent malfunctions. Conventional preventive and scheduled maintenance practices are currently employed, relying on FMECA (Failure Mode Effects and Criticality Analysis) to identify relevant failure modes. These failure modes are then analyzed using data from installed equipment sensors and measurement interfaces.

II. USE-CASE DESCRIPTION

Industrial systems deal with different types of anomalies that can turn into risks. This risk is a probability when there is deviation from the normal expected behavior. Considering a specific variable that is of concern according to the maintenance need through anomaly detection, the target variable is identified as 'liquid conductivity'. The electrical conductivity is a measure of the quality of the liquid used for cooling. Each cooling liquid has a relative constant range that is established as a baseline for comparing with the measurements during operations. The measurement of conductivity is very important in industrial application as it indicates the presence of the minerals, chemicals and other dissolved substances. When the dissolved ions increase, there is a high conductivity as ions carry electric charges in the liquid, that results in a certain amount of current passed through increasing the conductance level. This is harmful to the electronics in the object being cooled. In general, the conductivity is influence by changes temperatures. As temperature increases, the mobility of ions in the liquid changes which results in a change in conductivity. Thus, it is essential to maintain conductivity at stable levels through a

circuit for treatment where the impurities are removed. Then the treated water conductivity is again measured as it flows into the object being cooled. So the risks involved here are a) the treatments circuit may malfunction in performing the treatment process, b) the conductivity measurements may be high after the treatment circuit, c) there could be electric charges flowing to the electronics of the cooling object that is harmful. The use-case of the cooling system is already a complex circuit of operation with multiple variables of which some are derived and others measured. Any change in value of one variable, the values of the other variables also change, while few are designed to be stable. The related problem is with the assumption that there is knowledge of the causal structure behind data.

While identifying abnormal values through pre-defined thresholds is crucial, it is equally important to understand the root cause behind these deviations (high or low) to take appropriate actions. Current practices rely on manual analysis by domain experts, which can be time-consuming and resource-intensive. This highlights the need for a more efficient approach to enable faster interpretation and provide feature-level explanations for the model's predictions. By automatically pinpointing the cause of abnormalities, the system can facilitate quicker decision-making and improve overall process efficiency.

III. Related Work

This section summarizes the PdM related XAI works associated with anomaly detection and failure diagnosis. The work in [4, 17] applies KernelSHAP, to generate local explanations for each anomaly detected by autoencoders. The paper [18] introduces a global XAI method, allowing for a more comprehensive understanding of the model's decision-making process. The paper [20] presents a survey of various XAI methods on multivariate industrial data, indicating that SHAP based explainer correctly identified the root cause of the anomalies. The work in [6] evaluates the effectiveness of anomaly detection and explainability algorithms to supplement Decision Support System insights in PdM and Root Cause Analysis for hydroelectric power plants. The results show that isolation forest and autoencoder performed the best and the use of SHAP provided explanation to the root cause and guided the user towards features related to anomaly. The paper [11] applies advanced attention based convolutional autoencoder for anomaly detection in industrial use-case, to identify risks. The [23] applies LSTM-autoencoder for anomaly detection and explaining fault localization on multivariate data using ground truth provided by PLC. The paper [19] presents a framework for explainable anomaly detection and failure prognostic, combining a Bayesian DL model and Shapley additive explanations being effective on real-world gas turbine data. The research on bearings using DSASA (dynamic structure–adaptive symbolic approach), a cross-domain life prediction model that used historical run-to-failure data in [19] elaborates the RUL prediction. In [25] autoencoder anomaly detection is assisted by XAI. The paper [14] uses feed-forward neural network architecture with local interpretable model agnostic explainability for fouling prediction in the crossflow heat exchanger.

Despite a culture of continuous improvement, surprisingly, industrial, manufacturing, and energy sectors are lagging in adopting AI for daily operations [12]. This apparent contradiction points to a key hurdle: trust. Unlike other sectors facing potential ethical concerns, industrial actors struggle with trusting AI decisions due to performance anxieties. Regulations surrounding AI often focus on transparency, fairness, privacy, and data security of algorithms. Transparency plays a crucial role in minimizing malfunctions and achieving desired AI quality goals. This is especially important due to the "black-box" nature of some AI techniques. DL, currently the most powerful AI method, is a prime example of a black-box model. While highly effective, its internal workings in generating predictions remain obscure. This opacity hinders AI adoption in high-risk sectors like industry and energy, where incomprehensible outcomes could lead to disastrous consequences for life, safety, and finances. These domain experts require more than just point estimates to be convinced of a course of action.
The onus lies with the research community to bridge this trust gap. This is where XAI comes into play. XAI techniques aim to make AI models interpretable, allowing users to understand the rationale behind their predictions. By addressing the transparency issue, XAI can pave the way for wider AI adoption across various sectors. The field of XAI focuses on making AI models interpretable to humans. While the concept has existed for decades, recent years have witnessed a surge in global attention, as evidenced by initiatives from organizations like the Defense Advanced Research Projects Agency (DARPA) since 2016 [13]. This growing interest is further reflected in the increasing number of research articles dedicated to XAI. Therefore, the main objectives of this research are:

i. To process and prepare real industrial data for a case study use-case.
ii. To construct anomaly detection model suing an ensemble of weak and strong learners for unsupervised use-case
iii. To generate the prediction labels for anomaly detection
iv. To demonstrate the SHAP explanation’s ability to improve prognostic task’s performance, which was absent from previous works.
v. Derive the SHAP plot and interpret it.
vi. Apart from SHAP plots, a custom function to derive the anomaly contribution score makes it easy for onwards task to read them and use the value, impossible with visual chart.
vii. To evaluate the explainability

Note, that the goal of this paper is not to compare with other related algorithms, but to focus on highlighting the power of explainability in a complex and conservative power domain. Secondly, due to the novelty of the different models stacked with the meta-learner, there is one to one comparison. The secondary objectives are:

i. To add model agnostic explainability to the collection of PHM-XAI articles, which is still lacking currently.
ii. To validate the efficiency property of Shapley values and prove local accuracy of the explanation.

While the application domain of the proposed approach is novel, the use of all variants of autoencoder for learning

features from the same data, 5 model ensemble with meta learner, custom function to derive the explainability instead of the standard XAI SHAP plot (contribution vi) , also add to the novelty.

This rest of the paper is organized as follows. The methodology is presented in Section III, the results and discussion on the case study, in Section IV. Finally, the concluding remarks are given in Section V

IV. Methodology

This section describes the individual methods involved, followed by the overall proposed integrated framework.

A. Ensemble Method:

Ensemble is a collection of independent ML algorithms, have emerged as a powerful approach for anomaly detection, offering enhanced accuracy and robustness compared to individual algorithms for anomaly detection in real-world applications. A single anomaly detection algorithm may excel in identifying anomalies in specific datasets but struggle with others due to inherent limitations. Ensemble methods address this challenge by combining multiple algorithms, effectively mitigating the weaknesses of any single algorithm, and leveraging their collective strengths. This collective intelligence enables ensemble methods to provide more accurate and consistent anomaly detection results. By combining diverse algorithms, ensemble methods can achieve a more balanced trade-off between bias and variance, leading to improved generalization performance. Systematic literature reviews in [21, 22], highlight the effectiveness of ensemble methods to enhance the performance, in terms of generalization and robustness. This ability is crucial for anomaly detection, as it ensures that the model is not overly sensitive to noise or outliers while maintaining the ability to capture meaningful patterns in data.

B. Number and type of Base models

With an inspiration from [8] on ensemble deep learning, it is evident that two types of strategy are suitable for deep learning viz. (i) applying many different basic models using the same data and (ii) applying many different basic models using many different data samples. Building effective ensemble deep learning systems goes beyond just model selection. A well-designed architecture is key, requiring decisions on model types suited to the problem (often 3+ models) [8], optimal data splits (e.g., 80/20), and the entire deep learning pipeline. This pipeline includes data generation, training individual models, and choosing the best method to combine their outputs (fusion). Optimizing these elements is crucial for a powerful ensemble system. The homogeneous form of ensemble is challenged by the generation of diversity from the same learning algorithm. While heterogeneous ensembles consist of different numbers of baseline classifiers, each based on the same data and, the feature selection method is different for same training data.

C. Stacking Method:

In the stacking method the baseline learners are used simultaneously, as there is no data dependency, the fusion methods depends on the meta-learning method. Here, the predictions are from several models say m_1 to m_n to build a new model, and the new model makes predictions on the test dataset. Thus, it seeks to increase the predictive power of a model. The basic idea of stacking is to "stack" the predictions of m_1 to m_n. Here the baseliners adopted are autoencoder models and this is based on the success of autoencoder in various applications [11,23,24]. Autoencoders are neural networks that learn efficient representations of data. They compress the input data into a smaller hidden space (encoding) and then try to reconstruct the original data from that compressed version (decoding). By forcing the network to recreate the input, it learns useful features of the data. This difference , called the reconstruction error ϵ, is computed by the Euclidean distance given by "(1)".

$$\epsilon = \sqrt{(x_i - x_i')^T \cdot (x_i - x_i')} \quad (1)$$

The inputs are then classified as either "Anomaly" or "Normal" based on their reconstruction error, ϵ. If ϵ exceeds a predefined threshold value, the input is labeled as "Anomaly". Otherwise, it is labeled as "Normal". The threshold is determined as the max of the mean absolute error (mae) on the training data and is used as threshold on the test data to detect anomalies. The different base learners are different variation of the autoencoders. Two crucial components in the architecture of a stacking ensemble comprises of:

- Base Models: Here five individual DL algorithms based on autoencoders, and its variants are used, to generate initial predictions labels for anomalies. Base models applied are conventional Autoencoder (AE), variational autoencoder, and three hybrid autoencoders namely, Convolutional Autoencoder (CAE), LSTM Autoencoder (LSTM-AE) and LSTM Variational Autoencoder (LSTM-VAE). The choice of the number of neurons in each layer, the latent space size, optimization of parameters, loss function for monitoring the training and testing is made accordingly.
- Meta-Model: This is the super learner model that typically employs a simpler structure compared to the base models, allowing for a more interpretable and transparent representation of the ensemble's decision-making process. Here a 3-layer ANN model is used.

The base learning algorithms widely used in multiple experiments have been traditional ML methods like support vector machines (SVMs), decision trees (DTs), and ANNs. For unsupervised use-case autoencoders and its variants have been successful in learning the representations for anomaly detection, as in [22-25]. The specific choice of base learners depends on the characteristics of the anomaly detection task and the desired trade-off between accuracy, computational efficiency, and interpretability. To train the meta-model, the predictions made by the base models on a test dataset, to ensure that it is not biased and can generalize well to unseen instances. In this case each base model will output prediction (of anomaly or normal), which are then concatenated to form the stacked data input to the meta learner. The effectiveness of stacking hinges on the correlation between the predictions of the base models. When the predictions are uncorrelated or weakly correlated, the meta-model can effectively identify patterns and trends that individual models may miss. This synergy leads to improved prediction accuracy. Supporting the learning algorithms to better adapt to changing conditions, speed up learning processes by reducing the

number of experiments and optimizing hyperparameters to achieve optimal results are the benefits from the meta-learner. Note: How the anomalies are detected and a comprehensive explanation of approach using self supervision with autoencoders, its hybrid variants, their underlying concepts, or the mathematical formulas involved is not the focus of this paper, as they are covered in [9, 10].

D. SHAP

SHAP is a model-agnostic approach for generating explanations to the model predictions by computing the contribution of each features to the prediction. Shapley values stand out as a promising explanation methodology with a strong mathematical foundation and unique theoretical properties rooted in cooperative game theory, introduced by Lloyd Shapley in 1953 [7], where the key component is the "shapley value" that assess the influence of a player to form the coalitions. Thus, SHAP identifies the feature's (the players in the game theory) impact on the overall prediction. It focuses on explaining a model's 'f' prediction at a specific point, x' and these explanations are based on a value function, V_S, which represents the model's prediction at x' after setting a subset of variables, S, to specific values as in "(2)".

$$v_s = E[f(x)|x_s = x_s'] \tag{2}$$

We consider the individual contribution as well as the interaction between the features. There could also be different ways to share the contribution in a fair way using the shapley values. From the cooperative games theory the shapley values were originally proposed to distribute payouts fairly using the four axioms viz. efficiency, symmetry, dummy, and additivity. Efficiency dictates that the total payout across all features should equal the model's prediction. Symmetry demands that features with identical contributions receive equal payouts. Dummy asserts that a feature with no contribution should receive a payout of zero. Additivity ensures that the payout for a coalition of features is the sum of their individual payouts. To assess the importance of a feature 'i' is based on analyzing how the set S will affect the function Vs, if the feature 'i' is added to S. Contribution φ(i), of the feature 'i' on the prediction f(x) is:

$$\phi_i(f) = \beta_i x_i - E(\beta_i x_i) \tag{3}$$

where E(βjXj) is the mean effect estimate for feature. The contribution is the difference between the feature effect minus the average effect. Feature contributions can be negative. Thus, the contribution of the shapley value is to payout, weighted and summed over all possible feature value combinations given by "(4)":

$$\phi_i(v_s) = \sum_{S \subseteq \{1,..p\} \setminus \{i\}} \frac{|S|!(p-|S|-1)!}{p!} (v_s(S \cup \{i\}) - v_s) \tag{4}$$

with S being the subset of features in the model, x being the vector of feature values of the instance to be explained and the number of features is p. Shapley values have several desirable properties, such as ensuring that features that do not contribute to the prediction get an attribution of zero. The feature value represents the numerical or categorical value of a feature in a specific instance. It serves as the input to the Shapley value calculation, which determines the feature's contribution to the model's prediction. Pipeline to generate XAI explanations on the unsupervised ML model is shown in Fig. 1. With the stacked base model outputs as input to train the meta-model to generate the output target labels (y'ℰy). This final output labelled dataset 'd' (supervised data) from meta model is fed to a surrogate neural network models (M) that is trained with its own parameters (p). The key goal of this surrogate model is to minimize the loss function L(p,d) and is then used to generate the explanations using the feature attributions technique with shapely values. The benefit of this approach is that it can provide both local and global explanations.

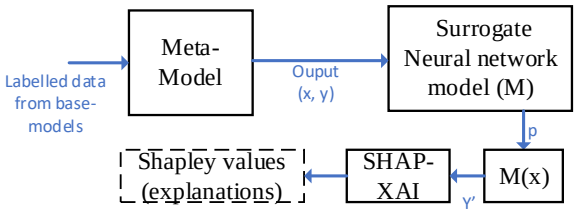


Fig. 1: Pipeline of XAI - SHAP explanation

However, Shapley values are computationally expensive, especially for large models or high-dimensional inputs. The SHAP Python library is used to calculate SHAP values, and also visualize the importance of features using plots.

E. Explainability

XAI methods are distinguished into model-specific interpretability technique that depend on the internal structure of the specific ML model, and mode-agnostic interpretability where the explanatory component and the ML models are independent from each other. The paper [2] demonstrates that many existing saliency methods fail to maintain consistency, leading to misleading attributions and encourage the community to develop additional tests to evaluate the reliability of saliency methods. XAI research seeks to mimic human explanation methods that rely on researcher intuition rather than from psychology and cognitive science. [3, 5] surveys explainability with autoencoder. A post-hoc explainability is being considered here where the explanation is generated after the model is trained. Based on the scope, XAI methods can also be classified as local or global explanations.

1) Local Explanations:

Local explanations provide insights into how specific features contribute to a particular prediction made by a ML model. They focus on understanding the relationship between the input features and the model's output at a single data point. This type of explanation is useful for understanding how the model makes decisions on individual instances. The SHAP values for each individual prediction is calculated to identify how the features contribute to that single prediction.

2) Global Explanations:

Global explanations, provides a broader understanding of how the model works as a whole with the overall patterns and relationships between the features and the model's predictions across the entire dataset. This type of explanation is useful for understanding the model's behavior in general and identifying the most influential features. SHAP values not only show feature importance but also show direction whether the feature has a positive or negative impact on predictions.

Local and global explanations serve different purposes and complement each other. In many cases, it is helpful to use a combination of both local and global explanations to gain a comprehensive understanding of a ML model. SHAP offers various approximations to suit different use cases, in this work, KernelSHAP that combines the Linear LIME [25] and

SHAP algorithms is applied. Multiple measures for evaluating explanations have been proposed in ISO [26]. One such property is the local accuracy of the SHAP examined using a waterfall plot. It establishes that the sum of the feature contributions, is equal to prediction of x or $f(x)$, minus the average prediction, E(f(x)).

The SHAP values translate to indicate a directionality to show how features impact the output in a more intuitive way, with a plus in red color indicating positive impact on the prediction and minus in blue means negative impact. For model explainability in machine learning, the SHAP values helps understand the model at row and feature level.

F. Explanation Visualization

This section details the visualization methods used to represent local and global explanations:

- Local Explanations: Force plots and waterfall plots are employed to illustrate the impact of individual features on a specific instance's prediction.
    - Force Plot: This plot uses colored bars to represent the contribution of each feature. The bar length corresponds to the strength (positive or negative) of the feature's influence on the prediction. Red bars indicate features pushing the prediction upwards, while blue bars indicate features pulling it downwards. This visualization is particularly useful for understanding anomalies.
    - Waterfall Plot: This plot arranges feature contributions in a bar-like format, with the most influential features at the top and the least influential at the bottom, resembling a waterfall. Color-coding remains consistent with the force plot, clearly showing the direction of each feature's influence. Waterfall plots are used to verify the local accuracy and consistency of explanations.
- Global Explanations: Summary plots are used to highlight the most influential features across a dataset. These plots rank features based on their contribution strength and the direction of their forces. This information was leveraged to improve prognostic accuracy by focusing on the most critical features.

In this paper, a custom function is also written to give a quantitative representation of the explanations.

G. Human-in-the-loop

A radical approach is to embed human experts into the ML workflow, integrating them into physical feedback loops [15,17], continuously providing feedback to the learning system to improve its performance and optimize its parameters. This synergistic approach harnesses the strengths of both human intelligence and machine learning, leading to more effective and reliable decision-making [16]. To ensure model confidence and reliability, a domain expert conducts an iterative validation process. This is to evaluates the model's outputs and provides feedback to refine the model parameters or training data. This loop continues until the model's performance meets the desired level of confidence. Thus, in this paper, apart from formal validation methods, the domain expert also validates the results for interpretation. Specific evaluations is done in each application domain with experts' supervision.

H. Overall Integerated Framework

Using all the individual methods and concept described, an end-to-end framework named as XES^{3}MaP: Explainable Risks Identified from Ensembled Stacked Self-Supervised Model to Augment Predictive Maintenance is proposed. The two main parts integrated are the anomaly detection model and the explainability model (depicted in Fig. 1). The overall approach with steps in the workflow numbered and marked is shown in Fig. 2

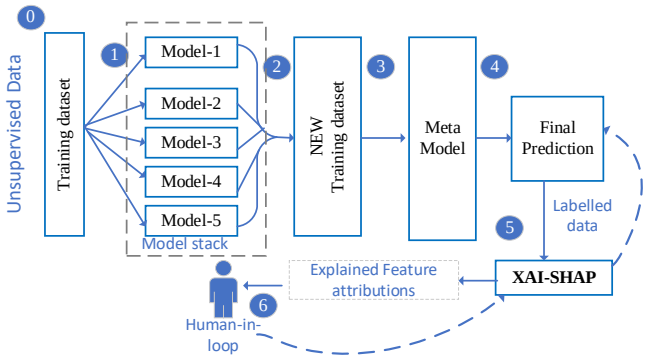


Fig. 2: Overall framework

1) Step 0 and 1: Data preparation

The operational data is aggregated from the sensor measurements of the cooling subsystem using SCADA system of a real HVDC station, anonymized various features, including temperature, cooling liquid conductivity, liquid level, pressure, and flow. The data collection period spans from January 2020 to June 2023 and was preprocessed to resample to uniform frequency, removing missing and noisy data with support of domain experts. The dataset is normalized using StandardScalar, to transform each feature to have a mean of 0 and a standard deviation of 1.

2) Step 2: Base Model Training

The five identified base models are trained individually on the processed dataset that is split for training and testing using 80-20% strategy. Experiment parameters are as in TABLE I.

TABLE I. EXPERIMENT PARAMETERS

parameters	***Value range used for best results***
TimeSteps	24 hr for convolutional & LSTM models
Batch size	16 to 64
Learning rate	0.01 to 0.0001
Epochs	50 to 500 with callback to montior the loss, with patience=5, on a validation split of 10%
Activiation function	Relu, Leakyreul,tanh & sigmoid
No. of layers	3 dense layer in each model

By using cross-validations (5-fold), the best hyperparameters that result in the best generalization performance are selected. This approach balances between training the model effectively and avoiding overfitting, which occurs when the model memorizes the training data too well.

3) Step 3 and 4: Meta Model training

The prediction outputs of the base models on the test data set are concatenated to form the new stacked input (step 3 in Fig. 2) to the meta model. Then

the meta-model is trained in isolation, to get its own final predictions of the anomalies.

4) Step 5 and 6: Explanations with SHAP

Once, the final prediction is made by the meta model, the pipeline steps of the XAI-SHAP process as shown in Fig. 1 is triggered. The explanation is derived for the anomalous instances where the difference (error) between the input and the reconstructed value is high. The explanatory models provide explanation for why an instance was marked as anomaly by the predicting model. The method here is to compute the SHAP values of the reconstructed features and relate them to the true values of the input that is anomalous.

V. Results and Discussions

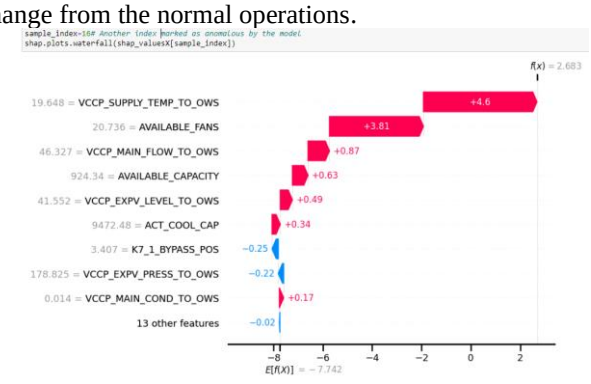
The overall results in TABLE II. shows that the mae score of meta models is higher than the individual base model.

TABLE II. Results from all models

Model Name	*AE*	*VAE*	*CAE*	*LSTM-AE*	*LSTM-VAE*	*Meta-Model*
MAE score	0.43	0.453	0.007	0.116	0.484	0.0064

Now to apply the SHAP XAI to explain why an anomaly is flagged as anomalous by the final model. SHAP Plot of few indexes of the test data samples that is flagged as an anomaly is shown in Fig. 3. Note, the name of the parameters are short abbreviation used in the experiment and random indexes of 2 samples [index 16 & 10] marked as anomalous by the model is taken for the SHAP interpretation. The features marked by the red in each of the plot cause an increase of the contribution to anomaly value while the features marked in blue causes a decrease. Taking the case of index 16 (first image), an increase in supply temperature is the main cause of the anomaly, and in turn the change in available number of fans to cool the system. This justifies the fact that increase in supply temperature, requires more number of fans to be turned ON to achieve required cooling. And here, to turn them ON (automatic by SCADA), sufficient fans may not be ready (among the groups of fans in the system) or may be unhealthy. The service engineer can check the status and take necessary actions accordingly.

Now, taking the index 10, the flag is due to changes in the liquid flow values (that is kept almost stable during operation). This was identified to be due to the opening of the bypass switch as also indicated as the key contributing parameter. As seen in Fig. 3 the features that assert positive impact push output value higher. In this case, the bypass opening/closing is normal operation process during different outdoor weather conditions. And needs to be accounted in the data for training. And thus, by domain this is not an anomaly as it is a sign of the actual operation process where there is a change from the normal operations.



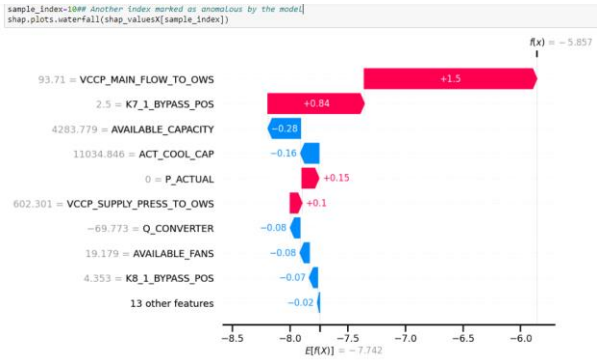


Fig. 3: SHAP Interpretation for 3 anomalous indexes

Now, validation of the local accuracy for index 16, using "(3)" and "(4)" is as follows:

f(x)=2.683, Ef(x)=-7.742, and $\phi_i = 2.683 - (-7.742) = 10.42$

$$\phi_i(v_s) = 4.6 + 3.81 + 0.87 + 0.63 + 0.49 + 0.34 + 0.17 - 0.25 - 0.22 - 0 - 02 = 10.42$$

Similarly for index 10, it is :

f(x)=-5.857, Ef(x)=-7.742, and $\phi_i = 1.885$

$$\phi_i(v_s) = 1.5 + 0.84 + 0.15 + 0.1 - 0.28 - 0.16 - 0.08 - 0.08 - 0.07 - 0.02 = 1.9$$

Alternatively, among the visual plot representations of shap value, a custom function to return a dictionary with explanations for top N records that contribute highest to the anomaly prediction as shown in Fig. 4.

```
1
{ 141: [('VCCP_MAIN_COND', -1),
 ('VCCP_RETURN_TEMP', 0.004146357662135036),
 ('MAX_COOL_CAP', 0.0033244182963671413),
 ('VALVE_LOSSES', 0.0028682962521391544),
 ('VCCP_N2_PRESSURE', 0.001784277824727361),
 ('VCCP_EXPV_LEVEL', 0.001368210596257032),
 ('VCCP_E1_BQ4_COND', -1),
 ('VCCP_EXPV_PRESS', 0.0012785825096122906)],
2  143: [('VCCP_E1_BQ4_COND', -1),
 ('VCCP_MAIN_COND', 0.0037743146629229595),
 ('MAX_COOL_CAP', 0.002866317016418176),
 ('VCCP_N2_PRESSURE', 0.0022480050375740354),
 ('VCCP_EXPV_LEVEL', 0.0017460985295906725),
 ('VALVE_LOSSES', 0.0013851627739505151),
 ('VCCP_EXPV_PRESS', 0.0013500520724260230)],
```

Fig. 4: Custom function showing highest contributor for anomaly

Here, taking one example of the explanation of a randomly picked index [141] marked as '1' in red indicates that the anomaly is mainly due to 2 features, that is values of conductivity both in main and treatment circuit (marked as VCCP_MAIN_COND and E1_BQ4_COND). While there are different ways one can validate the explainability, one of the reason could be that the treated liquid conductivity measurement is not as expected, due to which the main conductivity value is affected.

VI. Conclusion

This paper presents a method for interpreting anomaly prediction results from the model, that can be utilized in various real-world domains including other industrial sectors. It is necessary to be aware of the data-driven nature of the methods, as there is no automatic way to guarantee the correctness of the interpretation. While the decision-making processes is supported with this method, it does not replace the critical roles of human experts. XAI tries to overcome the

"black-box" apprehensions from the industrial actors who struggle with trusting AI decisions due to performance anxieties. It illuminates anomalies and provides explanations for their occurrence. This enhanced understanding can guide proactive maintenance interventions, minimizing downtime and maximizing equipment lifespan. The adaptation of model agnostic KernelSHAP XAI to explain the final anomalies detected. Such type of explanation either in form of visual plots or with numerical values explain the feature contributions to speed up service or maintenance activity predictively to diagnose abnormal scenarios rather than being reactive. Thus, bridging the gap between data-driven insights and human comprehension in the industrial application.

The results show how the post-hoc SHAP XAI technique describes why it made certain decisions of the anomaly detected. The case study example shows how feature contributions define the anomaly index to provoke a maintenance action or a further validation step (like in case of the sensor disturbance). The explainable AI algorithm SHAP makes use of the shap values. Each feature contributes differently to the prediction. The shap values is used to find out the marginal contribution of each feature in the prediction. The shapley values formula is adapted to explain models, by dividing the model prediction amongst it features. The shapley values tell us the average contribution of the features to the predictions. The multivariate features, have the correlations among the features and the interpretability also brings out these aspects. As a result, the XAI explanation gave insights into the reasonings by making the whole system transparent. Then, the explanation produced is evaluated utilizing local accuracy metric. As the goal is to show the value of the XAI in the industrial domain to ensure Trust in the Deep learning models predictive performance among the stakeholders, there is no comparison study carried with other related algorithms.

The diversity in ensemble deep learning with several data samples is limited by the computation cost and the availability of suitable data to be sampled. The computational complexity of the ensemble approach is an additional essential aspect to consider among others like predictive performance accuracy. The computational cost is distributed on two complexity metrics: cost of training and creating the ensemble model and the cost of predicting a new instance.

ACKNOWLEDGMENT

This work is supported by the Swedish Foundation for Strategic Research (Project: ID20-0019).

REFERENCES

[1] E. H. Shortliffe and B. G. Buchanan, “A model of inexact reasoning in medicine,” Mathematical Biosc, vol. 23, no. 3, pp. 351–379, Apr. 1975.

[2] P.-J. Kindermans et al., “The (Un)reliability of Saliency Methods,” in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., in Lecture Notes in Computer Science. , Cham: Springer International Publishing, 2019, pp. 267–280.

[3] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019.

[4] L. F. Antwarg, R. Miller, B. Shapira, and L. Rokach, “Explaining anomalies detected by autoencoders using shapley additive explanations,” Expert Syst. Appl.,vol.186,p.115736, 2021

[5] T. Lombrozo, Explanation and abductive inference, in: Oxford Handbook of Thinking and Reasoning, 2012, pp. 260–276.

[6] M. Fanan, C. Baron, R. Carli, M.-A. Divernois, J.-C. Marongiu, and G. A. Susto, “Anomaly Detection for Hydroelectric Power Plants: a Machine Learning-based Approach,” in 2023 IEEE 21st International Conference on Industrial Informatics (INDIN), Lemgo, Germany: IEEE, Jul. 2023, pp. 1–6.

[7] Shapley, L. (1953) A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317

[8] Mohammed, Ammar, and Rania Kora. “A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges.” Journal of King Saud University - Computer and Information Sciences 35, no. 2 (February 2023): 757–74.

[9] S.M.Naidu et al, “A Self-Supervised Stacked Ensemble Framework for Robust Anomaly Detection to Reduce False Alarms,” unpublished, in processing. International conference on optimization and learning (OLA2024)

[10] S.M.Naidu et al, “Self-Supervised Learning Framework with Dual Ensemble Voting Fusion for Maximizing Anomaly Prediction in Timeseries,” unpublished, in processing. International Conference on Machine Learning Technologies (ICMLT 2024)

[11] S.M.Naidu et al, “Attention based Convolutional Autoencoder for Risk Assessment,” unpublished, in processing. International Conference on Machine Learning Technologies (ICMLT 2024)

[12] Rao, A.S.; Verweji, G. Sizing the Prize: What’s the Real Value of AI for Your Business and How Can You Capitalise?online: www.pwc.com/gx/en/news-room/docs/report-pwc-ai-analysis-sizing-the-prize.pdf.

[13] Gunning, D.; Vorm, E.; Wang, J.Y.; Turek, M. DARPA’s explainable AI program: A retrospective. Appl. AI Lett. 2021, 2, e61.

[14] Sundar, S.; Rajagopal, M.C.; Zhao, H.; Kuntumalla, G.; Meng, Y.; Chang, H.C.; Shao, C.; Ferreira, P.; Miljkovic, N.; Sinha, S.; et al. Fouling modeling and prediction approach for heat exchangers using deep learning. Int. J. Heat Mass Transf. 2020, 159, 120112

[15] A. Holzinger et al., “Interactive machine learning: experimental evidence for the human in the algorithmic loop,” Appl Intell, vol. 49, no. 7, pp. 2401–2414, Jul. 2019, doi: 10.1007/s10489-018-1361-5.

[16] Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: the role of humans in interactive machine learning. AI Mag 35(4):105–120

[17] K. Roshan and A. Zafar, “Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP),” Int. J. Comput. Netw. Commun.,vol.13,no.6,pp.109−128, 2021.

[18] M. Schultz, N. Gnoss, and M. Tropmann-Frick, XAI in the Audit Domain − Explaining an Autoencoder Model for Anomaly Detection, Nuremberg, Germany, Wirtschaftsinformatik Proceedings. 1., 2022.

[19] Ding, P.; Jia, M.; Wang, H. A dynamic structure-adaptive symbolic approach for slewing bearings’ life prediction under variable working conditions. Struct. Health Monit. 2020, 20, 273–302. [Google Scholar].

[20] S. M. Tripathy, A. Chouhan, M. Dix, A. Kotriwala, B. Klöpper, and A. Prabhune, “Explaining Anomalies in Industrial Multivariate Time-series Data with the help of eXplainable AI,” in 2022 IEEE International Conference on Big Data and Smart Computing (BigComp), Jan. 2022, pp. 226–233.

[21] Y. Yang, H. Lv, and N. Chen, “A Survey on ensemble learning under the era of deep learning,” Artif Intell Rev, vol. 56, no. 6, pp. 5545–5589, Jun. 2023, doi: 10.1007/s10462-022-10283-5.

[22] M. Klaiber, “A fundamental overview of sota-ensemble learning methods for deep learning: a systematic literature review,” Science in Information Technology Letters, vol. 2, no. 2, Art. no. 2, Dec. 2021, doi: 10.31763/sitech.v2i2.549.

[23] S. Maleki, S. Maleki, and N. R. Jennings, “Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering,” Applied Soft Computing, vol. 108, p. 107443, Sep. 2021.

[24] C. Fan, F. Xiao, Y. Zhao, and J. Wang, “Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data,” Applied Energy, vol. 211, pp. 1123–1135, Feb. 2018, doi: 10.1016/j.apenergy.2017.12.005.

[25] Alfeo, A.L.; Cimino, M.G.C.A.; Manco, G.; Ritacco, E.; Vaglini, G. Using an autoencoder in the design of an anomaly detector for smart manufacturing. Pattern Recognit. Lett. 2020, 136, 272–278.

[26] ISO 24028:2020. 2020. Overview of Trustworthiness in Artificial Intelligence.Standard. International Organization for Standardization.