Trustworthy Anomaly Detection in Industrial IoT-Enabled Cyber-Physical Systems Using Explainable Ensemble Learning

Tamara Zhukabayeva[1], Zulfiqar Ahmad[2], Nurdaulet Karabayev[1], Yerik Mardenov[1], Dilaram Baumuratova[1], Hela Elmannai[3]

[1]Faculty of Information Technology, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan

[2]Department of Computer Science and Information Technology, Hazara University, Mansehra 21300, Pakistan

[3]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia

zhukabayeva_tk@enu.kz, zulfiqarahmad@hu.edu.pk, 020419501012@enu.kz, emardenov@gmail.com, d.b_va@mail.ru, hselmannai@pnu.edu.sa

Abstract

Cyber-physical manufacturing systems based on industrial IoT generate large amounts of diverse sensor data. This makes them more vulnerable to operational abnormalities that can lead to equipment failure, safety risks, and production downtime. Intelligent manufacturing systems must therefore have accurate and reliable anomaly detection. This paper suggests a credible system of anomaly detection with the use of explainable ensemble learning on cyber-physical systems of industrial IoT. The given solution combines multiple machine learning models, such as XGBoost, logistic regression, random forest, support vector machine, and multilayer perceptron, into the stacking ensemble architecture to consider the different characteristics of data and minimize the bias of individual models. The probabilistic outputs of the base learners are combined using a meta-model, which leads to better detection accuracy and high strength. Through experiments applied to a real-world smart manufacturing IoT dataset, it can be observed that the proposed ensemble model has an accuracy of 97.98%, which is the highest of all the single classifiers and balanced across normal and anomalous classes. Moreover, the explainable AI with SHAP is implemented to allow transparency into the decision-making process. The findings validate that the proposed framework provides a scalable, reliable, and interpretable solution to real-time anomaly detection in industrial IoT-enabled cyber-physical manufacturing systems.

1 Introduction

In the era of Industry 4.0, the integration of Industrial Internet of Things (IIoT) and Cyber-Physical Systems (CPS) is revolutionizing smart manufacturing by enabling reasl-time monitoring, automation, and optimization of industrial processes (Peter, Pradhan, and Mbohwa 2023; Singh 2021). These production environments are majorly dependent on large systems of interconnected sensors and devices, like temperature, vibration, humidity, pressure, and equipment health, to continually collect and relay information to centralized and edge-based computing platforms and make smart decisions (Shah et al. 2025). The growing complexity and size of these IIoT-enabled CPS systems presents them with high vulnerability, and as such, they become prone to anomalous behaviors that could point to system failures, malfunctioning, or even cyber-attacks (Ahmed et al. 2023; Alabadi, Habbal, and Wei 2022; Ali et al. 2025).

Anomaly detection is very important in such an environment to maintain reliability, safety, and continuous production. The common static threshold-based or rule-based detection systems simply do not work well with the dynamic and multidimensional sensor data and the wide variety of anomaly types that occur in industrial environments. Besides, the diversity of sensors and operation conditions additionally complicates the detection of anomalies. Therefore, there has been a growing trend in adaptive and data-driven machine learning due to its capacity to learn complex trends and identify unobtrusive deviations in real-time (Ali et al. 2020; Chevtchenko et al. 2023; Rafique et al. 2024).

Ensemble learning methods are among machine learning techniques that have been demonstrated to outperform with regard to accuracy and robustness where noisy and imbalanced industrial data are involved. The use of many encapsulated ensemble models is, however, constrained by the need to make them available in safety-critical manufacturing CPS, where user-friendliness and confidence in the model predictions are paramount to their human operators and maintenance engineers (AlHaddad et al. 2023; Mohy-eddine et al. 2022). Explainable Artificial Intelligence (XAI) methods like SHapley Additive explanations (SHAP) are designed to solve this problem by offering both general information about the importance of features and specific details about the process of making a prediction to an anomaly, therefore increasing the level of transparency and practical knowledge

(Alshammeri et al. 2025; Zhang et al. 2025). Out of these challenges, the study will come up with a detailed, explicable ensemble learning model for trustworthy anomaly detection in IIoT-enabled manufacturing CPS. The architecture brings the various base models, such as XGBoost, logistic regression, multilayer perceptron, random forest, and support vector machines, into use and merges their output using meta-learning to generate high precision and strong resilience. Moreover, XAI methods are combined to offer explanations that can be interpreted and help an operator to validate anomaly notifications.

The key contributions of this paper are as follows:

- We develop an explainable ensemble learning framework tailored for anomaly detection in complex IIoT-enabled manufacturing cyber-physical systems.
- We integrate multiple machine learning models to improve detection accuracy and robustness against diverse anomaly types.
- We apply SHAP-based XAI method to provide transparent interpretations of model decisions, supporting trust and operational insight.

The remainder of this paper is organized as follows: Section 2 reviews related works on anomaly detection in IIoT and CPS. Section 3 details the proposed explainable ensemble learning framework. Section 4 presents the experimental setup, results, and discussion. Finally, Section 5 concludes the paper and outlines future research directions.

2 Related Work

We considered the current research related to edge deployment, detection of anomalies using ensemble and deep learning, and the security of Industrial IoT-enabled cyber-physical systems in manufacturing. Advances in IIoT and CPS have triggered a rise in attention to latency-aware and scalable as well as trustful frameworks of anomaly detection to manage the vast and varied data produced by interconnected sensors in the industries (Ahammad 2023). Edge computing has become a vital facilitator of such systems that provide proximity-based data processing to reduce latency and improve real-time responsiveness at the manufacturing site. A cross-functional overview of the edge computing paradigm in industrial environments also classified the state-of-the-art research into application areas and some of the most critical challenges and open issues in implementation, scalability, and security (Zhukabayeva et al. 2025). But although these works focus on the performance of the system, they tend to ignore the incorporation of explainable and reliable anomaly detection systems that are key to the operator trust and the system transparency in the cyber-physical manufacturing systems.

The heterogeneity and complexity in the industrial sensor data are challenges to the conventional methods of detecting anomalies. Recent developments have discussed the ensemble methods of learning that integrate multiple machine learning models to enhance robustness and effectiveness in industrial anomaly detection. As an example, a hybrid anomaly detection approach was introduced in (Jeffrey, Tan, and Villar 2024). This hybrid approach is designed to identify threats to CPSs by combining the signature-based anomaly detection typically utilized in information technology networks, the threshold-based anomaly detection typically utilized in operational technology networks, and behavioural-based anomaly detection using ensemble learning, which leverages the strengths of multiple machine learning algorithms against the same dataset to increase accuracy. Even though these ensemble methods enhance the performance of detection, they are low-level methods and hence cannot be adopted in safety-critical manufacturing.

Explainable Artificial Intelligence has been on the rise as a remedy to the problem of interpretability. The use of techniques like SHapley Additive explanations (SHAP) has been used to give both global and local explanations of the anomaly detection models so that maintenance engineers can learn how some anomalies were detected and which sensors make the greatest contribution to the decision (Lundberg et al. 2019). Research on XAI for IIoT anomaly detection shows that understandable model features enhance trust and decision-making considerably, which is essential in responding to faults in manufacturing CPS in a timely and effective manner (Ahmad et al. 2025).

Recent studies have concentrated on the use of meta-learning and ensemble models to improve anomaly detection in IIoT-enabled CPS. The methods use several base learners that have complementary strengths and combine them with the help of meta-classifiers to enhance the overall detection accuracy and resistance to new anomalies (Zoppi et al. 2021). Not much has incorporated explainability into such ensemble systems to produce CPS. Moreover, it has been shown that hybrid edge-cloud architecture can be effective in real-time anomaly detection in an industrial setting by integrating lightweight edge inference with more profound cloud-based analytics (Sathupadi et al. 2024). These architectures facilitate scalable deployment and faster response times, but they need effective and practical models of anomaly detection. While significant progress has been made in IIoT anomaly detection, ensemble learning, and explainable AI, there remains a pressing need for comprehensive frameworks that integrate these aspects, especially within the complex and safety-critical domain of manufacturing CPS. Our proposed approach contributes to this by offering an explainable ensemble learning framework tailored for trustworthy and interpretable anomaly detection in industrial IoT-enabled cyber-physical manufacturing systems.

3 System Design and Model

This research study presents an explainable ensemble learning framework of anomaly detection of industrial IoT-enabled cyber-physical systems, as shown in Figure 1. The system comprises three basic layers, i.e., data preprocessing, explainable ensemble learning, and cloud-edge deployment. The first layer of the proposed framework is the Data Preprocessing Layer. This layer implies gathering multivariate sensor data of industrial devices (e.g., temperature, vibration, pressure, humidity, and metrics of useful life).

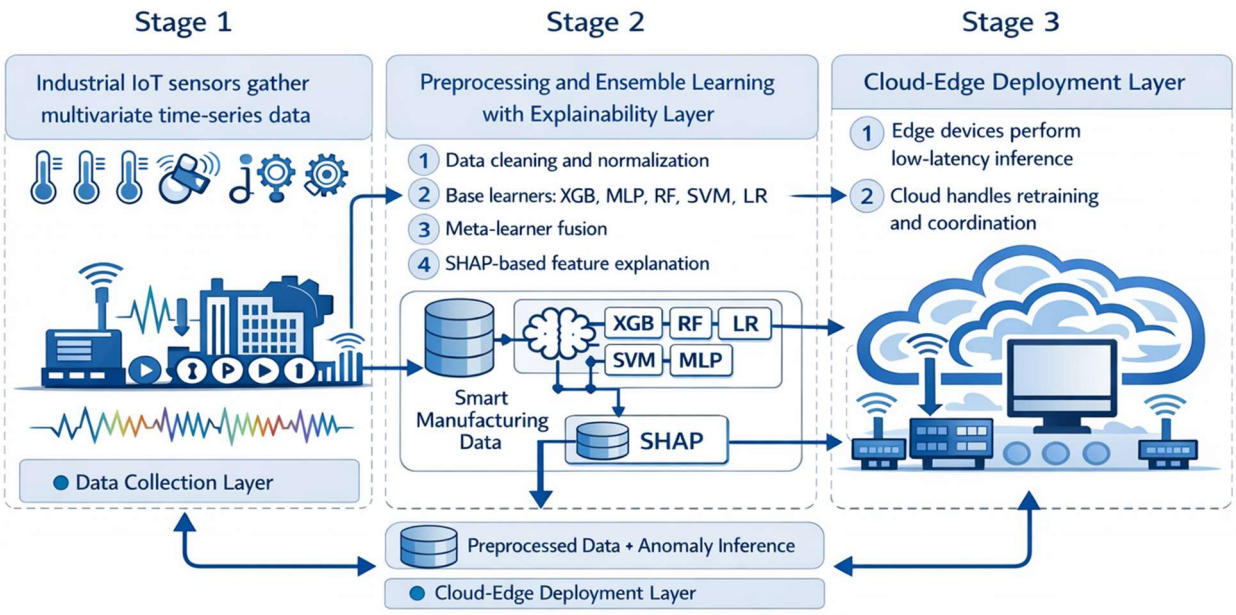

Figure 1: Proposed Explainable Ensemble Learning Framework for Anomaly Detection in IIoT-enabled CPS

The dataset was preprocessed using cleaning, scaling, and encoding techniques. The preprocessing process provides noise reduction and normalization to ensure that the models are fed similar input. The second layer is Ensemble Learning with Explainability Layer, where a number of machine learning models are trained separately, including XGBoost (XGB), logistic regression (LR), multilayer perceptron (MLP), random forest (RF), and support vector machine (SVM). Their outputs are fused through a meta-learning methodology that enhances the accuracy and strength of general prediction.

To overcome the interpretability challenge, SHAP (SHapley Additive explanations) values are calculated that not only give importance to global features but also give explanations of individual anomaly detection. This layer will guarantee transparency of the model and will create operator confidence. The third and last layer is called the Cloud-Edge Deployment Layer. The trained ensemble model is implemented in a hybrid cloud-edge setup in this layer. Edge devices perform real-time, lightweight anomaly inference near data sources to reduce the latency. At the same time, the cloud processes resource-oriented computational activities like retraining the model, mass data analytics, and coordination of system changes. The architecture is scalable, efficient, and of centralized management.

The ensemble learning method used in this study is a stacking method to improve anomaly detection in smart manufacturing IoT data. Let the input dataset be represented by equation 1.

$$D = \{(x_i, y_i)\}_{i=1}^{N} \qquad \text{Eq. 1}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the sensor feature vector and $y_i \in \{0,1\}$ represents normal and anomalous states. Five base classifiers, logistic regression, XGBoost, random forest, support vector machine, and multilayer perceptron, are trained independently on the original feature space. Each base model $f_k(\cdot)$ produces a probability score $p_k = f_k(\mathbf{x})$ indicating the likelihood of an anomaly.

These probability outputs are then combined to form a new meta-feature vector represented by equation 2.

$$\mathbf{z} = [p_1, p_2, \ldots, p_5] \qquad \text{Eq. 2}$$

The meta-model learns an optimal combination of the base model predictions to generate the final decision represented by equation 3.

$$\hat{y} = g(\mathbf{z}) \qquad \text{Eq. 3}$$

The stacking mechanism reduces model bias and limits overfitting by exploiting the complementary strengths of individual classifiers. The quality of the final model is determined by testing a set of data, and the indicators of the performance of this layered ensemble approach, as well as the visualization of the confusion matrix, prove the efficiency of the chosen strategy. The probabilistic outputs of the base learners are used as input features to a logistic regression meta-classifier. Logistic regression learns the optimal weighting of individual model predictions to generate the final anomaly classification. Hyperparameters were selected based on preliminary tuning experiments to achieve optimal model performance. To avoid data leakage during stacking, a k-fold cross-validation strategy (k=5) was used. Base learner models were trained on the training folds and then out-of-fold predictions were produced for the validation fold. These predictions were then used to train the meta-classifier.

Algorithm for the proposed framework

The proposed anomaly detection framework, as outlined in Algorithm 1, enforces an organized pipeline combining sensor data preprocessing, multi-model ensemble learning, explainable AI, and hybrid cloud-edge implementation to avert industrial IoT-enabled cyber-physical systems. The raw multivariate sensor data produced by the various industrial sensors is first cleansed and normalized and then coded to a uniform format that is acceptable to machine learning. Several base models, such as XGBoost, logistic regression, multilayer perceptrons (MLP), random forest, and support vector machines (SVM), are then individually trained on the preprocessed data to determine normal and abnormal behavior. These base learners are used to make individual predictions that are combined together to produce a meta-feature vector, which is used as input to the meta-classifier that then optimizes the combined output by reducing bias and increasing generalization.

To be interpretable, SHapley Additive exPlanations (SHAP) values are calculated to give both global and local explanations of model choices, allowing domain experts to gain knowledge about the impact of every sensor feature on the outcome of anomaly detection. In deployment, the base models can be deployed to low-latency, real-time anomaly inference edge devices near the physical machines constrained by resource availability. The meta-classifier is deployed on the cloud to achieve centralized decision fusion, retraining models, and updates, as well as scalability and flexibility. Such hierarchical structure has latency, resource usage, accuracy, and transparency, which is appropriate in critical anomaly detection in manufacturing CPS settings.

Algorithm 1: Anomaly Detection in Industrial IoT-Enabled Cyber-Physical Systems Using Explainable Ensemble Learning

Begin
Input: Multivariate sensor data samples
$S = \{s_1, s_2, \ldots, s_n\}$
Base models
$M = \{XGBoost, LR, MLP, RF, SVM\}$
Output: Predicted anomaly class $\hat{y}$
Procedure: Cloud–Edge Ensemble Anomaly Detection
Data Preprocessing:
  for each sensor sample $s_i \in S$:
    Clean, normalize, and encode $s_i \rightarrow s_{iprocessed}$
  end for
Training Base Models:
  for each base model $m_k \in M$:
    Train m_k on $\{s_{iprocessed}\} \rightarrow predictions\ \hat{y}_k$
  end for
Collect Base Predictions:
  $P = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5\}$
Meta-Model Construction:
  for each sample s_i:
    Construct meta-feature vector $\Phi_i = \{\hat{y}_{1i}, \hat{y}_{2i}, \hat{y}_{3i}, \hat{y}_{4i}, \hat{y}_{5i}\}$
  end for
  Train meta-classifier M_{meta} on $\Phi \rightarrow final\ prediction\ \hat{y}_i$
Explainability:
  Calculate SHAP values for feature importance
Performance Metrics Calculation:
  $Accuracy = (TP + TN) / (TP + TN + FP + FN)$
  $Precision = TP / (TP + FP)$
  $Recall = TP / (TP + FN)$
  $F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall)$
Cloud-Edge Deployment:
  Deploy base models M on edge devices for real-time inference
  Deploy meta-classifier M_{meta} on cloud for decision fusion, retraining, and updates
Inference:
  For new input s_{new}:
    Preprocess $s_{new} \rightarrow s_{new_{processed}}$
    Generate base predictions $\{\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5\}$
    Create meta-feature vector Φ_{new}
    Compute final prediction $\hat{y} = M_{meta}(\Phi_{new})$
End

4 Experiments, Results and Discussions

We experimentally assessed the proposed explainable ensemble anomaly detection framework on multivariate sensor dataset on industrial IoT-enabled cyber-physical systems. The framework uses five base frameworks that include XGBoost, logistic regression, multilayer perceptron, random forest, and support vector machine to examine independently preprocessed sensor information. These base learners are then combined with a meta-classifier that would enhance the overall classification accuracy and strength. The effectiveness of all the individual base models and the meta-model ensemble was measured according to standard metrics such as accuracy, precision, recall, and F1-score (macro and weighted). The experiment was conducted by dividing the dataset into cross-validation, a training and testing set, to ascertain the strength of the findings.

In order to interpret the model decisions, SHAP values were calculated for better understanding of the importance of features. This explainable factor is an added value to the visibility and credibility of the detection system, which is vital in the industry. A hybrid cloud-edge concept was provided with base models being deployed on edge, while cloud nodes are used to perform meta-classification.

The findings confirmed that the ensemble meta-model has been shown to be significantly better than each of the underlying base learners in all metrics with high accuracy and with a balance between precision-recall scores. The cloud-edge deployment architecture was found to be useful to minimize the detection latency at the cost of scalability.

Dataset

The data used in this study has been obtained on the Kaggle platform with the title of Smart Manufacturing IoT-Cloud Monitoring Dataset (Kaggle 2025). It is multivariate time-series sensor data of many industrial machines and equipment

of a smart manufacturing facility. It has various sensor signals, which include temperature, vibration, humidity, pressure, and predicted remaining useful life (RUL), that record the operating conditions and health status of the machinery. The data includes normal working states and different abnormal states showing defects, wear, or even failure. These aberrations are associated with the various types of faults that can occur in the production processes and can include mechanical degradation, sensor failure, or some unforeseen environmental disruption. The multifaceted character of the dataset makes it possible to design and test machine learning models to identify faults early and predict maintenance. The diversity of sensor types and anomaly classes in the dataset reflects the complexity and heterogeneity of modern industrial IoT-enabled cyber-physical systems.

Data Preprocessing

The raw data that was used in this research was found in the Kaggle Smart Manufacturing IoT Cloud Monitoring dataset, which is multivariate sensor data and operational parameters of industrial machinery. In order to achieve a high quality and consistency of data, the first action was to clean the data via eliminating a row that has anomaly labels (anomaly_flag) missing. This has been done to ensure that all the samples used in training and evaluation were valid in terms of having anomaly status information. The main numeric characteristics that were included were temperature, vibration, humidity, pressure, energy consumption, and estimated remaining useful life (RUL), which were chosen according to the importance of these indicators to monitor machine condition. There was also a categorical variable in the dataset, machine_status, which was one-hot encoded and thus converted into a set of machine learning-consuming numerical indicator variables. Following feature selection as well as encoding, any other rows with missing values in the selected features or target variable were eliminated in an effort to keep the data intact.

The cleaned data was split into 80-20 to create training and testing subsets to be used in training and evaluating the model. To maintain the balance in the model performance, stratification in terms of the anomaly label was implemented during the split so that the ratio of normal samples and anomalous ones is repeated in both groups. The StandardScaler method was used to normalize the numerical features to a mean value of zero and unit variance. The scaling process was necessary to make sure that every feature contributed in proportion to the learning process and assisted in the model convergence and accuracy. The resulting preprocessed data was an effective and clean foundation on which the ensemble learning models can identify anomalies in the industrial IoT environment.

Anomaly Detection

Table 1 shows a comparison of individual base learners, namely, XGBoost (XGB), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), and the proposed ensemble learning meta-model in the task of detecting anomalies in an industrial IoT-enabled cyber-physical system. In the case of Normal class, all the models demonstrate good performance, which means that they are reliable in terms of identifying healthy operating states. RF achieves perfection in terms of precision (1.0000), and SVM and MLP are also characterized by high precision, ranging at 0.9957 and 0.9554, respectively. The ensemble learning model has reached a precision of 0.9799 with a very high recall of 0.9984 and, accordingly, the highest F1-score (0.9890) for normal class compared to any other model. This proves the usefulness of the ensemble in reducing false positives, and this is essential in the manufacturing setting, where the unnecessary alarms can break the production processes.

Class	Precision						Recall						F1 Score						Accuracy					
	XGB	LR	RF	SVM	MLP	Ensemble Learning	XGB	LR	RF	SVM	MLP	Ensemble Learning	XGB	LR	RF	SVM	MLP	Ensemble Learning	XGB	LR	RF	SVM	MLP	Ensemble Learning
Normal	0.9421	0.9384	1.0000	0.9957	0.9554	0.9799	1.0000	0.9998	0.9198	0.9266	0.9958	0.9984	0.9702	0.9681	0.9582	0.9599	0.9752	0.9890	0.9440	0.9401	0.9270	0.9294	0.9538	0.9798
Anomalous	1.0000	0.9949	0.5496	0.5610	0.9241	0.9798	0.3724	0.3292	1.0000	0.9591	0.5255	0.7902	0.5427	0.4947	0.7094	0.7079	0.6700	0.8749						
Macro average	0.9711	0.9666	0.7748	0.7784	0.9398	0.9798	0.6862	0.6645	0.9599	0.9428	0.7606	0.8943	0.7565	0.7314	0.8338	0.8339	0.8226	0.9320						
Weighted average	0.9473	0.9434	0.9598	0.9569	0.9526	0.9798	0.9440	0.9401	0.9270	0.9294	0.9538	0.9798	0.9321	0.9259	0.9360	0.9374	0.9480	0.9789						

Table 1. Performance Comparison of Base Models and Meta-Model

Variation of performance between models is higher in the Anomalous class, which is indicative of the natural difficulty of identifying rare fault conditions in industrial systems. The values of precision of XGB and LR are very high (1.0000 and 0.9949, respectively), but the values of recall are very low (0.3724 and 0.3292), which means that the models missed certain instances of anomalies. RF and SVM, in their turn, have a perfect and nearly perfect recall of 1.0000 and 0.9591 with a lower precision. The ensemble learning model has the optimal trade-off, with a precision of 0.9798 and a recall of 0.7902 and the maximum F1-score of 0.8749 of the anomalous class. This performance balance proves that it is better capable of spotting real anomalies as well as managing false alarms.

These macro-average results also illustrate the strength of the ensemble approach. Although single models like RF and SVM have a macro F1-score of 0.8338 and 0.8339, respectively, the ensemble learning model achieves higher scores in macro-average F1-score of 0.9320 and a macro recall of 0.8943 than the others. This shows steady performance in the case of both normal and anomalous classes, as well as efficient control of the problem of class imbalance. On the same note, the weighted-average measures indicate that even though all the models enjoy the superiority of the dominance of the Normal class, the ensemble model has the highest weighted precision (0.9798), weighted recall (0.9798), and weighted F1-score (0.9789). Besides, the ensemble learning model has the best overall accuracy of 0.9798, which is higher than XGB (0.9440), LR (0.9401), RF (0.9270), SVM (0.9294), and MLP (0.9538). Table 1 clearly shows that individual machine learning models are effective, particularly in particular areas, but the proposed ensemble learning framework has consistent, better, and reliable performance. This renders it very applicable to the real-world industrial IoT and manufacturing cyber-physical systems, where correct detection of anomalies is vital to fault prevention, predictive maintenance, and reliability to operate.

The confusion matrix of the proposed stacking ensemble meta-model for detecting anomalies in industrial IoT-enabled cyber-physical systems is shown in Figure 2. The matrix makes a comprehensive understanding of the classification behavior of the ensemble by depicting the true positives, the true negatives, the false positives, and the false negatives of the normal and the anomalous operating conditions. The ensemble model, as indicated in the figure, correctly predicts that 18,188 normal cases are normal, which indicates that the model has a very high rate of true negatives, and it indicates the high capacity of the model to identify normal and normal behavior of the system. There are actually 29 false positives, which are wrongly identified as being anomalous, meaning that the false positive rate is incredibly low. This is due to the fact that in an industrial manufacturing setting, false alarms may result in unnecessary maintenance measures, downtime, and higher operating expenses. In the anomalous category, the ensemble correctly identifies 1,409 unusual instances (true positives) that indicate the usefulness of the model in detecting abnormal machine behavior that can infer faults or failures. There are 374 misclassified instances as normal, and they are false negatives. Although these false negative instances indicate that there is the possibility of improvement, the collective still attains a good tradeoff between sensitivity and specificity. The confusion matrix demonstrates the ability of the ensemble learning method to make robust decisions whereby the incorporation of a number of base models makes the ensemble models make fewer misclassification mistakes than the single classifiers. The high true positive rate, as well as the low rate of false positives, proves that the suggested framework is appropriate in the context of real-time anomaly detection in industrial IoT and cyber-physical manufacturing systems, where reliability, safety, and trustworthiness are essential.

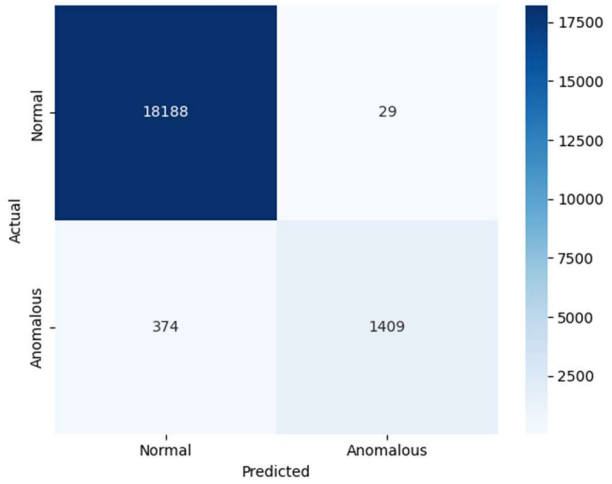

Figure 2: Confusion matrix (ensemble learning)

Explainability using SHAP

In order to enhance the openness and credibility of the suggested ensemble learning model, SHAP (SHapley Additive exPlanations) was used to examine the decision-making pattern of the meta-model. The SHAP summary plot (Figure 3) demonstrates the overall significance of the probability outputs of the base learners, i.e., RF probability, LR probability, XGB probability, SVM probability, and MLP probability, in determining the overall prediction of the final anomaly. RF_prob and LR_prob are the most influential features, as revealed in the SHAP summary plot, with the most extensive range of SHAP values and the biggest effects on the meta-model output. Both are high probability values (shown as red points), and both models play a major role in pushing the prediction towards the anomalous class, and low probability values (blue points) pull the prediction towards the normal class.

The XGB_prob and SVM_prob characteristics are also noteworthy to the decision-making process of the meta-model, nevertheless, with a bit lower significance than RF and LR. They are complementary to each other, as their distribution of SHAP values indicates a positive contribution

at increased values of probability. This is consistent with previous findings on experiments, where XGBoost and SVM showed high recall ability on anomalous samples. Contrary to this, MLP_prob has relatively smaller SHAP magnitudes, which indicates a smaller contribution to the final prediction but does not disappear. Although it might not be very strong individually, the addition of MLP does increase diversity in the ensemble, and this allows the meta-model to generalize better to different operational conditions. The SHAP analysis substantiates that the meta-model is not dependent on only one classifier, but rather it is learning an adaptive weighting scheme, which dynamically favors base models in accordance with their trustworthiness on a particular input. This property of explainability justifies the design decision of ensemble learning and shows that the proposed framework does not only provide great detection but also interpretability, which is a key requirement to operate in an industrial internet of things and smart manufacturing setting.

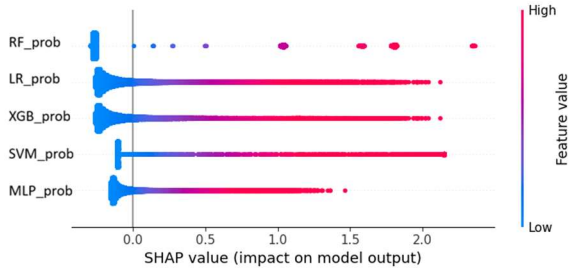

Figure 3: SHAP value (impact on model output)

Discussion on Results

It is evident that the proposed reliable framework of trustworthy anomaly detection is effective and reliable in terms of its use in the context of the Industrial IoT-based cyber-physical manufacturing systems. The proposed method helps to overcome the drawbacks of component models and enhances overall detection results through the combination of heterogeneous base learners, including XGBoost, logistic regression, random forest, support vector machine, and multilayer perceptron, into a stacking ensemble architecture. Based on Table 1, we can see that by using individual base models, we can achieve high precision and recall rates and successfully identify normal operating conditions. Nevertheless, their performance deteriorates with the anomalous class. This imbalance indicates a major predicament in the area of industrial anomaly detection, where anomalous events are few and hard to capture by simply applying stand-alone classifiers. The ensemble meta-model is a major solution towards this problem because it exploits a relationship between complementary decision boundaries that the base models learn. The ensemble itself has an accuracy of 97.98, which is better than all the individual classifiers. It is worth noting that the ensemble achieves the macro F1-score of 93.20, meaning that the performance is balanced between the normal and abnormal classes and that the ensemble is resistant to the imbalance in the classes. The weighted F1-score of 97.89% also indicates that the method has a constant performance throughout the dataset, which proves the suitability of the suggested framework in the practical manufacturing context.

These findings are also confirmed by the confusion matrix in Figure 2. The groupings assign most normal samples properly with a very low false positive error rate, which will not cause much disturbance to the industrial activities. Meanwhile, a significant part of anomalous cases is properly identified, which is the key to early fault detection and predictive maintenance. The fact that there are only a few false negatives means that even though some anomalies can be hard to detect, the given system can greatly minimize the misdetection as compared to single-model options. Besides a high level of predictive performance, the framework focuses on trustworthiness and transparency with explainable AI. Basing the analysis on SHAP, it can be stated that the decisions made by the ensemble are mostly affected by the probability outputs of powerful base learners like XGBoost, logistic regression, and SVM. This validates the assertion that the meta-model is highly effective in focusing on trusted predictions and having repressed less informative signals. This interpretability is essential to the industrial stakeholders because it will allow the engineers to know the reason behind the generation of certain anomaly alerts and therefore enhance trust in automated decision-making systems.

5 Conclusion

In this paper, we provided a credible framework of anomaly detection in industrial IoT-enabled cyber-physical manufacturing systems using explainable ensemble learning. The suggested framework combines several machine learning models in a stacking ensemble model to properly detect abnormal operation patterns using heterogeneous IoT sensor data. The strengths of various base learners and using a meta-model to make a decision effectively overcome the limitations of imbalance of classes, model bias, and limited generalization that are usually witnessed with standalone classifiers. Larger-scale experimental analysis of a real-world smart manufacturing dataset has proven that the proposed ensemble model has much better accuracy, precision, recall, and F1-score than the single classifiers. The ensemble has a high macro F1-score as well as an overall accuracy of 97.98, which indicates its robustness both in normal and anomalous operating conditions. The reliability of the framework was further supported by the confusion matrix analysis indicating a very low rate of false alarms and inefficient detection of the anomalous events that is very important in minimizing production disruptions and operational safety in the industrial surroundings. The framework was integrated with Explainable AI (XAI) to understand the decisions of the meta-model with the help of SHAP to achieve the process of transparency and trust. According to the explainability findings, one can see that the ensemble intelligently focuses on the predictions of stronger base models, which gives both global and local information about the decisions made on picking up anomalies. Such interpretability allows engineers and operators of the system

to comprehend automated alerts more effectively and confirm their usefulness, which will lead to more trust in the implementation of AI-based monitoring systems.

In future work, we aim to address multiclass faults and attack classification in the proposed framework to make a distinction among various types of anomalies in industrial cyber-physical systems. The integration of online and incremental learning mechanisms will enable the model to adjust to the concept of drift and changing manufacturing operational patterns in real-time manufacturing environment. The implementation of the framework on edge devices with resource constraints and its validation in large-scale and real-world industrial applications will also prove its scalability.

Acknowledgments

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No AP23489127).

References

Ahammad, I. 2023. Fog Computing Complete Review: Concepts, Trends, Architectures, Technologies, Simulators, Security Issues, Applications, and Open Research Fields. SN Computer Science 4(6):765. doi.org/10.1007/s42979-023-02235-9.

Ahmad, J.; Latif, S.; Khan, I. U.; Alshehri, M. S.; Khan, M. S.; Alasbali, N.; and Jiang, W. 2025. An Interpretable Deep Learning Framework for Intrusion Detection in Industrial Internet of Things. Internet of Things 33:101681. doi.org/10.1016/j.iot.2025.101681.

Ahmed, S. F.; Alam, M. S. B.; Hoque, M.; Lameesa, A.; Afrin, S.; Farah, T.; Kabir, M.; Shafiullah, G. M.; and Muyeen, S. M. 2023. Industrial Internet of Things Enabled Technologies, Challenges, and Future Directions. Computers and Electrical Engineering 110:108847. doi.org/10.1016/j.compeleceng.2023.108847.

Alabadi, M.; Habbal, A.; and Wei, X. 2022. Industrial Internet of Things: Requirements, Architecture, Challenges, and Future Research Directions. IEEE Access 10:66374–66400. doi.org/10.1109/ACCESS.2022.3185049.

AlHaddad, U.; Basuhail, A.; Khemakhem, M.; Eassa, F. E.; and Jambi, K. 2023. Ensemble Model Based on Hybrid Deep Learning for Intrusion Detection in Smart Grid Networks. Sensors 23(17):7464. doi.org/10.3390/s23177464.

Ali, T.; Khan, Y.; Ali, T.; Faizullah, S.; Alghamdi, T.; and Anwar, S. 2020. An Automated Permission Selection Framework for Android Platform. Journal of Grid Computing 18(3):547–561. doi.org/10.1007/s10723-018-9455-1.

Ali, W.; Amin, M.; Alarfaj, F. K.; Al-Otaibi, Y. D.; and Anwar, S. 2025. AI-Enhanced Differential Privacy Architecture for Securing Consumer Internet of Things (CIoT) Data. IEEE Transactions on Consumer Electronics 71(2):5201–5215. doi.org/10.1109/TCE.2025.3573519.

Alshammeri, M.; Ahmad, Z.; Humayun, M.; and Alamri, M. 2025. Explainable Cluster-Based Predictive Framework for Early Diagnosis of Autism Spectrum Disorder Using Behavioral Biomarkers. Diagnostics 15(24):3241. doi.org/10.3390/diagnostics15243241.

Chevtchenko, S. F.; Rocha, E. D. S.; Dos Santos, M. C. M.; Mota, R. L.; Vieira, D. M.; De Andrade, E. C.; and De Araújo, D. R. B. 2023. Anomaly Detection in Industrial Machinery Using IoT Devices and Machine Learning: A Systematic Mapping. IEEE Access 11:128288–128305. doi.org/10.1109/ACCESS.2023.3333242.

Jeffrey, N.; Tan, Q.; and Villar, J. R. 2024. Using Ensemble Learning for Anomaly Detection in Cyber–Physical Systems. Electronics 13(7):1391. doi.org/10.3390/electronics13071391.

Kaggle. 2025. Smart Manufacturing IoT-Cloud Monitoring Dataset. https://www.kaggle.com/datasets/ziya07/smart-manufacturing-iot-cloud-monitoring-dataset.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2019. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv preprint. http://arxiv.org/abs/1905.04610.

Mohy-Eddine, M.; Guezzaz, A.; Benkirane, S.; and Azrour, M. 2022. An Effective Intrusion Detection Approach Based on Ensemble Learning for IIoT Edge Computing. Journal of Computer Virology and Hacking Techniques 19(4):469–481. doi.org/10.1007/s11416-022-00456-9.

Peter, O.; Pradhan, A.; and Mbohwa, C. 2023. Industrial Internet of Things (IIoT): Opportunities, Challenges, and Requirements in Manufacturing Businesses in Emerging Economies. Procedia Computer Science 217:856–865. doi.org/10.1016/j.procs.2022.12.282.

Rafique, S. H.; Abdallah, A.; Musa, N. S.; and Murugan, T. 2024. Machine Learning and Deep Learning Techniques for Internet of Things Network Anomaly Detection—Current Research Trends. Sensors 24(6):1968. doi.org/10.3390/s24061968.

Sathupadi, K.; Achar, S.; Bhaskaran, S. V.; Faruqui, N.; Abdullah-Al-Wadud, M.; and Uddin, J. 2024. Edge-Cloud Synergy for AI-Enhanced Sensor Network Data: A Real-Time Predictive Maintenance Framework. Sensors 24(24):7918. doi.org/10.3390/s24247918.

Shah, B.; Junaid, M.; Rustam, H.; Habib, M.; and Anwar, S. 2025. AI-Driven Fog-Edge Computing for IoMT Systems: Architecture and Use Cases. Proceedings of the AAAI Symposium Series 6(1):42–48. doi.org/10.1609/aaaiss.v6i1.36023.

Singh, H. 2021. Big Data, Industry 4.0 and Cyber-Physical Systems Integration: A Smart Industry Context. Materials Today: Proceedings 46:157–162. doi.org/10.1016/j.matpr.2020.07.170.

Zhang, K.; Zheng, B.; Xue, J.; and Zhou, Y. 2025. Explainable and Trust-Aware AI-Driven Network Slicing Framework for 6G IoT Using Deep Learning. IEEE Internet of Things Journal 1–1. doi.org/10.1109/JIOT.2025.3619970.

Zhukabayeva, T.; Ahmad, Z.; Karabayev, N.; Baumuratova, D.; and Ali, M. 2025. An Intrusion Detection System for Multiclass Classification Across Multiple Datasets in Industrial IoT Using Machine Learning and Neural Networks Integrated with Edge Computing. In *Data, Information and Computing Science*, 98–110. IOS Press.

Zoppi, T.; Gharib, M.; Atif, M.; and Bondavalli, A. 2021. Meta-Learning to Improve Unsupervised Intrusion Detection in Cyber-Physical Systems. ACM Transactions on Cyber-Physical Systems 5(4):1–27. doi.org/10.1145/3467470.