Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 59-10 (2025) 1498–1503

Explaining Manufacturing Anomalies: Transformer-Based Detection with xAI for Imbalanced Process Data

Abdullah Al Noman * **Anton Zitnikov** **
Firoj Ahmmed Patwary *** **Aaron Heuermann** *
Klaus-Dieter Thoben *

* *BIBA-Bremer Institut für Produktion und Logistik GmbH, 28359 Bremen, Germany (e-mail: nom@biba.uni-bremen.de, her@biba.uni-bremen.de, tho@biba.uni-bremen.de).*
** *Faculty of Production Engineering, University of Bremen, 28359 Bremen, Germany (e-mail: zitnikov@uni-bremen.de)*
*** *Freie Universität Berlin, 14195 Berlin, Germany (e-mail: firoj.patwary@fu-berlin.de)*

Abstract:
The manufacturing industry is increasingly adopting a computational approach that relies heavily on process data for operational insight. Anomaly detection plays a crucial role in providing a thorough understanding of process behavior, helping operators determine if their production systems are operating optimally or if proactive intervention is required. A significant challenge with machine learning-based solutions is their lack of interpretability, making it difficult to understand the reasoning behind model predictions. This paper addresses the need for interpretability in anomaly detection using Transformer networks, achieving 82% accuracy in the experiments conducted for this study, where the minority class required half the data augmentation applied to the majority class for balance. The explainable AI framework known as Local Interpretable Model-agnostic Explanations (LIME) is used to elucidate the critical features and their interactions that influence individual predictions. Traditionally used to interpret sequential neural network models, this study extends the application of LIME to Transformer models for anomaly detection in manufacturing processes. This method not only enables anomaly detection but also helps to identify key features that signal anomalies, thereby improving process management and control.

Keywords: Anomaly Detection, Transformer Model, Explainable AI, LIME, Feature Extraction, Imbalanced Data, Process Data, Manufacturing Data, Model Interpretability.

1. INTRODUCTION

In the era of Industry 4.0, the digital transformation of manufacturing is significantly increasing the volume and complexity of production data (Hughes et al., 2022). This evolution necessitates the development of advanced analytical tools to effectively manage and derive valuable insights from such data, particularly in identifying anomalies that are essential for maintaining product quality, operational efficiency, and minimizing downtime and costs. Anomaly detection identifies rare, unusual patterns in data and is a key field in machine learning (Shon and Moon, 2007; Salima Omar, 2013). Traditional anomaly detection systems, often designed for balanced datasets or based on specific assumptions about data distribution, struggle with the highly imbalanced nature of manufacturing data, where anomalies are rare (Hajjami et al., 2020). This challenge is further compounded by a lack of interpretability, which is crucial for providing actionable insights into the causes of anomalies. This paper introduces an approach to enhance anomaly detection in manufacturing environments where data are highly imbalanced by utilizing Transformer-based architectures. To address the challenge of data imbalance, the synthetic minority over-sampling technique (SMOTE) is utilized to generate additional minority samples within training sets, helping to correct the imbalance and improve model accuracy. Beyond model performance, model interpretability is essential for decision-making, particularly in the manufacturing field, where strict inspection requirements for product quality and safety must be met (Hayes and Shah, 2017; Puggini and McLoone, 2018). To improve model interpretability, this approach integrates Local Interpretable Model-agnostic Explanations (LIME). LIME provides insight into the model's decision-making processes by identifying the features most significantly influencing anomaly detection. This integration enables manufacturers to not only detect anomalies but also understand the underlying causes, facilitating proactive and informed decision-making in fast-paced production environments. This paper's primary contribution is the integration of Transformer-based architectures with the SMOTE to ad-

10.1016/j.ifacol.2025.09.252

dress the challenges of imbalanced manufacturing data in anomaly detection. It investigates the optimal percentage of over-sampling that enhances model performance without overfitting. Furthermore, the inclusion of LIME significantly improves the interpretability of the detection process. Together, these advancements allow for the effective identification and comprehensive understanding of anomalies, providing manufacturers with actionable insights critical for optimizing production processes.

The rest of this paper is structured as follows: Section 2 reviews related work in anomaly detection. Section 3 describes the methodology, including the overall workflow and details of the implementation. Section 4 presents the analysis, results, and discussion. Finally, Section 5 draws the main conclusions.

2. LITERATURE REVIEW

Anomaly detection is critical in manufacturing, where it involves identifying deviations from normal process behavior. Traditional approaches such as statistical process control (SPC) and rule-based systems have been widely used but face significant limitations when dealing with high-dimensional datasets or variables with nonlinear relationships. Machine learning (ML) models, including support vector machines (Yokkampon et al., 2021), decision trees (Douiba et al., 2023), and autoencoders, have been adopted to address these challenges. Autoencoders, for example, excel at compressing process data to identify anomalies, but they often struggle with imbalanced datasets, leading to missed anomalies and high false-negative rates (González and Dasgupta, 2003). Transformers have emerged as a powerful alternative for handling sequential and time-series data due to their ability to model long-term dependencies and contextual information (Vaswani, 2017). Studies have demonstrated the superiority of Transformers over traditional models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks in detecting anomalies in time-series data (Reza et al., 2022). Furthermore, their ability to process both high-dimensional and sequential data makes them well-suited for anomaly detection in manufacturing environments (Ogunfowora and Najjaran, 2023). Imbalanced datasets are a pervasive issue in manufacturing anomaly detection. Oversampling, undersampling, and synthetic data generation methods such as the Synthetic Minority Oversampling Technique (SMOTE) have been commonly employed to address this imbalance (Chawla et al., 2002). While effective in some cases, these techniques can lead to overfitting or fail to capture the complexity of anomalies in rapidly changing environments. Transformers, on the other hand, leverage their attention mechanisms to naturally focus on minority class features, making them suitable for highly imbalanced datasets without explicit data balancing (Ahmed et al., 2023). Despite the accuracy of ML models, their ”black box” nature limits their adoption in critical manufacturing contexts, where interpretability is essential (Hassija et al., 2024). xAI addresses this limitation by providing insights into model predictions. LIME can explain model decisions by analyzing the impact of different input features on predictions (Ribeiro et al., 2016). LIME has been successfully applied to enhance interpretability in anomaly detection models, including random forests and neural networks, in industrial contexts (Alvarez Melis and Jaakkola, 2018). Combining Transformers with LIME offers a dual advantage: accurate anomaly detection and interpretable decision-making.

3. METHODOLOGY

The data used in this study were obtained from a Fine Edge Manufacturing Process, which involves the utilization of advanced and sophisticated machinery. These machines are equipped with innovative technologies designed to maintain high production efficiency while seamlessly integrating various components to produce the final product. This study employs a Transformer model to detect anomalies in production processes, with the goals of identifying key contributors to defects, improving defect detection accuracy, and reducing error rates. The dataset used for the study consisted of 34883 samples, divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The training set was utilized to develop the Transformer model, while the validation set was employed to evaluate its performance and select the model yielding optimal results. The test set was then used for final evaluation to determine the model's accuracy and robustness. To gain insights into the decision-making process of the Transformer model, this study incorporates xAI tools, specifically LIME. The workflow of the study encompasses data preprocessing, model training, validation, testing, and explainability analysis, as illustrated in Figure 1.

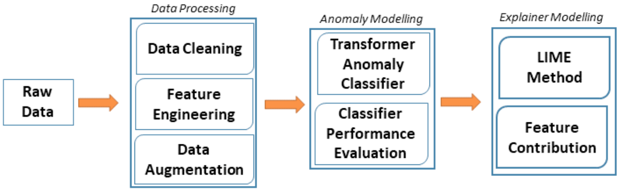

Fig. 1. Overview of the anomaly detection and explainability workflow

3.1 Data Processing

The initial phase of this study focused on data preprocessing to prepare raw sequential data for analysis and modeling. The distribution and uniqueness of each feature were carefully examined to identify and resolve issues such as missing values, redundancy, and inconsistencies. These data cleaning steps were critical for ensuring data quality and improving model accuracy. During feature engineering, statistical metrics such as mean, maximum, minimum, and standard deviation were calculated for each raw feature, resulting in the construction of 20 derived features to be used as model inputs. With the dataset including labels categorizing products as normal or defective, the anomaly detection task was framed as a binary classification problem to predict defective products. The dataset comprised 33,402 normal entries and 1,481 defective entries, reflecting a significant imbalance with a ratio of approximately 22:1. To address this imbalance, the data augmentation technique SMOTE was applied. SMOTE generates synthetic samples by interpolating between existing data points in the minority class, providing a more robust and effective solution than simple duplication methods. To ensure consistency in feature scales and mitigate potential biases during model training, all feature values were standardized using the z-score method. This normalization process improved the model's learning efficiency and predictive performance.

3.2 Anomaly Model Description

The Transformer is a foundational encoder-decoder architecture designed for sequence-to-sequence tasks, such as machine translation (Vaswani, 2017). It consists of two main components: an encoder, which processes the input (source) sequence, and a decoder, which generates the output (target) sequence. The encoder is composed of a stack of identical layers, each with two sublayers: a multi-head self-attention mechanism and a positionwise feed-forward network (FFN). The self-attention mechanism allows each position in the input sequence to attend to all other positions, capturing global dependencies. Queries, keys, and values for self-attention are derived from the outputs of the previous encoder layer. To ensure stable gradients and efficient training, each sublayer is wrapped in a residual connection followed by layer normalization. The encoder ultimately outputs a fixed-dimensional vector representation for each position in the input sequence. The decoder is also a stack of identical layers, but includes an additional sublayer for encoder-decoder attention. Each layer consists of three sublayers: masked multi-head self-attention, encoder-decoder attention, and a positionwise feed-forward network. The masked attention mechanism ensures that each position in the decoder can only attend to earlier positions or itself, maintaining the autoregressive property needed for sequential output generation. The encoder-decoder attention sublayer allows the decoder to focus on relevant parts of the encoded input sequence by using queries from the decoder's self-attention output and keys and values from the encoder's output. Like the encoder, each sublayer in the decoder is surrounded by residual connections and layer normalization. The mask attention in the decoder ensures predictions depend only on previously generated tokens, preserving the autoregressive property. The Transformer's parallelization capability, enabled by the self-attention mechanism, allows it to process sequences efficiently, making it computationally faster than traditional RNNs.

3.3 Explainer Model Description

Model-agnostic explanation methods treat machine learning models as black-box functions, relying solely on their outputs to provide explanations. LIME is a flexible algorithm designed to explain the predictions of any classifier or regressor in a faithful and interpretable manner (Ribeiro et al., 2016). The core idea behind LIME is to approximate the behavior of a complex model locally around a specific input by using a simpler, interpretable model. These interpretable models, such as decision trees or linear regression, are trained on small perturbations of the original data, such as adding noise, removing words, or obscuring parts of an image. By analyzing how these perturbations impact the model's predictions, LIME constructs an explanation for the input's prediction. The process begins by modifying the input data sample and generating new samples in its neighborhood with tweaked feature values. These perturbed samples are evaluated using the target model, and the resulting outputs are used to train the surrogate model. This surrogate model provides a good local approximation of the target model's behavior near the original input. The output of LIME is a set of explanations, often in the form of feature contributions, that represent how each feature influenced the prediction for the specific sample. This allows users to gain local interpretability by understanding which features were most significant in the model's decision-making for that instance.

3.4 Experimental Setup

This study evaluated Transformer models through five experiments, with data augmentation. Data augmentation was achieved by upsampling the minority class by 20% - 60%. The overall experimental pipeline is depicted in Figure 1. The models were trained for up to 50 epochs, using early stopping to halt training if validation performance showed no improvement for five consecutive epochs. Training was performed in batches of 64. Augmented datasets were used to evaluate the impact of augmentation on model generalization, supported by learning curves to assess potential overfitting or improved robustness. Geometric augmentations, as detailed in the methodology, were designed to simulate real-world variations, enhancing the diversity and robustness of the dataset. The models were implemented using the TensorFlow framework. Training loss was calculated using Binary Cross-Entropy, and the Adam optimizer, with a learning rate of 0.01, was employed to ensure efficient convergence. The performance of the model was evaluated on the test set using standard classification metrics: accuracy, precision, recall, and F1 score. To further enhance the analysis, LIME explanations highlighted important features influencing the model's decisions for individual instances, offering transparency into the reasoning behind predictions. These insights supported the diagnosis of the causes of the anomaly, complementing quantitative metrics with actionable information for improving product quality and defect detection strategies.

4. RESULTS AND DISCUSSION

The experiments conducted in this study addressed the significant class imbalance in the dataset, with a ratio of 22:1 between the majority (Normal) and minority (Anomaly) classes. Such a severe imbalance challenges the Transformer model's ability to effectively learn and predict minority class instances, as the model tends to bias toward the majority class to optimize overall accuracy. To mitigate this issue, oversampling through SMOTE was applied to ensure adequate representation of the minority class during training. This strategy enabled a thorough evaluation of the Transformer model's performance at different augmentation levels (20%, 30%, 40%, 50%, and 60%) by increasing the minority class data to these proportions of the majority class size using the SMOTE algorithm, facilitating the identification of the optimal configuration for anomaly detection.

Table 1. Performance Metrics for Transformer Models with Different Minority Class (Anomaly) Augmentation Levels

Model	Acc.	ROC	Labels	Precision	Recall	F1
20%	0.87	0.85	Normal	0.0.87	1.00	0.93
			Anomaly	0.92	0.24	0.38
30%	0.84	0.86	Normal	0.84	0.99	0.91
			Anomaly	0.91	0.34	0.50
40%	0.82	0.86	Normal	0.84	0.92	0.88
			Anomaly	0.74	0.55	0.63
50%	0.82	0.88	Normal	0.83	0.92	0.87
			Anomaly	0.79	0.63	0.70
60%	0.81	0.88	Normal	0.84	0.86	0.85
			Anomaly	0.76	0.72	0.74

4.1 Results on Transformer Modelling

Table 1 presents the classification metrics obtained for the Transformer models trained on datasets with different augmentation levels. The model's performance was evaluated across five augmentation levels, revealing key trends in sensitivity, specificity, and generalization. At 20% augmentation, the model achieved 87% accuracy, with a high precision of 92% but a low recall of 24% for the minority class, reflecting difficulty in detecting anomalies. Increasing augmentation to 30% improved recall while maintaining high precision, although accuracy dropped slightly. At 40%, the model achieved a better balance with 82% accuracy, a substantial recall improvement to 55%, and a trade-off in precision, indicating enhanced sensitivity at the cost of more false positives. The 50% augmentation level maintained 82% accuracy, with further recall improvement to 63% and precision stabilizing at 79%, providing the best balance between sensitivity and specificity. At 60%, recall peaked at 72%, but precision dropped to 76%, and fluctuations in the loss curves indicated overfitting due to excessive noise in the data.

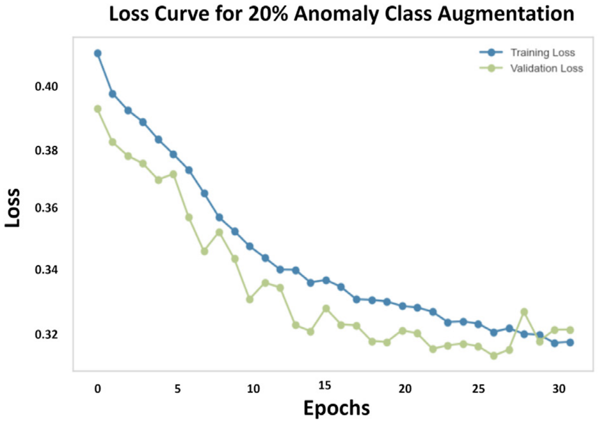

Fig. 2. Training and validation loss curves for 20% of anomaly class data augmentation.

In figure 2 and 3, the loss curves for 20% and 50% augmentation showed smooth convergence, suggesting effective generalization, while in figure 4 shows the augmentation of 60% introduced erratic behavior, reflecting reduced robustness. In figure 5, the Receiver Operating Characteristic (ROC) analysis showed Area Under the Curve (AUC) values ranging from 0.85 to 0.88, with diminishing benefits at higher augmentation levels due to overfitting.

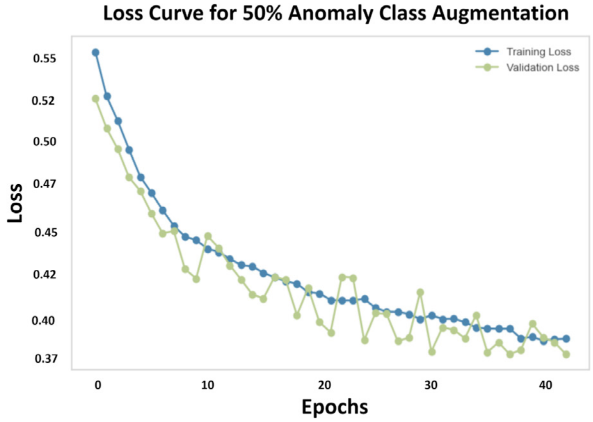

Fig. 3. Training and validation loss curves for 50% of anomaly class data augmentation.

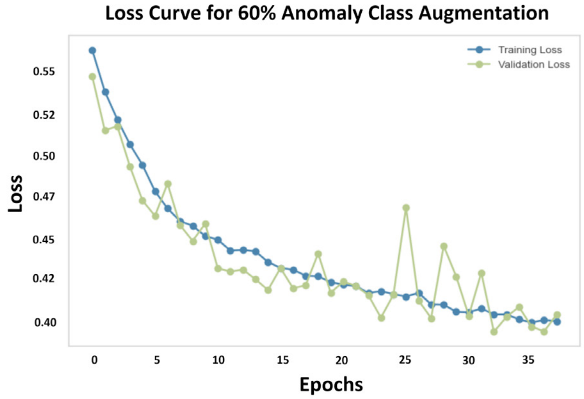

Fig. 4. Training and validation loss curves for 60% of anomaly class data augmentation.

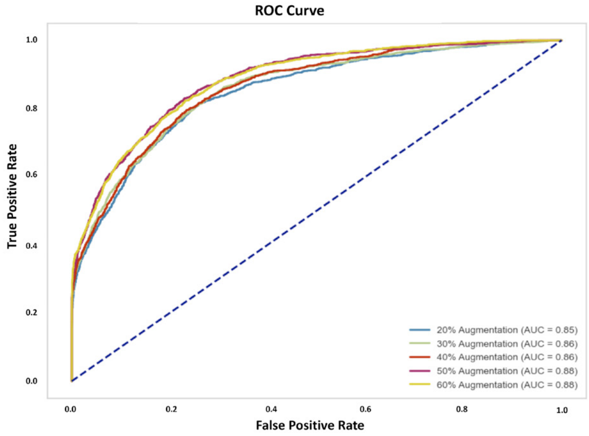

Fig. 5. ROC curve for different percentages of data augmentation for model performance evaluation.

4.2 Explanability using LIME xAI Technique

The application of transformers in manufacturing processes introduces significant advantages in terms of predictive power but also presents challenges related to interpretability.

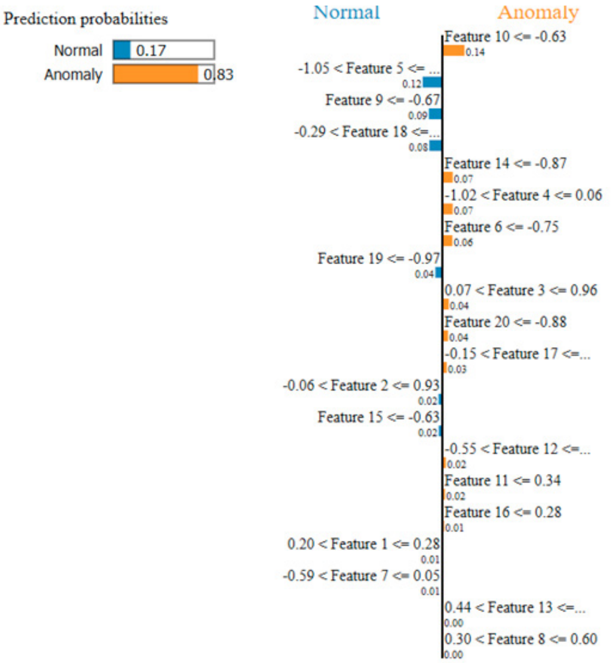

Fig. 6. Prediction probability and contribution analysis on anomaly 1

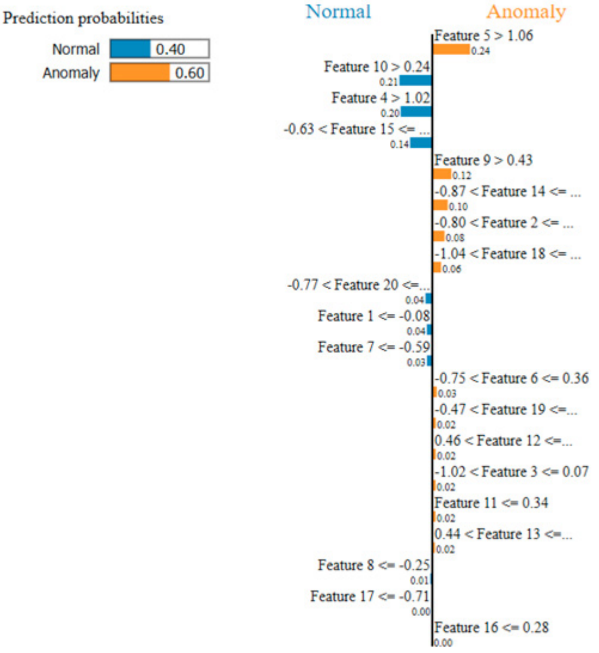

Fig. 7. Prediction probability and contribution analysis on anomaly 2

Figure 6 illustrates the LIME analysis for the first anomaly instance, demonstrating high confidence in the anomaly prediction with a probability score of 0.83. This score indicates a strong deviation of certain features from normal operational ranges. The primary contributors to this anomaly prediction were Features 10, 14, and 4, with Feature 10 showing the highest positive contribution, suggesting that

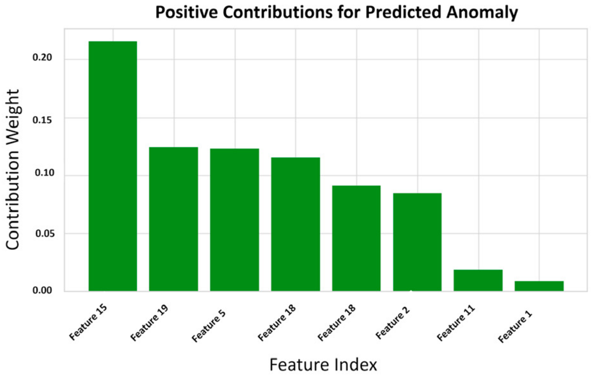

Fig. 8. Feature contribution analysis on anomaly 3

its significant deviation from normal values is a strong indicator of potential anomalies. Figure 7 presents the second instance, where the LIME output shows a moderate anomaly probability score of 0.60. This reflects a potential onset of anomalous conditions, where the features have not yet fully deviated beyond normal thresholds but are indicative of emerging issues. Key contributing features included Features 5, 9, and 14, with Feature 5 identified as having the most substantial impact on this prediction. This emphasizes its role in identifying disruptive operational conditions that may require closer monitoring. For the third instance, which results in an absolute anomaly prediction with a probability score of 1.00, it is suggested that definitive anomalous conditions are present. Figure 8 shows that Features 15, 19 and 5 were the most critical in this determination, with Feature 15 providing the largest positive contribution. This underscores the significance of these features in signaling clear and severe deviations from typical manufacturing processes, necessitating immediate and targeted corrective actions.

4.3 Discussion

The results from the experiments conducted in this study provide valuable insights into the challenges and opportunities associated with the use of Transformers for anomaly detection in manufacturing processes. These insights emphasize the intricate balance required in data preprocessing, specifically in handling class imbalance through data augmentation, and the critical role of model interpretability in operational settings. The augmentation helped address the severe class imbalance, which is often detrimental to model performance, particularly for minority classes that represent critical anomalies. The optimal augmentation level of 50% provided the best trade-off between detecting true positives and avoiding false positives, which is crucial in a manufacturing context where both precision and recall have significant operational implications. The diminishing returns observed at higher levels of augmentation, particularly at 60%, highlight the risks associated with overfitting. This suggests that while augmentation is beneficial, excessive use can lead to models that are too closely fitted to the training data, reducing their effectiveness in generalizing to new data. The integration of LIME for interpretability addresses the opaque nature of models. Understanding the ”why” behind predictions enables more informed decision-making and facilitates trust between the users and the automated systems. For instance, the identification of Features 4, 5, 9, 10, 14, 15, and 19 in various instances as key contributors to anomaly predictions provides actionable insights. These features, associated with specific operational parameters, can be monitored more closely for deviations that precede anomalies.

5. CONCLUSION

This study demonstrates the potential of Transformer-based models combined with LIME for anomaly detection in manufacturing processes. The proposed approach effectively addresses challenges such as class imbalance, interpretability, and the complexity of industrial data, achieving a balance between detection performance and actionable insights. By integrating LIME, this framework enhances predictive capabilities while providing manufacturers with a deeper understanding of the causes of anomalies, enabling informed decision-making and improving operational reliability. However, this study has certain limitations. While SMOTE effectively addressed class imbalance, the synthetic nature of the augmented data may not fully reflect real-world anomalies, potentially affecting generalizability. The computational cost of transformers poses challenges for resource-constrained environments, and LIME, though effective, has limitations in capturing feature interactions comprehensively. Additionally, the framework has yet to be validated in real-world manufacturing settings, where variability in operational conditions may affect its robustness. Future work will focus on exploring computationally efficient transformer variants and developing more accurate local self-explainable AI (sAI) techniques to enhance interpretability. Finally, testing the framework in live manufacturing scenarios will validate its scalability, robustness, and practical impact, further bridging the gap between AI-driven analytics and industrial application.

ACKNOWLEDGEMENTS

The activities described in this manuscript were conducted as part of the AI Automated Industrial Analytics submitted to and funded by EIT Manufacturing within the second Call for Innovation activities for the Business Plan 2023–2025 (project ID 24094).

REFERENCES

Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A., and Gandomi, A.H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11), 13521–13617.

Alvarez Melis, D. and Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.

Douiba, M., Benkirane, S., Guezzaz, A., and Azrour, M. (2023). An improved anomaly detection model for IoT security using decision tree and gradient boosting. *The Journal of Supercomputing*, 79(3), 3392–3411. doi: 10.1007/s11227-022-04783-y.

González, F.A. and Dasgupta, D. (2003). Anomaly detection using real-valued negative selection. *Genetic Programming and Evolvable Machines*, 4, 383–403.

Hajjami, S.E., Malki, J., Berrada, M., and Fourka, B. (2020). Machine Learning for anomaly detection. Performance study considering anomaly distribution in an imbalanced dataset. In *2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, 1–8. doi: 10.1109/CloudTech49835.2020.9365887.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1), 45–74.

Hayes, B. and Shah, J.A. (2017). Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 6586–6593. IEEE Press, Singapore, Singapore. doi:10.1109/ICRA.2017.7989778.

Hughes, L., Dwivedi, Y.K., Rana, N.P., Williams, M.D., and Raghavan, V. (2022). Perspectives on the future of manufacturing within the industry 4.0 era. *Production Planning & Control*, 33(2-3), 138–158.

Ogunfowora, O. and Najjaran, H. (2023). A transformer-based framework for multi-variate time series: A remaining useful life prediction use case. *arXiv preprint arXiv:2308.09884*.

Puggini, L. and McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. *Engineering Applications of Artificial Intelligence*, 67, 126–135. doi: 10.1016/j.engappai.2017.09.021.

Reza, S., Ferreira, M.C., Machado, J.J., and Tavares, J.M.R. (2022). A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications*, 202, 117275.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). " 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*, 1135–1144.

Salima Omar, Asri Ngadi, H.H.J. (2013). Machine learning techniques for anomaly detection: An overview. *International Journal of Computer Applications*, 79(2), 33–41. doi:10.5120/13715-1478.

Shon, T. and Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. *Inf. Sci.*, 177(18), 3799–3821. doi:10.1016/j.ins.2007.03.025.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Yokkampon, U., Chumkamon, S., Mowshowitz, A., Fujisawa, R., and Hayashi, E. (2021). Anomaly Detection Using Support Vector Machines for Time Series Data:. *Journal of Robotics, Networking and Artificial Life*, 8(1), 41. doi:10.2991/jrnal.k.210521.010.