CrowdSense: Interpretable and Efficient Multivariate Crowd Forecasting with Active Learning

Anahid Wachsenegger*[1], Anita Graser[1], Axel Weißenfeld[1] and Melitta Dragaschnig[1]

[1] *AIT Austrian Institute of Technology, Vienna, Austria*

Abstract

Accurate forecasting of multivariate time series is essential for high-stakes industrial applications, where real-time decisions rely not only on predictive accuracy but also on transparency and human oversight. In this work, we present a novel Explainable Active Learning (XAL) framework for multivariate time series forecasting that integrates human expertise into the learning loop while enhancing interpretability. Our approach is specifically designed for complex and dynamic environments, such as crowd density prediction in urban settings, where high-impact decisions depend on anticipating critical events. We combine classical and deep learning models—including XGBoost, Temporal Convolutional Networks, Temporal Fusion Transformers, and TimeGPT—within an active learning loop that selects the most informative data points for expert review. Using SHAP-based explanations, our framework provides actionable insights into model behavior, allowing domain experts to iteratively refine predictions through guided feedback. Applied to real-world crowd density data over an 11-day horizon, our method demonstrates superior performance: XGBoost augmented with XAL achieves an R^2 of 0.8491 and the lowest RMSE of 0.3126, while increasing recall for high-density events by 27%. By bringing humans into the loop and ensuring explainability in multivariate forecasting, this work addresses key challenges in industrial domains, where understanding why a model makes a prediction is as important as the prediction itself. The proposed XAL framework offers a promising direction for deploying trustworthy AI in environments where safety, efficiency, and accountability are paramount.

Keywords

multivariate timeseries, interpretability, explainability, crowd density forecasting, machine learning,

1. Introduction

Forecasting rare events in multivariate time series data is a significant challenge across various domains, especially when these events impact operational decisions. In industrial settings, predicting rare events like machinery failures or safety risks requires models that are both accurate and interpretable [1]. We propose an explainable active learning framework to improve the forecasting and understanding of such events. This approach is highly relevant for industrial applications, where explaining predictions can enhance decision-making. We demonstrate its effectiveness using a crowd density dataset, highlighting its potential for addressing complex, rare event forecasting challenges.

While deep learning models like LSTMs [2], GRUs [3], and Transformers [4] have advanced multivariate time series forecasting, their black-box nature limits their practical use in safety-critical domains. These models often fail to provide interpretable explanations for their outputs, hindering error diagnosis and human oversight. Although explanation tools like SHAP [5] and LIME [6] offer model-agnostic insights, their adaptation to time-dependent, multivariate inputs remains limited.

Recent innovations, including LT2D [7], GCFs [8], and Informer-based models [9], have improved forecasting over long horizons and spatial grids. Yet, the trade-off between accuracy and explainability is rarely addressed systematically in these studies. Interpretability-focused architectures like the Temporal

1st Workshop on Leveraging SEmaNtics for Transparency in Industrial Systems (SENTIS), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

*Corresponding author.

anahid.wachsenegger@ait.ac.at (A. Wachsenegger*); anita.graser@ait.ac.at (A. Graser); axel.weissenfeld@ait.ac.at (A. Weißenfeld); melitta.dragaschnig@ait.ac.at (M. Dragaschnig)

0000-0002-8889-0735 (A. Wachsenegger*); 0000-0001-5361-2885 (A. Graser); 0000-0002-7246-2744 (A. Weißenfeld); 0000-0001-5100-2717 (M. Dragaschnig)

Fusion Transformer (TFT) [4] offer some progress, but robust frameworks that integrate both human feedback and model introspection are still underdeveloped.

In this work, we propose a hybrid approach that combines high-performing forecasting models (such as XGBoost, TCN, TFT, and TimeGPT) with a human-centered interpretability workflow. Central to our method is an *Explainable Active Learning* (XAL) loop that enables domain experts to interact with SHAP-based dashboards, diagnose prediction failures, and apply targeted corrections. This iterative refinement leads to significant gains in forecasting rare high-risk events, while enhancing model transparency and usability. We evaluate models using both standard accuracy metrics and a dual-layer interpretability framework: (1) SHAP value analysis to trace temporal feature contributions, and (2) cluster-based surrogate decision trees to understand prediction regimes. Our results show that XGBoost, when paired with XAL, offers the best balance of precision, interpretability, and operational robustness, particularly in forecasting extreme crowding scenarios.

Our key contributions are as follows:

- We introduce an explainable active learning (XAL) workflow that integrates SHAP-based visual diagnostics with expert-in-the-loop feedback to refine crowd density forecasts.
- We demonstrate how XGBoost, when augmented with XAL, achieves strong forecasting accuracy on multivariate urban crowd data, particularly improving recall for critical high-risk events.
- We provide a dual-layer interpretability framework combining temporal SHAP attributions with cluster-based surrogate models to diagnose and explain forecasting behavior at both global and local scales.
- We develop an interactive dashboard to support expert corrections and guide model retraining, showing measurable performance gains across several evaluation rounds.
- We contribute visual analyses (including risk heatmaps and confusion matrices before and after XAL) that illustrate the practical impact of explainability in safety-critical forecasting scenarios.

The rest of the paper is organized as follows. Section 2 reviews related work in time series forecasting and explainable AI. Section 3 describes our modeling pipeline, interpretability tools, and XAL framework. Section 4 presents experimental results and expert-in-the-loop evaluations. Section 5 concludes with limitations and directions for scalable, real-time crowd forecasting systems.

2. State of the Art

Crowd density forecasting has become essential for urban planning and public safety, with recent advancements shifting from traditional statistical models to machine learning and deep learning approaches. This evolution emphasizes spatiotemporal modeling, multimodal data integration, and improved model interpretability and generalizability. This section outlines key advances in multivariate time series forecasting for crowd prediction, along with current explainability and active learning techniques.

2.1. Multivariate Time Series for Crowd Forecasting

Traditional statistical methods such as ARIMA and GARCH have historically been used for short-term crowd forecasting tasks [10, 11]. However, their reliance on assumptions of stationarity and linearity limits their ability to model the nonlinear, high-variance patterns characteristic of real-world urban environments. To overcome these limitations, the field has increasingly shifted toward deep learning techniques, including Long Short-Term Memory networks (LSTM) [2], Gated Recurrent Units (GRU) [3], and Bidirectional LSTMs (BiLSTM) [12]. Extensions such as ConvLSTM [13] and LT2D [7] further improve long-range forecasting by leveraging multi-resolution temporal inputs and spatial structure.

To better capture the spatiotemporal dynamics of crowd behavior, recent models have incorporated mobile phone signaling data and adopted convolutional or attention-based mechanisms for spatially irregular urban regions [14]. Graph Neural Networks (GNNs), such as the Graph-based Crowd Forecaster

(GCF), have extended this capability by modeling crowd dynamics at micro, meso, and macro scales [8]. Transformer-based models like Informer have also been adapted for urban forecasting tasks, with applications such as MobCovid integrating exogenous variables (e.g., COVID-19 case rates and mobility policies) to enhance accuracy [9].

Complementary to these, fuzzy cognitive maps (FCMs) have gained attention as an interpretable tool for capturing causal relationships between variables, especially in video-based crowd monitoring systems [15].

Despite the robustness of these approaches, many suffer from high computational demands, limited scalability, or lack of interpretability – key barriers for real-time decision-making in operational environments.

In this work, we address these gaps by systematically benchmarking models that balance predictive performance with computational efficiency and explainability. Our evaluation spans traditional interpretable models like XGBoost, deep learning architectures such as Temporal Convolutional Networks (TCNs) [16] and Temporal Fusion Transformers (TFTs) [4], and emerging foundation models like TimeGPT [17, 18]. We specifically assess these models' suitability for deployment in crowd forecasting scenarios, intending to bridge the gap between academic modeling innovations and the practical demands of safety-critical, high-density urban settings.

2.2. Explainable AI for Time Series Forecasting

Applying Explainable AI (XAI) to multivariate time series (MTS) forecasting presents unique challenges due to the inherent temporal dependencies, high dimensionality, and complex inter-feature interactions of time series data. Deep neural network (DNN) architectures, such as Long Short-Term Memory (LSTM) networks and Transformers, are widely used in this domain for their ability to model intricate temporal and contextual relationships across multiple variables. These models often outperform traditional statistical approaches such as AR and ARIMA in domains such as traffic forecasting, financial modeling, and weather prediction. However, their black-box nature hinders transparency and interpretability, particularly in critical decision-making settings.

In time series forecasting, interpretability is not only about understanding what the model predicts, but also when specific inputs influence predictions and why. This understanding is essential in high-stakes applications, such as public safety, infrastructure management, and healthcare, where accountability, trust, and error diagnosis are paramount.

Despite increasing interest, interpretability in MTS forecasting remains a relatively underexplored area. Much of the existing work focuses on post hoc local explanation techniques, which provide instance-specific insights and can be integrated with existing forecasting pipelines with minimal architectural changes [19, 20, 21]. Perturbation-based methods are among the most widely used local explanation techniques [22]. These methods estimate the importance of input features by altering them—typically by replacing values with noise or statistical aggregates—and measuring the resulting impact on model predictions. While intuitive, these approaches face difficulties in time series contexts, where "removing" or perturbing timestamped inputs can distort the underlying temporal structure, leading to unrealistic or misleading interpretations.

Attribution-based methods offer a complementary approach by directly quantifying each input's contribution to the model's output [5]. Gradient-based attribution techniques, including SHAP, have shown promise in time series classification, yet their application to time series forecasting is still limited and often lacks domain-specific adaptations [20]. Recent developments in SHAP-based approaches are expanding, as highlighted by several key contributions to the field. For example, the FI-SHAP approach, which improves feature engineering for time series forecasting by enhancing SHAP explanations, thus addressing some of the limitations in feature selection for forecasting [23]. Another approach is C-SHAP, which is a method specifically designed to offer high-level temporal explanations for time series forecasting using Prophet decomposition and SHAP values [24]. In a non-forecasting application, an unsupervised feature selection approach using SHAP for industrial time series anomaly detection was presented to showcase its application to real-world industrial datasets [25]. These advancements demonstrate the increasing relevance and potential of SHAP for both explaining and improving time series forecasting.

Among recent advancements, the Temporal Fusion Transformer (TFT) architecture has emerged as a promising interpretable solution for MTS forecasting [26, 27]. TFT integrates variable selection and temporal attention mechanisms to provide built-in interpretability while maintaining high forecasting accuracy. Its ability to capture long-range dependencies and highlight important temporal patterns makes it particularly suitable for operational deployment in time-sensitive, high-risk environments.

2.3. Active Learning

Recent research on active learning (AL) for time series data primarily focuses on classification and anomaly detection tasks, driven by the need to efficiently label large volumes of unlabeled sequential data where manual annotation is costly and time-consuming. This emphasis is particularly evident in domains such as industrial monitoring, healthcare, and cybersecurity, where timely detection of rare or abnormal events is critical to prevent failures and losses.

For example, RLAD [28] introduces a semi-supervised anomaly detection algorithm combining deep reinforcement learning with active learning to continuously adapt to new anomaly patterns without assumptions on data generation, achieving significant improvements over state-of-the-art unsupervised and semi-supervised methods with minimal labeled data. Similarly, a white-box anomaly detector using moving averages and prediction intervals optimized via active learning and Bayesian methods offers interpretable results for univariate time series anomaly detection in IT infrastructure monitoring [29]. In healthcare, the ActDP framework [30] leverages a combination of data programming and active learning for ECG beat classification, iteratively refining labels through expert feedback and boosting classification accuracy substantially on large datasets. Industrial applications are addressed through active learning frameworks that incorporate pre-clustering and advanced feature extraction to overcome the cold start problem and reduce labeling efforts, achieving over 90% accuracy by labeling only 10% of data in vibration and process control time series [31].

Reviews of deep learning approaches highlight the challenges of anomaly detection in multivariate time series due to the need to model temporal dependencies and variable interactions, while stressing the importance of domain knowledge and expert input facilitated by active learning [32]. Additionally, novel active learning methods that include class balancing strategies help mitigate bias in imbalanced time series datasets, demonstrating effectiveness in texture recognition and industrial fault detection tasks by significantly reducing labeled data requirements [33].

Despite these advances, active learning for multivariate time series forecasting remains under-explored, with most existing work focusing on classification or anomaly detection. This gap suggests opportunities for further research to develop AL strategies that address the unique challenges of forecasting in complex, high-dimensional time series data.

2.4. Synergy Between Explainable AI and Active Learning

We propose that integrating Explainable AI (XAI) with Active Learning (AL) offers a powerful, interactive framework for improving multivariate time series forecasting. XAI builds model transparency and trust, while AL optimizes data labeling by focusing on uncertain or informative samples. Despite AL's success in other domains, its application to time series forecasting is still limited, especially in complex, high-stakes contexts like urban crowd prediction. Our work addresses this gap by introducing an explainable AL (XAL) framework designed specifically for multivariate forecasting in urban settings:

- **Limited spatial generalization:** Current approaches often apply explanations uniformly across regions, overlooking the localized and context-specific drivers of crowd behavior. *Our framework supports region-sensitive refinement through expert-in-the-loop feedback that captures spatial variation.*
- **No support for feedback incorporation:** Human-in-the-loop forecasting remains largely unexplored in spatiotemporal settings, with little to no mechanisms for incorporating expert

corrections or suggestions into the learning cycle. *XAL directly integrates expert feedback (such as re-weighting features, correcting anomalies, and contextual insights) into both model updates and explanation adjustments.*
- **Disconnect between XAI and AL:** In active learning workflows, query selection is rarely guided by interpretability metrics, resulting in suboptimal sampling and inefficient learning in data-scarce or high-risk regions. *Our approach bridges this gap by using explanation uncertainty and domain relevance to inform the active sampling process.*

To address these challenges, we propose **XAL** – an explainable active learning framework that tightly integrates human interactions into the multivariate time series forecasting pipeline.

3. Methodology

This section outlines the experimental workflow used to forecast hourly crowd density and to iteratively improve model performance through our XAL loop. Our methodology integrates historical data, contextual features, and expert feedback to enhance both predictive accuracy and model transparency.

3.1. Problem Formulation

Given a multivariate time series dataset $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, where each $\mathbf{x}_t \in \mathbb{R}^d$ represents d features (e.g., crowd density, weather, mobility indices) observed at time t, the task is to predict future crowd density values $\hat{\mathbf{y}}_{t+h}$ over a forecast horizon h. Formally, we learn a forecasting function

$$f : \{\mathbf{x}_{t-w+1}, \dots, \mathbf{x}_t\} \rightarrow \hat{\mathbf{y}}_{t+1:t+h},$$

where w is the size of the historical input window, and $\hat{\mathbf{y}}_{t+1:t+h}$ denotes predicted crowd densities for the next h time steps.

3.2. Dataset

Our dataset is a multimodal time series compiled from diverse sources in the Scheveningen region of the Netherlands, spanning from 1 May 2022 to 31 October 2024. The primary data source consists of hourly crowd density estimates extrapolated from a voluntary mobile application used by regional visitors. These serve as our ground truth for regional crowd levels.

In addition to crowd data, we incorporate high-frequency parking occupancy records collected every 15 minutes across three major parking facilities in the city. These facilities exhibit a “waterfall” usage pattern: as Parking A approaches full capacity, drivers overflow to Parking B, and subsequently to Parking C. Notably, Parking C, with the highest capacity (approximately 1,700 spaces), serves as a strong proxy indicator for extreme visitor density. To align with other data sources, we up-sample this data to an hourly frequency.

To account for external influences on visitor behavior, we integrate hourly meteorological data—including temperature, wind speed, cloud coverage, and precipitation probability—as well as a structured event calendar that flags public holidays and cultural events known to drive surges in attendance.

Forecasting crowd density under these conditions is challenging. As shown in Figure 1, high-density events are rare, creating an imbalanced target distribution that turns peak periods into anomaly-like cases. Specifically, 1.47% of the training set (from 2022-04-01 to 2024-04-01) contains high-risk events, while 3.46% of the test set (from 2024-04-01 to 2024-10-01) includes high-risk events. Accurate forecasting, therefore, requires sensitivity to contextual cues that may precede such outliers. Additionally, missing values and irregular sampling are present in some data sources. We address this using K-Nearest Neighbors (KNN) imputation, with the number of neighbors tuned to preserve temporal structure while avoiding data leakage.

Combined with the inherent variability in environmental and behavioral features (Figure 2), these factors contribute to the complexity and noise inherent in short-term crowd forecasting.

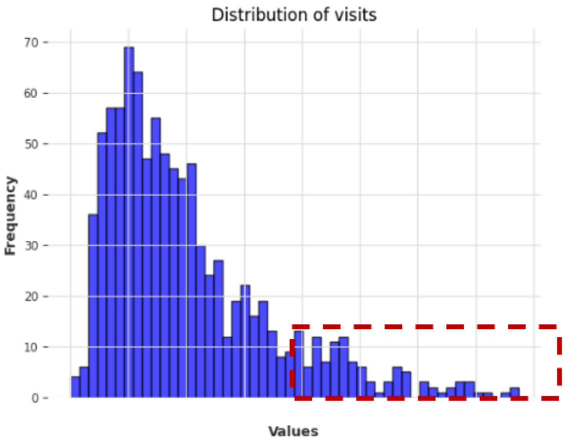

Figure 1: Distribution of crowd density showing the rarity of high-density events, which poses challenges in predicting anomalies.

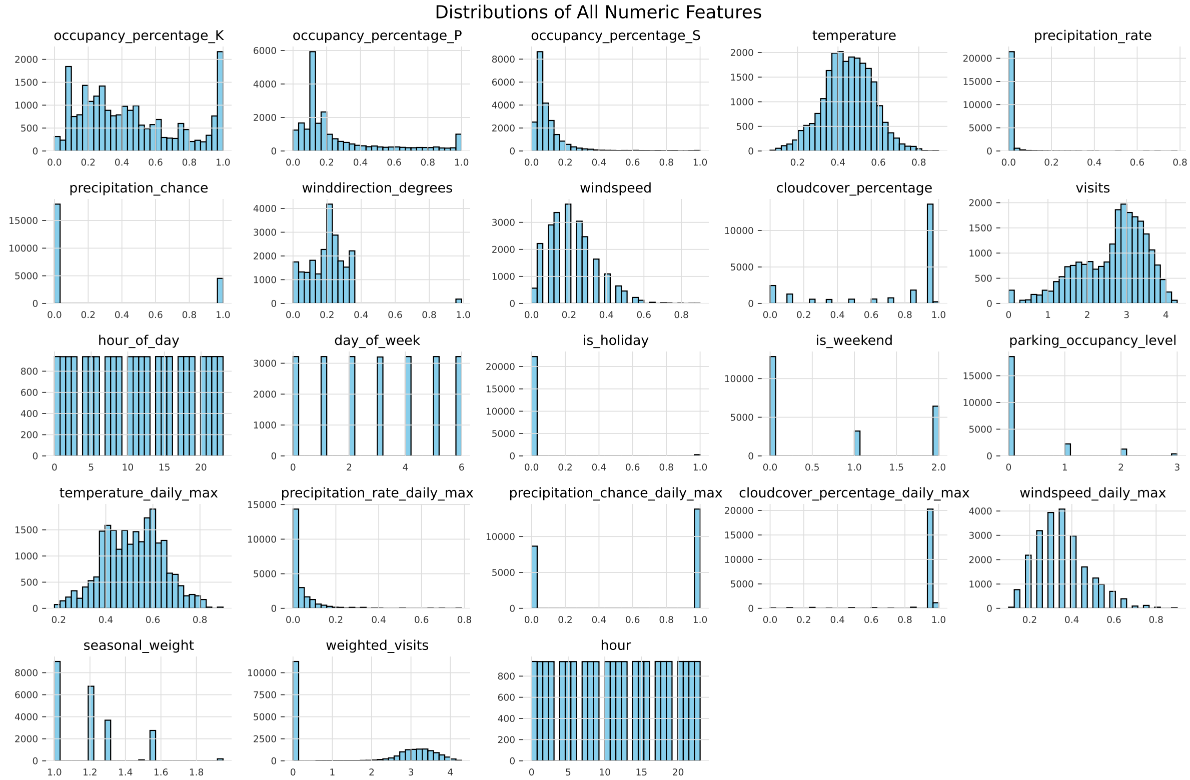

Figure 2: Distribution of key dataset features after preprocessing and normalization. (X-axis values removed for data protection reasons.)

All data sources are temporally aligned and resampled to a consistent hourly resolution, i.e., the 15-minute frequency of the parking data is resampled to an hourly max value, forming a unified multivariate time series that serves as input for the models evaluated in this study.

3.3. Preprocessing and Feature Engineering

To prepare the dataset for modeling, we standardize the time index by converting all timestamps to UTC and ensuring consistent datetime formatting. Meteorological and occupancy features are renamed and normalized to a [0,1] scale using domain-relevant min-max ranges to facilitate model convergence and comparability. The target variable, visitor count, is log-transformed to reduce skewness and stabilize variance.

We engineer several derived features to incorporate domain knowledge and temporal effects. A categorical parking occupancy level is computed to reflect the “waterfall” pattern across the three parking lots, indicating parking occupancy severity on a scale from 0 to 3. Daily maximum weather statistics are aggregated to capture extreme environmental conditions impacting visitor behavior. Additionally, a seasonal weighting factor is introduced, increasing during holidays, weekends, and summer months to reflect expected crowd density fluctuations. Finally, a time-of-day weight emphasizes forecast accuracy during peak hours (08:00–20:00) by scaling visits accordingly.

3.4. Forecasting Models

We benchmarked multiple forecasting approaches, including:

- **Baseline**: A naïve repeat of the previous week’s hourly values.
- **XGBoost**: A tree-based gradient boosting model optimized for tabular data.
- **Temporal Convolutional Network (TCN)**: A deep learning model that captures long-range dependencies using dilated convolutions.
- **Temporal Fusion Transformer (TFT)**: A neural architecture that uses attention mechanisms for multi-horizon forecasting.
- **TimeGPT**: A foundation model fine-tuned for time series generation.

All models (except TimeGPT, which was used in a zero-shot manner) were trained using the same feature set to ensure a fair comparison. Model hyperparameters were optimized via grid search or framework-specific tuning procedures, depending on the architecture. Evaluation was conducted on the 11-day test horizon using standard metrics such as RMSE, MAE, and R^2.

3.5. Explainable Active Learning (XAL) Loop

To iteratively refine the model and improve both forecast accuracy and interpretability, we developed a human-in-the-loop Explainable Active Learning (XAL) workflow. The process begins with an initial model trained on the full dataset. We then generate explanations using SHAP (SHapley Additive exPlanations), a model-agnostic attribution method that quantifies the contribution of each input feature to the forecast. Specifically, we adapt SHAP to multivariate time series by computing feature- and time-wise attributions, highlighting temporal patterns and key drivers behind forecast outcomes [5].

These attributions are visualized in an interactive dashboard using the *Plotly* library and *Panel* 4, which allows domain experts to explore temporal trends, identify discrepancies, and interpret model behavior across regions and time. Based on these visual insights, experts can apply structured corrections. These include reweighting temporal or contextual features, correcting noisy inputs such as erroneous parking data, or manually labeling atypical high-density events. The corrected inputs are used to augment or revise the training data, after which the model is retrained. Each iteration concludes with performance monitoring, where metrics related to forecast accuracy and risk detection are evaluated before and after updates. This closed-loop refinement forms the core of our XAL pipeline, enabling dynamic model improvement and user-aligned interpretability.

3.6. Implementation Details

Our experiments were implemented in Python using several key libraries. We employed the `darts` framework for time series forecasting, which provides implementations of models such as XGBoost, Temporal Fusion Transformer (TFT), and Temporal Convolutional Networks (TCN). For interpretability, we used the shap library to compute SHAP values and analyze feature contributions over time. To support the interactive human-in-the-loop workflow, we developed custom dashboards using `plotly` for visualization and `panel` for layout and control components. Together, these tools enable efficient exploration, correction, and retraining within our XAL framework.

Table 1
Performance comparison of forecasting models (11-day horizon)

Model	R2	RMSE
Baseline	0.3843	0.7614
XGBoost	0.7028	0.3671
XGBoost (after XAL)	***0.8491***	***0.3126***
TCN	0.6941	0.3721
TFT	0.5973	0.4431
TimeGPT	0.7568	0.3477

4. Results

This section presents the performance of multiple time series forecasting models and evaluates the impact of our explainable active learning (XAL) approach for iterative model refinement. The goal is to forecast hourly crowd density for the next 11 days (264 time steps) using a combination of historical features, engineered context-aware covariates, and expert-guided corrections. The expert-in-the-loop feedback mechanism plays a critical role in refining the model throughout its development. As shown in Figure 4, the user interface (UI) of the dashboard allows domain experts and ML developers to perform various corrections and adjustments to the forecasting model. The corrections possible in this interface include, but are not limited to, adjusting seasonal weighting factors for holidays and weekends, correcting outlier patterns in e.g., parking occupancy data, and refining logic for time-of-day importance. Therefore, the domain and ML experts can adjust weights for specific columns to highlight the importance of certain features, ensuring that the model prioritizes the most relevant data. In cases where certain periods have inaccuracies—such as imputed data periods that deviate from expected ranges—experts can modify the values by replacing them with median values derived from corresponding days of the week from previous and future years, ensuring consistency and accuracy.

Regarding the model parameterization, we applied hyperparameter tuning via Optuna. The final XGB model parameters were set as follows: random_state=7, gamma=0.3, booster='gbtree', eta=0.01, max_depth=10, n_estimators=100. For the lags, we chose 24 lags for past covariates, specifically from [-24, -23, ..., -1]. The future covariates included lags from [-24, -23, ..., -1, 0, 1, ..., 24].

For all other models in the study, we used the default parameter values provided by their respective libraries. This choice was aligned with our data-centric approach, where we prioritized focusing on feature engineering and debugging rather than dedicating significant resources to fine-tuning the model parameters. By doing so, we aimed to optimize our resources to address the inherent challenges in the data and improve its quality.

4.1. Forecasting Performance

We evaluated five forecasting models (naïve Baseline, XGBoost, Temporal Fusion Transformer (TFT), Temporal Convolutional Network (TCN), and TimeGPT) using standard error metrics (MAE, MSE, RMSE, MAPE, and R^2. Table 1 shows the performance comparison of different forecasting models over an 11-day horizon of Summer 2024.

The baseline model achieved the lowest performance, with an R^2 of 0.3843 and an RMSE of 0.7614, indicating limited predictive capability. Among the machine learning models, XGBoost significantly improved the results, achieving an R^2 of 0.7028 and an RMSE of 0.3671. Notably, applying XAL to XGBoost further enhanced performance, yielding the highest R^2 of 0.8491 and the lowest RMSE of 0.3126, demonstrating superior forecasting accuracy.

Deep learning models also showed competitive performance. The TCN obtained an R^2 of 0.6941 and an RMSE of 0.3721, slightly below the standard XGBoost model but outperforming the TFT, which recorded an R^2 of 0.5973 and an RMSE of 0.4431. The TimeGPT model achieved robust results with an R^2 of 0.7568 and an RMSE of 0.3477, outperforming both TCN and TFT but not surpassing XGBoost after XAL.

Importantly, the superior performance of XGBoost, particularly when enhanced with active learning, can be attributed to its efficient handling of structured data and ability to capture complex, non-linear relationships without requiring extensive training time. In contrast, the deep learning models evaluated, while state-of-the-art, generally require significantly longer training periods and computational resources. Our results indicate that XGBoost with active learning provides an effective and computationally efficient solution for crowd density forecasting, outperforming both traditional baselines and more computationally intensive deep learning approaches.

4.2. Impact of XAL on Forecast Accuracy and Interpretability

To address the inherent complexity and unpredictability of short-term crowd density forecasting, we applied the XAL loop – an iterative workflow that combines model interpretability with targeted data and feature refinement. The XAL framework was specifically developed to tackle challenges in dynamic urban environments, where data variability, rare crowd surges, and contextual dependencies limit the effectiveness of static black-box models.

As visualized in Figure 3, initial forecasts from our XGBoost model exhibit notable discrepancies, particularly during high-density periods such as summer weekends and public holidays. These errors are most pronounced when models fail to capture nonlinear interactions or misweight key temporal drivers like seasonal patterns or parking saturation thresholds.

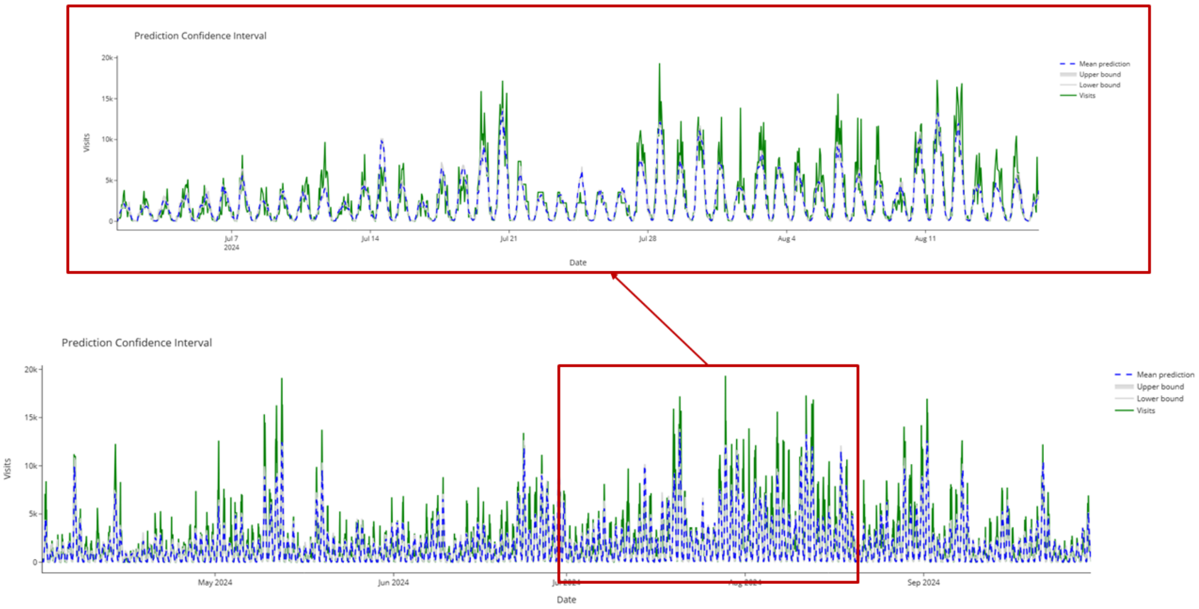

Figure 3: XGBoost forecast vs. true visitor count for Summer 2024 before applying the XAL workflow. Note significant underestimation during peak periods. (Y-axis values removed for data protection reasons.)

The XAL loop introduces a human-in-the-loop feedback mechanism supported by an interactive SHAP-based dashboard. This tool enables users to examine model predictions (top panel), the evolution of input features (middle panel), and SHAP-derived explanations over time (bottom panel), as shown in Figure 4. By clearly distinguishing between observed and future covariates, the dashboard provides actionable insight into which features influenced predictions, and when.

Through this interface, domain experts identified problematic regions in the training data and applied targeted corrections, such as: adjusting seasonal weighting factors for holidays and weekends, correcting outlier patterns in parking occupancy data, and refining logic for time-of-day importance.

The application of the XAL feedback and retraining cycle led to consistent and measurable improvements in forecasting performance. As shown in Table 2, the initial XGBoost model achieves an R^2 of

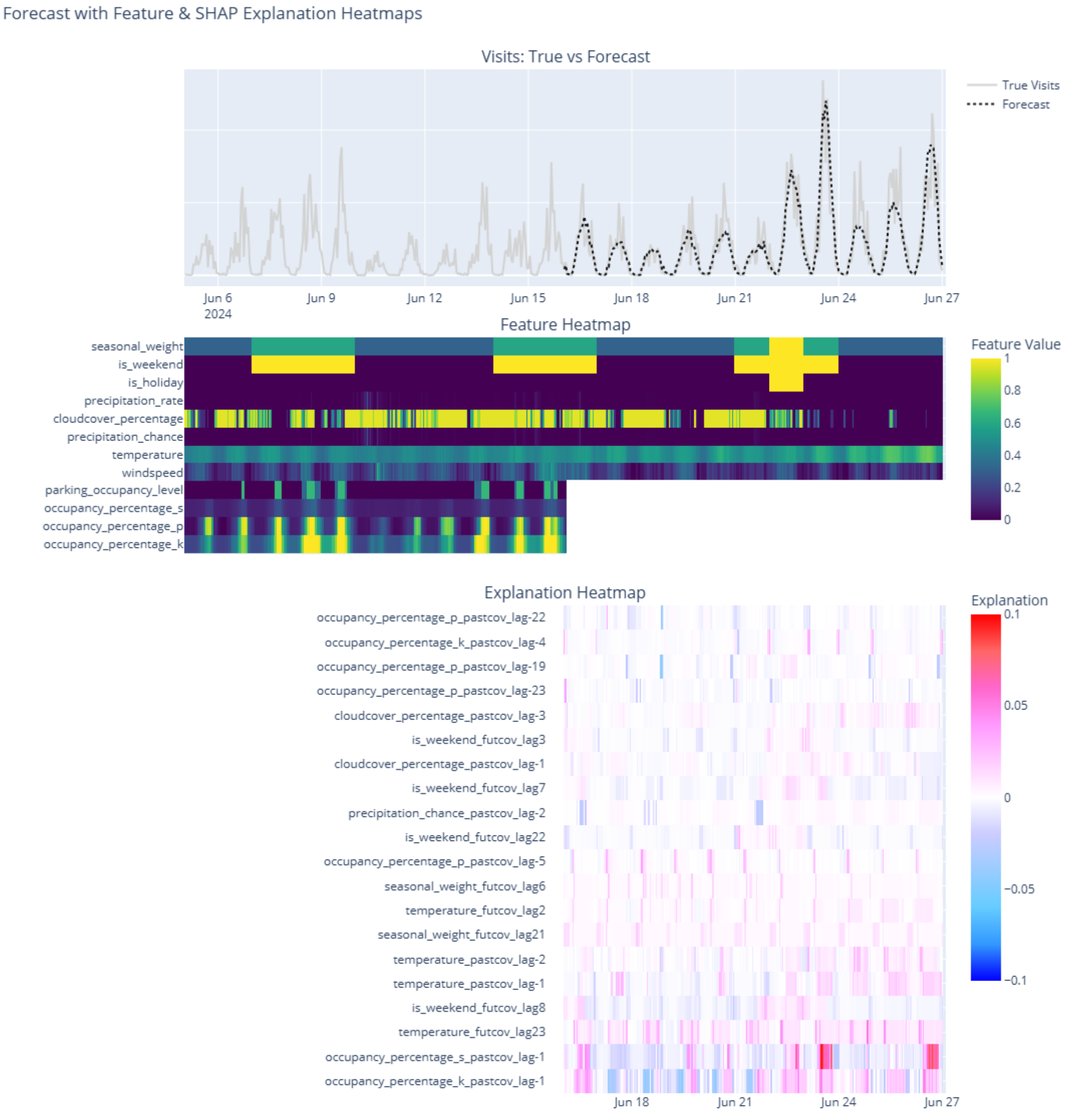

Figure 4: Interactive SHAP dashboard linking model forecasts, input features, and attributions. The shaded area indicates future covariates used during prediction. (y axis values removed for data protection reasons.)

0.7028 and an RMSE of 0.3671. After the first two XAL iterations, performance improves markedly, reaching an R^2 of 0.8351 and an RMSE of 0.3196. The best results are observed after the third iteration, with an R^2 of 0.8491 and an RMSE of 0.3126. Notably, these gains are most significant during high-density periods – rare and imbalanced events that pose challenges for conventional forecasting models (see Figure 1). A slight performance decline is observed in the fourth iteration, suggesting diminishing returns and highlighting the importance of targeted correction rather than excessive re-tuning.

Beyond improvements in forecasting accuracy, the SHAP-based explanations are also more actionable compared to the Local explanation of LIME, or Integrated Gradients (IG), which has proved to be unstable [34]. Standard SHAP visualizations (as shown in Figure 5) often present aggregated attributions that are difficult to interpret in a time series context, especially when trying to understand how specific features influence predictions over a given period. For instance, the contribution of `temperature_lag24` varies across samples, making it challenging to link its influence to concrete time intervals or contextual events.

To address this, we developed a custom visualization that retains the underlying SHAP values and

Table 2

Performance comparison of forecasting models after applying XAL workflow (11-day horizon)

Model	R2	RMSE
Baseline	0.3843	0.7614
XGBoost	0.7028	0.3671
XGBoost after XAL round 1	0.8096	0.3435
XGBoost after XAL round 2	0.8351	0.3196
XGBoost after XAL round 3	**0.8491**	**0.3126**
XGBoost after XAL round 4	0.8358	0.3190

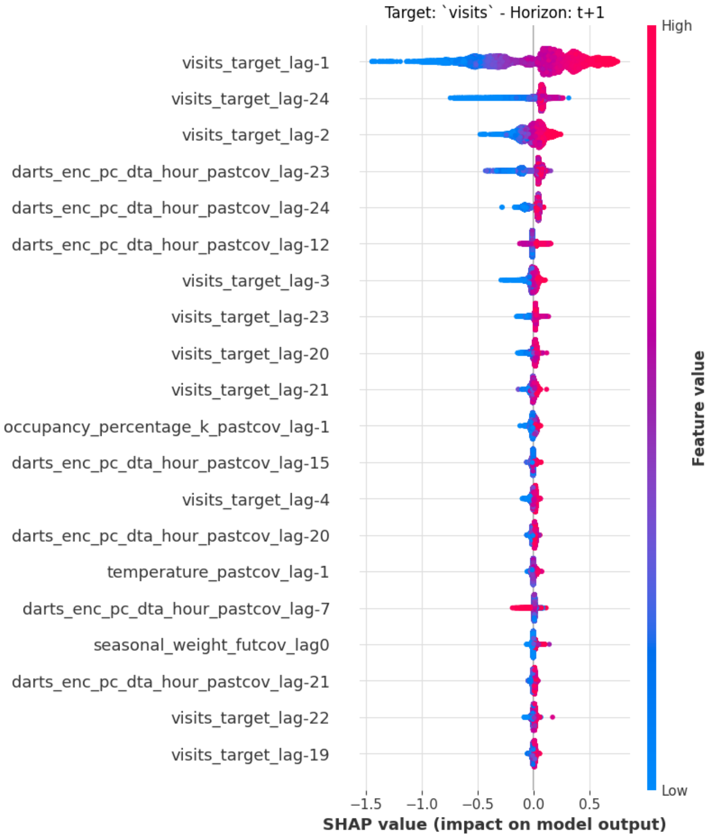

Figure 5: SHAP summary plot ($t + 1$ horizon) highlighting dominant influence of visit lags, parking features, and holiday indicators post-correction.

provides clearer insights into both feature contributions and their temporal dynamics. As shown in Figure 4, this enhanced dashboard allows users to examine SHAP attributions across three distinct zones: the 11-day historical input window, the forecast covariates, and the predicted future period. By aligning SHAP values with the exact timing of input features, the visualization helps users better understand not only "which" features drive the predictions, but also "when" they matter most. This design supports a more intuitive and diagnostic interpretation of model behavior, especially in high-stakes or error-prone intervals.

To study how our XAL workflow improves the identification of high-risk crowd events, we illustrate the results using a traffic-light risk categorization approach, using thresholds defined by domain experts.

After users examined SHAP explanations, identified misleading patterns, and made targeted corrections (e.g., increasing the weight of weather and holiday-related features), the retrained model better captured high-risk periods. These improvements align model behavior with operational constraints, demonstrating how human feedback can effectively close the model–reality gap.

As shown in Table 4, before XAL, the model struggled to identify high-risk (red) events, achieving only 0.33 recall and an F1-score of 0.46, while misclassifying many high-risk (red) periods as medium-risk

Table 3
Confusion matrices for risk level prediction: before (left) and after (right) XAL.

Before XAL					After XAL				
	green	2296	355	0		green	2298	353	0
True	orange	211	1337	16	True	orange	235	1307	22
	red	0	102	51		red	0	88	65
		predicted					Predicted		

(orange).

The confusion matrices in Table 3 further emphasize these gains: pre-XAL, only 51 out of 153 true red events were correctly classified, while 102 were misclassified as medium-risk (orange). Post-XAL, high-risk (red) recall rose to 42% with 65 correct classifications (14 additional high-risk hours flagged accurately). Importantly, no high-risk events were ever classified as low-risk, maintaining operational safety margins.

After incorporating feedback through the explainability interface – particularly by correcting feature attributions, reweighting red events, and adjusting prediction logic – the model's recall on the high-risk (red) class improved by 27%, reaching 0.42, and the F1-score increased to 0.54 (see Table 4). Precision remained stable, ensuring that improved detection of high-risk events did not come at the cost of false alarms.

Table 4
Risk Categorization Performance Before and After XAL

	Before XAL			After XAL		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Green	0.92	0.87	0.89	0.91	0.87	0.89
Orange	0.75	0.85	0.80	0.75	0.84	0.79
Red	0.76	0.33	0.46	0.75	0.42	0.54
Accuracy		0.84			0.84	
Macro avg	0.81	0.68	0.72	0.80	0.71	0.74
Weighted avg	0.85	0.84	0.84	0.84	0.84	0.84

This iterative XAL approach exemplifies how integrating explainability throughout the modeling life cycle can not only improve model transparency but also directly enhance forecasting performance. By empowering users to act on explanation insights – through either manual correction or strategic re-weighting – XAL creates a virtuous loop between interpretation and learning, tailored for complex, high-stakes prediction tasks in urban mobility.

5. Conclusion

In this work, we proposed a hybrid approach combining traditional machine learning with an explainable, human-in-the-loop framework for urban crowd forecasting. Among tested models, XGBoost integrated with our Explainable Active Learning (XAL) loop delivered the best and most robust performance, especially for rare, high-risk crowd events. The use of interactive SHAP dashboards allowed experts to iteratively improve the model by refining feature importance and correcting temporal errors.

Our XAL method not only boosts accuracy but also enhances transparency and usability, increasing trust and recall of critical crowd scenarios. The traffic-light risk framework illustrates how explainability-driven refinement supports urban safety planning.

However, the approach depends on timely expert feedback, which may limit scalability. SHAP visualizations, while helpful, require some technical expertise and do not fully capture complex feature interactions or missing contextual knowledge. Future work will focus on automating parts of the XAL loop, expanding to multi-region forecasting, integrating real-time data sources, and deploying the system in operational decision-support tools for urban stakeholders.

Acknowledgments

This work was funded by the EU's Horizon Europe research and innovation program under Grant No. 101093051 EMERALDS. We gratefully acknowledge the valuable collaboration and support of all project partners. In particular, we would like to thank TU Delft for their academic guidance and technical contributions, and Argaleo for providing domain expertise and insights that were essential to the applied implementation and evaluation of our forecasting framework.

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT-4o and Grammarly in order to paraphrase and reword the text. After using ChatGPT/Grammarly, the authors reviewed and edited the content as needed and took full responsibility for the manuscript's content.

References

[1] A. Jalali, A. Graser, C. Heistracher, Towards explainable ai for mobility data science, arXiv preprint arXiv:2307.08461 (2023). doi:https://doi.org/10.48550/arXiv.2307.08461.

[2] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1724–1734. URL: https://aclanthology.org/D14-1179/. doi:10.3115/v1/D14-1179.

[4] B. Lim, S. O. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting 37 (2021) 1748–1764. URL: https://www.sciencedirect.com/science/article/pii/S0169207021000637. doi:https://doi.org/10.1016/j.ijforecast.2021.03.012.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[6] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[7] K. H. Poon, P. K.-Y. Wong, J. C. Cheng, Long-time gap crowd prediction using time series deep learning models with two-dimensional single attribute inputs, Advanced Engineering Informatics 51 (2022) 101482. URL: https://www.sciencedirect.com/science/article/pii/S1474034621002329. doi:https://doi.org/10.1016/j.aei.2021.101482.

[8] C.-Z. T. Xie, J. Xu, B. Zhu, T.-Q. Tang, S. Lo, B. Zhang, Y. Tian, Advancing crowd forecasting with graphs across microscopic trajectory to macroscopic dynamics, Information Fusion 106 (2024) 102275. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000538. doi:https://doi.org/10.1016/j.inffus.2024.102275.

[9] J. Chen, X. Shi, H. Zhang, W. Li, P. Li, Y. Yao, S. Miyazawa, X. Song, R. Shibasaki, Mobcovid: Confirmed cases dynamics driven time series prediction of crowd in urban hotspot, IEEE Transactions on Neural Networks and Learning Systems 35 (2024) 13397–13410. doi:10.1109/TNNLS.2023.3268291.

[10] A. Andreoni, M. N. Postorino, et al., A multivariate arima model to forecast air transport demand, Proceedings of the Association for European Transport and Contributors (2006) 1–14.

[11] J.-F. Determe, U. Singh, F. Horlin, P. De Doncker, Forecasting crowd counts with wi-fi systems: Univariate, non-seasonal models, IEEE Transactions on Intelligent Transportation Systems 22 (2021) 6407–6419. doi:10.1109/TITS.2020.2992101.

[12] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures, Neural Networks 18 (2005) 602–610. URL: https://www.sciencedirect.com/science/article/pii/S0893608005001206. doi:https://doi.org/10.1016/j.neunet.2005.06.042, iJCNN 2005.

[13] U. Singh, J.-F. Determe, F. Horlin, P. D. Doncker, Crowd forecasting based on wifi sensors and lstm neural networks, IEEE Transactions on Instrumentation and Measurement 69 (2020) 6121–6131. doi:10.1109/TIM.2020.2969588.

[14] X. Fu, G. Yu, Z. Liu, Spatial–temporal convolutional model for urban crowd density prediction based on mobile-phone signaling data, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 14661–14673. doi:10.1109/TITS.2021.3131337.

[15] T. Goktug Altundogan, M. Karaköse, O. Yaman, S. Tanberk, F. Mert, A. Egemen Yılmaz, Dynamic fuzzy cognitive maps-based crowd analysis using time series obtained from video processing, IEEE Access 13 (2025) 33813–33833. doi:10.1109/ACCESS.2025.3542190.

[16] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, Y. Liu, Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station, Soft Computing 24 (2020) 16453–16482.

[17] A. Garza, C. Challu, M. Mergenthaler-Canseco, Timegpt-1, arXiv preprint arXiv:2310.03589 (2023).

[18] A. Graser, Timeseries foundation models for mobility: A benchmark comparison with traditional and deep learning models, 2025. URL: https://arxiv.org/abs/2504.03725. arXiv:2504.03725.

[19] W. Jo, D. Kim, Neural additive time-series models: Explainable deep learning for multivariate time-series prediction, Expert systems with applications 228 (2023) 120307.

[20] F. Yaprakdal, M. Varol Arısoy, A multivariate time series analysis of electrical load forecasting based on a hybrid feature selection approach and explainable deep learning, Applied Sciences 13 (2023) 12946.

[21] R. Saluja, A. Malhi, S. Knapič, K. Främling, C. Cavdar, Towards a rigorous evaluation of explainability for multivariate time series, arXiv preprint arXiv:2104.04075 (2021).

[22] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3145–3153. URL: https://proceedings.mlr.press/v70/shrikumar17a.html.

[23] Y. Zhang, O. Petrosian, J. Liu, R. Ma, K. Krinkin, Fi-shap: explanation of time series forecasting and improvement of feature engineering based on boosting algorithm, in: Proceedings of SAI intelligent systems conference, Springer, 2022, pp. 745–758.

[24] A. Jutte, F. Ahmed, J. Linssen, M. van Keulen, C-shap for time series: An approach to high-level temporal explanations, arXiv preprint arXiv:2504.11159 (2025).

[25] Q. Li, Y. Ji, M. Zhu, X. Zhu, L. Sun, Unsupervised feature selection using chronological fitting with shapley additive explanation (shap) for industrial time-series anomaly detection, Applied Soft Computing 155 (2024) 111426.

[26] B. Lim, S. Ö. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting 37 (2021) 1748–1764.

[27] B. Wu, L. Wang, Y.-R. Zeng, Interpretable wind speed prediction with multivariate time series and temporal fusion transformers, Energy 252 (2022) 123990.

[28] T. Wu, J. Ortiz, Rlad: Time series anomaly detection through reinforcement learning and active learning, 2021. URL: https://arxiv.org/abs/2104.00543. arXiv:2104.00543.

[29] R. van Leeuwen, G. Koole, Anomaly detection in univariate time series incorporating active learning, Journal of Computational Mathematics and Data Science 6 (2023) 100072.

[30] P. Gupta, M. Gupta, V. Kumar, An active learning enhanced data programming (actdp) framework for ecg time series, Machine Learning: Science and Technology 5 (2024) 035016.

[31] S. M. del Campo Barraza, W. Lindskog, D. Badalotti, O. Liew, A. Toyser, Active learning framework

for time-series classification of vibration and industrial process data, in: Annual Conference of the PHM Society, volume 13, 2021.

[32] K. Choi, J. Yi, C. Park, S. Yoon, Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines, IEEE access 9 (2021) 120043–120065.

[33] S. Das, An active learning framework with a class balancing strategy for time series classification, arXiv preprint arXiv:2405.12122 (2024).

[34] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: 2019 IEEE-CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 4197–4201.