electronics

MDPI

Article

Large Language Models for Predictive Maintenance in the Leather Tanning Industry: Multimodal Anomaly Detection in Compressors

Giulia Palma *, Gaia Cecchi and Antonio Rizzo

Dipartimento di Scienze Sociali Politiche e Cognitive, Università degli Studi di Siena, 53100 Siena, Italy; gaia.cecchi2@unisi.it (G.C.); antonio.rizzo@unisi.it (A.R.)
* Correspondence: giulia.palma2@unisi.it

Abstract: Predictive maintenance in industrial settings increasingly demands systems capable of integrating heterogeneous data streams while balancing computational efficiency and contextual reasoning. This paper introduces a novel framework leveraging Large Language Models (LLMs) to address these challenges in compressor monitoring, demonstrating their potential to enhance anomaly detection accuracy and operational cost-effectiveness. We evaluate Qwen 2.5-32B against traditional machine learning models (ANN, CNN, LSTM), achieving superior recall (92.3%) and AUC-ROC (0.991) through transformer-based architectures optimized for multimodal data fusion. A financial case study reveals operational cost reductions of 18% via reduced downtime and optimized maintenance schedules, while a real-time monitoring dashboard validates scalability for industrial deployment. Our findings highlight the transformative role of LLMs in bridging technical innovation with domain-specific operational constraints, offering a blueprint for predictive maintenance in niche industries.

Keywords: predictive maintenance; Large Language Models; multimodal data fusion; anomaly detection; industrial compressors; real-time monitoring; machine learning

check for updates

Academic Editor: Marcin Witczak

Received: 29 April 2025
Revised: 14 May 2025
Accepted: 16 May 2025
Published: 20 May 2025

Citation: Palma, G.; Cecchi, G.; Rizzo, A. Large Language Models for Predictive Maintenance in the Leather Tanning Industry: Multimodal Anomaly Detection in Compressors. *Electronics* **2025**, *14*, 2061. https://doi.org/10.3390/electronics14102061

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Predictive maintenance (PdM) has garnered substantial attention in recent years, propelled by rapid advancements in artificial intelligence (AI), machine learning (ML), and deep learning (DL) techniques. The integration of heterogeneous data streams from diverse industrial sensors and systems necessitates sophisticated frameworks capable of handling multimodal information while ensuring computational efficiency and contextual reasoning. Recent studies have demonstrated the efficacy of transformer-based architectures for multimodal data fusion, which leverage self-attention mechanisms to capture complex interdependencies among different data modalities, thereby enhancing anomaly detection and predictive accuracy [1–3]. For instance, Fidma et al. [1] proposed a hybrid data-driven and knowledge-based predictive maintenance framework that combines ontologies with ML and data mining techniques to address interoperability challenges in Industry 4.0 manufacturing environments, achieving improved real-time anomaly prediction and decision-making. Similarly, Ak et al. [2] introduced a transformer architecture that effectively integrates textual and visual modalities through both early and late fusion strategies, resulting in superior classification performance without compromising runtime efficiency. Furthermore, Shvetsova et al. [3] developed a modality-agnostic fusion transformer for

video retrieval tasks, which can be adapted to industrial settings to process variable-length multimodal inputs, facilitating scalable and robust predictive maintenance solutions.

The application of these transformer-based models extends beyond traditional image and video domains into industrial IoT (IIoT) contexts, where they enable seamless fusion of sensor data, operational logs, and maintenance records [4]. This multimodal approach addresses the limitations of conventional ML models such as ANN, CNN, and LSTM by capturing richer contextual information and temporal dynamics inherent in industrial processes [5]. Moreover, the financial impact of adopting such advanced predictive maintenance frameworks is significant, with empirical studies reporting operational cost reductions through optimized maintenance scheduling and reduced downtime [1,4]. Collectively, these advancements underscore the transformative potential of Large Language Models and transformer-based architectures in predictive maintenance, providing a blueprint for integrating technical innovation with domain-specific operational constraints in niche industrial sectors.

In light of this, predictive maintenance has emerged as a cornerstone of Industry 4.0, enabling manufacturing industries to mitigate downtime risks, optimize resource allocation, and enhance operational resilience [6]. Traditional machine learning (ML) approaches have advanced PdM by analyzing structured data from sensors, historical maintenance logs, and operational parameters to forecast equipment failures [7]. However, these methods often struggle with unstructured data integration, contextual adaptation, and dynamic scenario modelling—critical limitations in sectors like leather tanning, where domain-specific operational nuances significantly influence maintenance outcomes [8].

The leather tanning industry presents unique challenges for PdM. Compressors, which power pneumatic systems essential for hide processing, operate in harsh environments characterized by high humidity, chemical exposure, and variable load cycles. These conditions accelerate mechanical wear, yet traditional ML frameworks frequently overlook critical insights embedded in unstructured data such as technician notes, maintenance reports, and equipment manuals [9]. For instance, studies highlight that ML models trained solely on sensor data often fail to capture failure patterns linked to chemical corrosion or operator-specific maintenance practices [10], resulting in suboptimal predictive accuracy.

The advent of Large Language Models (LLMs) offers transformative potential for addressing these limitations. By integrating natural language processing (NLP) capabilities with structured data analysis, LLMs enable holistic PdM frameworks that contextualize equipment behavior within operational realities [11].

While this study focuses on compressors in a specific Italian tannery, representing a challenging niche environment, the core objective extends beyond this single application. We aim to introduce and validate a methodological framework leveraging Large Language Models (LLMs) for advanced predictive maintenance through multimodal data fusion. This framework is designed to be adaptable to other equipment types and similar industrial settings facing complex operational realities and diverse data streams, thereby offering a potentially transformative approach for sectors often underserved by complex AI solutions.

Our study demonstrates the feasibility and benefits of this LLM-based methodology in a real-world scenario, introducing an innovative application of LLMs to predictive maintenance for compressors in a tannery located in Santa Croce sull'Arno, Italy, a region globally recognized for its leather manufacturing tradition. This research marks a significant step forward by addressing critical challenges in the field. First, LLMs excel at integrating multimodal data, such as unstructured maintenance logs, technical manuals, and sensor telemetry, to uncover failure patterns that traditional machine learning methods often overlook. For instance, by correlating vibration data from compressors with technician

annotations regarding chemical exposure, LLMs can identify degradation mechanisms that remain undetected by numerical models alone.

Additionally, LLMs demonstrate an exceptional ability to adapt to the specific operational contexts of the leather tanning industry. By learning from historical data unique to this domain, they can account for sectoral nuances such as seasonal production cycles and the impact of chemical treatments on equipment performance. This adaptability enhances their capacity to detect anomalies in dynamic and complex environments. Finally, LLMs contribute to human-centric decision-making by generating maintenance reports in natural language. These reports simplify complex predictive insights, making them accessible and actionable for operational teams and bridging the gap between technical analysis and practical implementation.

The novelty of this research lies in its vertical application of LLMs to a traditionally underserved sector. While prior studies have explored LLMs in general manufacturing contexts [12], few have addressed their role in integrating tacit knowledge and structured data in niche industries like leather tanning. Our findings demonstrate that LLMs can transform unstructured data into competitive advantages, enabling tanneries to transition from reactive maintenance paradigms to proactive, data-driven strategies.

The integration of LLMs into predictive maintenance frameworks represents a paradigmatic shift beyond conventional analytical approaches. Rather than functioning as passive data processors, these models demonstrate an unprecedented capacity to synthesize quantitative sensor data with contextual operational expertise, effectively replicating the nuanced reasoning of human technicians. This capability assumes particular relevance in traditional industries such as leather tanning, where domain-specific knowledge and tacit operational practices remain critical to maintaining competitiveness in global markets. By validating LLM-driven predictive maintenance in a real-world tannery environment, this study establishes a replicable methodology for industries seeking to leverage advanced AI while preserving the operational wisdom accumulated through decades of sectoral specialization. The demonstrated ability of LLMs to contextualize failure patterns within the unique constraints of leather processing—such as chemical exposure variability and seasonal production cycles—provides a blueprint for industries navigating similar challenges in balancing technological innovation with domain-specific operational realities.

This paper is structured as follows. Section 2 reviews the evolution of predictive maintenance methodologies, emphasizing the transition from traditional machine learning models to Large Language Models (LLMs). Section 3 details the dataset description, project workflow, model architectures of LLMs, and their mathematical justification compared to traditional machine learning models. Section 4 presents and discusses the experimental results, including comparisons between various LLMs, evaluations against traditional models, a case study on the financial advantages of LLMs in predictive maintenance, and the user interface implementation. Finally, Section 5 collects some conclusions and delineates potential future perspectives.

2. Related Works in Predictive Maintenance

Predictive maintenance has gained significant attention in recent years, driven by advancements in artificial intelligence, machine learning, and deep learning models. Various studies have explored different approaches to enhance predictive maintenance strategies across industrial applications.

Text-based methods leveraging Large Language Models (LLMs) have demonstrated promising results. For instance, Lowin [11] proposed an LLM-based approach, specifically investigating how LLMs can be coupled with association rule mining for predictive maintenance in facility management by analyzing textual maintenance requests to identify

upcoming needs. Similarly, Postiglione and Monteleone [13] explored linguistic text mining techniques, focusing on finite automata to analyze log messages from industrial machinery, with the objective of improving the precision and promptness of predictive maintenance strategies in the manufacturing sector.

Artificial intelligence (AI) and deep learning techniques, including Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, have been extensively studied for predictive maintenance applications [7,14,15]. In [15], Shoorkand et al. developed a hybrid CNN-LSTM model for the joint optimization of production and imperfect predictive maintenance planning in multi-state systems, aiming to minimize overall costs by accurately predicting the health condition of multiple machines. Kim et al. [7] focused on optimizing design parameters in LSTM models for predictive maintenance of bearing components in machinery, using a genetic algorithm to tune hyperparameters for improved Remaining Useful Life (RUL) estimation. Borré et al. [14] introduced a hybrid CNN-LSTM attention-based model for machine fault detection, specifically applied to electrical motors, with the objective of enhancing anomaly detection accuracy through quantile regression to manage uncertainties in vibration data. Hybrid architectures combining CNN and LSTM models have been particularly effective in fault detection. For example, Wahid et al. [16] proposed a hybrid CNN-LSTM framework with skip connections for multivariate time-series forecasting to predict machine failures in an Industry 4.0 context, demonstrating high prediction accuracy on a real-world dataset from manufacturing equipment.

Multimodal data fusion has emerged as a key area of research. For instance, Alsaif et al. [17] introduced a fault detection and diagnosis framework utilizing multimodal LLMs (specifically GPT-4-Preview) within the Industry 4.0 context, aiming to improve scalability and generalizability by processing diverse data types from industrial systems and leveraging synthetic datasets for augmenting knowledge. Zhang et al. [18] demonstrated an LLM-enhanced wavelet packet transformation, where LLMs automatically optimize WPT parameters for improved fault prediction in synchronous condensers within UHVDC transmission systems, addressing limitations of manual parameter selection.

Recent works have pioneered novel frameworks combining sensor data with textual maintenance records. In [19], Dalzochio et al. introduced a BiLSTM-CRF network with attention mechanisms that achieves 16.9% accuracy improvement in press line maintenance (an industrial system) through event log–sensor fusion for the objective of predicting equipment failures. For energy systems, ref. [20] proposed a federated learning framework with neural circuit policies demonstrating 92% anomaly detection accuracy in distributed compressor networks (specifically in leather tanning operations), aiming to enhance predictive maintenance while maintaining data privacy.

Moreover, in compressor monitoring, Aminzadeh et al. [9] presented a machine learning-based framework employing Linear Regression for predictive maintenance of industrial air compressors, showcasing the effectiveness of multi-sensor data acquisition (temperature, pressure, flow rate) and SQL databases for real-time monitoring and alert generation. Similarly, Qu et al. [10] proposed a multimodal fusion approach combining an attention-based autoencoder and a Generative Adversarial Network (MFGAN) for industrial anomaly detection, specifically applied to roller press equipment in a cement plant, with the objective of fusing DCS data and acoustic signals to improve robustness and generalization.

Recent studies have highlighted the impact of generative AI on maintenance strategies. Lee et al. [21] examined the cognitive effects of generative AI, specifically focusing on how LLMs impact human critical thinking and cognitive effort in predictive maintenance decision-making within knowledge work contexts by surveying knowledge workers.

Salierno et al. [22] discussed the role of generative AI (GAI) and LLMs in shaping smarter sustainable cities and sustainable smart manufacturing within the Industry 5.0 paradigm, targeting industrial optimization, predictive maintenance, and urban management through selected case studies.

Recent advances in structural reliability analysis have focused on developing adaptive and computationally efficient methods to accurately estimate failure probabilities in complex engineering systems. For example, in [23], a novel learning function for adaptive surrogate-model-based reliability evaluation was proposed to enhance model accuracy in structural reliability analysis while reducing computational cost by dynamically adjusting the surrogate model. This approach addresses the challenge of balancing precision and efficiency in high-dimensional reliability problems, enabling more robust predictions under uncertainty. Complementing this, Yang et al. [24] introduced an enhanced soft Monte Carlo simulation (EMCS) framework combined with a generalized enhanced simulation scaling formula and support vector regression (SVR) to efficiently evaluate failure probabilities in structural systems. This method outperforms traditional Monte Carlo simulations by reducing the required number of samples while maintaining high accuracy, particularly in scenarios involving low failure probabilities. Additionally, the integration of SVR as a surrogate model in [25] was used for predicting tunnel collapse probability, facilitating nonlinear mapping between random variables and performance functions, further improving computational efficiency in geotechnical engineering. These recent contributions highlight the potential of hybrid simulation and machine learning approaches to advance predictive maintenance and reliability evaluation in industrial applications, offering scalable solutions that reconcile accuracy with computational feasibility.

Advancements in Named Entity Recognition (NER) have also contributed to maintenance optimization. For instance, Cao et al. [26] developed a maintenance-specific NER method based on multi-level clue-guided prompt learning (MCP) using LLMs (specifically GPT-3.5-turbo) for shearer maintenance in the coal mining industry, aiming to improve fault detection accuracy by extracting entities from unstructured technical texts and maintenance manuals.

Several reviews have provided insights into the evolution of predictive maintenance methodologies. For example, Ucar et al. [27] presented a comprehensive survey on AI-based predictive maintenance, analyzing its key components (such as sensors, data preprocessing, AI algorithms), trustworthiness, and future trends for various industrial systems. Similarly, Vithi and Chibaya [6] conducted a bibliometric review focusing on diagnostic models (ANNs and SVMs) that utilize machine learning techniques for predictive maintenance, primarily in railway systems for fault detection in wheelset axle bearings. While Çınar et al. [8] provided a comprehensive review on the advancements of Machine Learning (ML) techniques applied to Predictive Maintenance (PdM) for sustainable smart manufacturing in Industry 4.0, classifying existing research based on ML algorithms, machinery types, and data characteristics to offer guidelines for automated fault detection and diagnosis. Myöhänen [12] presented a comprehensive systematic literature review for the objective of summarizing predictive maintenance and process optimization applications leveraging LLMs across various industrial sectors, identifying use cases (like root cause analysis, anomaly detection), and implementation factors (like model training, organizational readiness).

While these studies collectively underscore the transformative potential of AI, machine learning, and LLMs in predictive maintenance, a notable gap exists in the vertical application of advanced LLMs to niche, traditionally underserved industries like leather tanning. Many existing works focus on general manufacturing contexts [12] or specific ML techniques without deep LLM integration for multimodal data specific to harsh operational

environments. Our research differentiates itself by specifically applying a State-of-the-Art LLM, Qwen 2.5-32B, to the predictive maintenance of compressors within the unique and challenging environment of a leather tannery. The novelty lies in (1) the tailored multimodal data fusion approach, integrating sensor telemetry with unstructured domain-specific data such as technician notes and detailed environmental stressors prevalent in tanneries (e.g., chemical exposure, high humidity), (2) the demonstration of superior anomaly detection performance of LLMs against traditional ML models in this specific context, and (3) the holistic evaluation including a financial impact analysis and a real-time monitoring dashboard designed for this niche application. This provides a targeted blueprint for similar industries facing complex operational constraints.

Distinguishing Features and Innovations of the Proposed Methodology

The growing body of literature on predictive maintenance and reliability evaluation has introduced a variety of data-driven and hybrid approaches that leverage machine learning, surrogate modeling, and simulation techniques to improve anomaly detection and failure prediction [1,23,24]. While these methods have demonstrated promising results within specific domains or under constrained conditions, they often face limitations in adaptability, scalability, and interpretability when applied to heterogeneous industrial environments. The methodology proposed in this work diverges from existing approaches by emphasizing a domain-agnostic, systematic framework that exploits the unique capabilities of LLMs for multimodal anomaly detection and contextual reasoning. This section outlines the key differentiators and novel contributions of our approach relative to the current State of the Art.

Firstly, unlike traditional predictive maintenance models that are typically designed and optimized for narrowly defined industrial applications—such as structural reliability analysis using surrogate models [23] or enhanced Monte Carlo simulations coupled with SVR [24]—our methodology is inherently generalizable. It does not aim to produce a domain-specific solution but rather proposes a flexible framework adaptable to diverse application areas, including control systems, manufacturing, and maintenance operations. The central innovation lies in harnessing the pattern recognition and contextual sensitivity of LLMs to systematically explore the "field of the possible". This enables detection of subtle anomalies and outlier conditions that may not be captured by conventional models relying on fixed feature sets or handcrafted rules. By focusing on the methodological paradigm rather than a fixed application, our work opens pathways for cross-domain transferability and rapid adaptation to emerging industrial challenges.

Secondly, the proposed approach introduces a reproducible and modular architecture that integrates multimodal data inputs, anomaly detection mechanisms, and pattern-based reasoning within a unified transformer-based framework. This contrasts with many existing methods that treat these components in isolation or rely on domain-specific feature engineering [2,3]. Our framework's modularity ensures that it can be replicated and customized across different operational contexts without extensive redesign, facilitating scalability and ease of deployment. Furthermore, the use of LLMs enables a novel form of "reasoning" that, while distinct from human cognitive processes, allows for a sophisticated combinatorial exploration of data patterns through progressive contextual enrichment of token representations. This capability enhances the system's ability to generate meaningful anomaly hypotheses that support human decision-making rather than attempting to replace it.

Lastly, a fundamental conceptual novelty of our work is the explicit recognition and operationalization of the difference between deduction and inference within the context of LLM-based reasoning. Unlike traditional AI systems that often conflate these processes, our

methodology clarifies that LLMs perform a form of deduction-explicating and reorganizing information already implicit in the data-rather than true inference or induction, which would generate new knowledge or hypotheses independently [25]. This distinction informs the design of the framework as a human-in-the-loop system, where the LLM-generated outputs serve as exploratory causal scenarios that require expert evaluation and refinement. This approach respects the limitations of current LLM architectures while leveraging their strengths in pattern recognition and contextualization, thereby fostering a collaborative synergy between artificial intelligence and human expertise.

3. Materials and Methods

This section outlines the methodology and components of the predictive maintenance framework. The subsections below address critical aspects of the study. Dataset characteristics (see Section 3.1) are detailed to ensure transparency in data composition and preprocessing, as the quality of input data directly impacts model performance. Project workflow (see Section 3.2) is presented to clarify the end-to-end pipeline, from data ingestion to real-time deployment, emphasizing reproducibility and scalability. Then, model architectures (see Section 3.3) are analyzed to highlight the technical advantages of LLMs, such as transformer layers and multimodal processing, which enable contextual anomaly detection. Moreover, comparison with traditional models (see Section 3.4) is included to contrast LLMs with ANN, CNN, and LSTM architectures, addressing trade-offs in computational efficiency and feature engineering requirements. Finally, mathematical justification (see Section 3.5) provides theoretical foundations for selecting LLMs, emphasizing their capacity to handle unstructured data and contextual reasoning, which traditional models lack.

3.1. Dataset Description

The dataset was collected between March and May 2024 at a tannery facility in Santa Croce sull'Arno, Italy, with hourly sampling frequency. This temporal granularity balances analytical resolution with operational representativeness, capturing typical leather processing load dynamics while avoiding excessive noise from overly frequent sampling. The three-month window ensures coverage of key stages in the production cycle, including periods of stable operation and seasonal peaks. Data spans two critical phases: a moderate load phase (March–April) characterized by steady production, and a seasonal peak (May) driven by increased spring hide demand. This design allows testing model robustness under both routine and high-activity scenarios, ensuring representativeness for real-world validation.

The dataset is stored in a SQLite database (ver. 3.49.1), chosen for its efficiency in handling large operational and meteorological data volumes, surpassing traditional Python (ver. 24.3.1) memory limits and ensuring scalability for industrial applications. It is structured into four primary tables (see Table 1), each capturing different aspects of compressor behavior and environmental factors for targeted analysis of operational patterns, anomaly detection, and contextual influences critical to predictive maintenance:

compressor_data stores hourly sensor readings (vibration, temperature, pressure), aligned with known compressor degradation mechanisms in leather processing, aiding early failure detection.

anomalies contains confirmed failure events, labeled using validated thresholds (e.g., vibration > 4.0 mm/s, temperature > 100 °C), ensuring model predictions are based on real-world maintenance records and addressing a common limitation in industrial datasets where anomalies are often inferred or underreported.

false_positives focuses on distinguishing transient fluctuations from true failures. By retaining borderline cases where sensor values exceed thresholds but correspond to normal operation, this table helps enhance the model's robustness, minimizing unnecessary maintenance stops that could disrupt tannery operations.

compressor_data_with_weather integrates meteorological variables such as humidity and temperature, accounting for environmental stressors affecting compressor performance. For example, humidity levels above 80% can accelerate corrosion, a factor often overlooked in traditional datasets, but critical in chemical-intensive environments.

Table 1. Primary dataset tables.

Table Name	Purpose
compressor_data	Core sensor readings
anomalies	Confirmed anomaly events
false_positives	Borderline false positive cases
compressor_data_with_weather	Sensor data enriched with meteorological data

These tables enable a three-phase analytical workflow: baseline characterization to establish operational norms, failure validation by cross-referencing anomalies with maintenance records, and contextual refinement using environmental factors to interpret performance variations. Finally, robustness testing evaluates false positives to ensure model reliability and minimize unnecessary interventions. This approach improves predictive maintenance models, balancing technical accuracy with practical applicability in environments where mechanical and environmental stressors interact.

Each dataset record is timestamped (YYYY-MM-DD HH:MM:00) and includes 17 parameters divided into compressor-specific and meteorological categories (Table 2). This multidisciplinary selection integrates mechanical, electrical, and environmental factors relevant to compressor degradation in tanneries.

The selection of the parameters included in the dataset was primarily guided by domain expertise concerning compressor degradation mechanisms, particularly those relevant to the harsh operating conditions found in tanneries, such as high humidity and chemical exposure impacting corrosion and component wear. Data preprocessing steps were tailored to the specific requirements of the different modeling approaches compared in this study.

Table 2. Record parameters.

Category	Variables
Compressor Parameters	Temperature (°C), Vibration (mm/s), Operating Pressure (bar), Rotation Speed (RPM), Electrical Current (A), Voltage (V), Power Factor (CosPhi), Energy Consumption, Reactive Energy
Meteorological Variables	Ambient Temperature, Humidity, Wind Speed, Precipitation, Atmospheric Conditions

For analysis using traditional Machine Learning models (ANN, CNN, LSTM), the structured numerical data from sensors (Temperature, Vibration, Operating Pressure, etc.) and meteorological variables (Ambient Temperature, Humidity, etc.) were normalized using Z-score standardization to ensure features were on a comparable scale. Specific feature engineering steps for the ANN are detailed in Section 3.4. For the Large Language Model (LLM) analysis, these same numerical data streams were converted into a structured textual format. This involved serializing the time-series readings, including their timestamps and units (e.g., '2024-05-26 08:00:00, Temperature: 78.5 °C, Vibration: 1.2 mm/s...'), creating

descriptive strings (compressor_str, weather_str) that were then incorporated directly into the LLM's input prompt (see Section 3.3).

To illustrate the dataset structure, Table 3 shows the selected portion of the dataset, which includes both normal operating conditions and anomalies. The dataset includes various sensor parameters relevant to predictive maintenance. Temperature is measured in degrees Celsius (°C), humidity in percentage relative humidity (% RH), wind speed in meters per second (m/s), and precipitation in millimeters (mm). Electrical parameters include current measured in amperes (A), power factor ($\cos\varphi$) as a dimensionless quantity, energy consumption in kilowatt-hours (kWh), reactive power in kilovolt-amperes reactive (kVAR), and voltage in volts (V). Mechanical and operational parameters include vibration velocity measured in millimeters per second (mm/s), pressure in bars (bar), and rotational speed in revolutions per minute (rpm).

Table 3. Preview of a selected portion of the dataset used in this study, showing both normal operating conditions and an anomaly event (voltage fluctuation on 26 May 2024 at 14:00). The dataset includes features such as sensor telemetry, operational parameters, and environmental conditions, providing a comprehensive basis for anomaly detection.

DateTime	Weather					Current		Voltage			Other Metrics			
	Temp	Hum	Wind	Precip	Condition	Current	CosPhi	Energy	Reactive	Voltage	Vibr	Temp	Press	Speed
2024-05-26 08:00:00	19.5	80.0	5.5	0.0	Fair	50.0	0.972	3.39	2784.73	380.0	2.16	87.3	6.78	2800.0
2024-05-26 09:00:00	20.5	78.0	6.0	0.0	Partly Cloudy	50.0	0.980	2.14	1198.67	380.0	1.94	87.5	6.77	2800.0
2024-05-26 10:00:00	21.5	68.0	10.0	0.0	Partly Cloudy	50.0	0.964	5.01	3381.59	380.0	2.56	100.1	6.78	2800.0
2024-05-26 11:00:00	22.0	66.0	15.0	0.0	Fair	50.0	0.980	1.89	1165.59	380.0	1.84	88.1	6.84	2800.0
2024-05-26 12:00:00	22.0	69.0	18.0	0.0	Fair	50.0	0.980	1.89	1961.80	380.0	2.35	82.6	6.63	2800.0
2024-05-26 13:00:00	22.5	62.0	17.0	0.0	Fair	50.0	0.974	4.14	2830.79	380.0	1.96	82.9	6.89	2800.0
2024-05-26 14:00:00	22.0	71.0	18.0	0.0	Fair	50.0	0.980	1.51	760.48	380.0	1.61	84.6	6.48	2800.0
2024-05-26 15:00:00	22.0	73.0	16.0	0.0	Fair	50.0	0.968	3.76	2635.95	380.0	2.08	78.2	6.69	2800.0
2024-05-26 16:00:00	22.0	76.0	18.0	0.0	Fair	50.0	0.980	2.26	1263.77	380.0	2.04	84.4	6.85	2800.0
2024-05-26 17:00:00	21.0	78.0	14.0	0.0	Fair	50.0	0.980	1.89	1256.60	380.0	1.74	86.4	6.87	2800.0

Technical parameters were selected to track key degradation mechanisms. Thermal indicators help detect overheating risks linked to prolonged chemical exposure. Mechanical stressors, such as vibration patterns and rotational speed variations, signal bearing wear and imbalances. Pressure fluctuations indicate seal degradation, while electrical metrics (current, voltage stability, power factor (CosPhi)) reveal motor inefficiencies affecting pneumatic systems. Energy consumption patterns enable efficiency assessments, supporting data-driven maintenance decisions.

Environmental parameters contextualize compressor performance in tannery conditions. Humidity over 80% accelerates corrosion, especially where chemical residues exacerbate moisture effects. Temperature fluctuations impact pressure ratios and efficiency, while heat from drying operations adds mechanical stress. Extreme weather events (e.g., heavy rain, strong winds) threaten outdoor compressors, increasing risks of moisture ingress and thermal shocks. Categorical labels (e.g., "Rain") allow systematic analysis of climate-driven operational trends. Together, these parameters emphasize the need for adaptive compressor designs and maintenance strategies suited to tannery environments, where industrial and environmental stressors intersect.

The dataset comprises 2208 records. The class distribution reflects real-world operational realities in tannery environments, where anomalies are rare but critical (Table 4). The dataset contains:

- Normal values: 98.73% (2172 records)
- Confirmed anomalies: 0.50% (11 records)
- False positives: 0.77% (17 records)

Table 4. Class distribution.

Class	Percentage
Normal values	98.73%
Confirmed anomalies	0.50%
False positives	0.77%

The 98.73% normal values represent stable compressor operation under typical tannery conditions, including chemical exposure and humidity fluctuations, providing a baseline for modeling healthy performance patterns.

Confirmed anomalies (0.50%) correspond to validated failure events such as bearing degradation or overheating, critical for training predictive models to recognize early warning signs. Their scarcity mirrors real-world industrial challenges but ensures alignment with maintenance logs.

False positives (0.77%) represent borderline cases where sensor readings exceed thresholds but correspond to normal behavior. Retaining these cases forces models to develop robust discrimination capabilities, reducing unnecessary downtime from false alarms—a key requirement in tanneries where unscheduled maintenance disrupts chemical treatment cycles.

This distribution necessitates specialized modeling strategies to address class imbalance while maintaining operational relevance. The inclusion of all three categories ensures models balance accuracy on rare anomalies with reliability in distinguishing transient fluctuations from true failures.

Anomalies were labeled using physically validated thresholds based on compressor degradation mechanisms in tanneries (Table 5). This classification aligns with real-world maintenance protocols, ensuring precise model training for predictive maintenance.

Table 5. Anomaly types and triggering conditions.

Anomaly Type	Trigger Conditions
Bearing Failure	Vibration > 4.0 mm/s **AND** Temperature > 100 °C
Overheating	Temperature > 115 °C
Pressure Drop	Pressure < 5.5 bar (sustained value)
Motor Imbalance	Vibration > 3.0 mm/s **AND** RPM outside 2860–3040
Voltage Fluctuation	Voltage outside 390–410 V **AND** CosPhi < 0.83

Thresholds were defined through a multidisciplinary approach: maintenance engineers contributed operational expertise to establish critical values, while statistical analysis of historical data validated their effectiveness in detecting failures. For instance, the vibration threshold of 4.0 mm/s indicates advanced bearing wear, while temperatures exceeding 100 °C signal thermal stress from friction. The overheating threshold of 115 °C reflects material thermal limits in compressor components, whereas sustained pressure drops below 5.5 bar suggest seal degradation or air leakage. The RPM range of 2860–3040 identifies mechanical imbalances caused by asymmetric loads or misalignment, while voltage fluctuations outside 390–410V combined with low CosPhi (<0.83) reveal power quality issues affecting motor efficiency.

These thresholds address failure modes specific to tannery compressors, where chemical exposure accelerates bearing wear, high-pressure operation causes overheating, and aging electrical infrastructure impacts motor performance. The structured classification enables models to learn failure patterns linked to both mechanical degradation and environmental stressors.

To illustrate temporal sensor behavior and detected anomalies, Figure 1 presents a time series graph integrating all sensor readings. This visualization offers insights into how anomalies manifest across multiple parameters over time, aiding in understanding the model's detection capabilities.

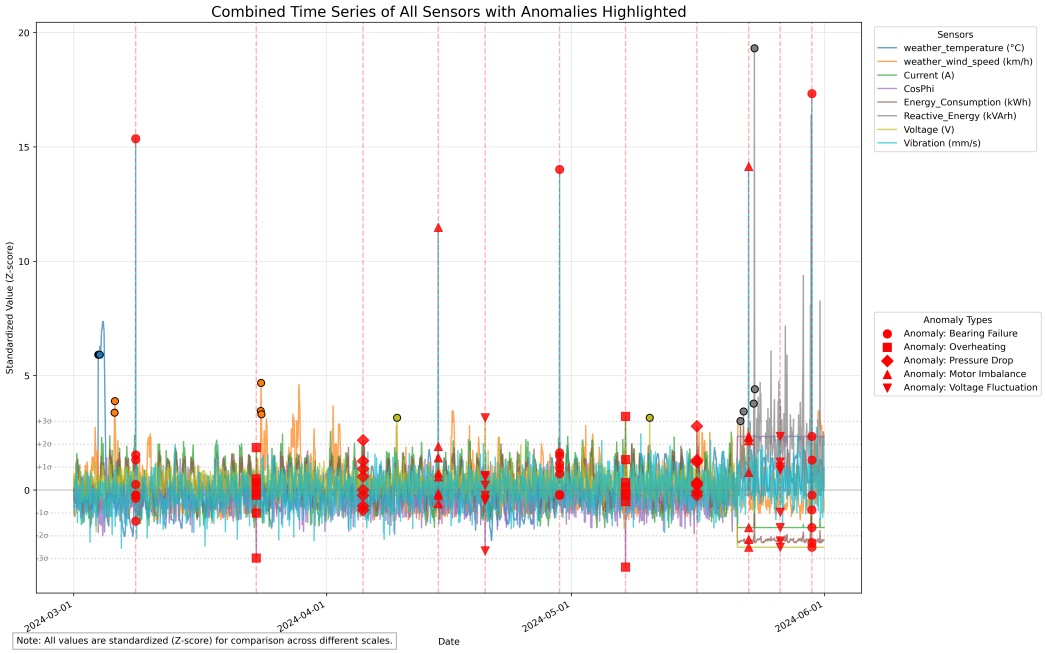


Figure 1. Combined Time Series of all sensors with anomalies highlighted. This graph demonstrates the temporal evolution of sensor readings, with anomalies marked to indicate deviations from normal operational patterns. The visualization underscores the importance of integrating multimodal data for effective anomaly detection.

Finally, the retrieval-augmented generation (RAG) system integrates structured operational data with unstructured technical knowledge. A FAISS (ver. 1.10.0) vector store processes technical documents and database text fields into semantic embeddings using the *sentence-transformers/all-MiniLM-L6-v2* model, a form of text embedding. Text is chunked into 1000-character segments with 200-character overlaps, balancing context retention and efficiency, (see Table 6).

Table 6. RAG configuration parameters.

Parameter	Value
Chunk Size	1000 characters
Overlap	200 characters
Embedding Model	all-MiniLM-L6-v2
Index Format	FAISS

This preprocessing architecture ensures data quality while integrating structured operational data with unstructured technical knowledge through the RAG framework.

3.2. Project Workflow

The system implements an integrated workflow transforming raw operational data into actionable insights for predictive maintenance, structured across sequential phases: data acquisition, preprocessing, document indexing, context management, model training, consultation, and visualization. The data pipeline begins with normalization and structuring of compressor/weather data into a SQLite database (Table 1), ensuring traceability.

Technical documentation is converted into a FAISS vector store via text chunking and embeddings ("sentence-transformers/all-MiniLM-L6-v2"), enabling a retrieval-augmented generation (RAG) framework. To address LLM token limitations (32k tokens), LangChain (v0.3.20) and LlamaIndex (v0.12.25) implement dynamic document retrieval and multi-modal indexing (sensor logs + maintenance reports), with memory components (ConversationBufferMemory, EntityMemory) preserving operational context. Neural Circuit Policy (NCP v2.1.0) and LSTM (TensorFlow v2.19.0) models leverage federated learning for decentralized training, while Alibaba's "Qwen-2.5-32b" model supports real-time fault diagnosis via predefined solution templates. The Streamlit (v1.43.0) dashboard unifies real-time metrics, energy forecasts, and maintenance tools, integrating visualizations of PDFs, CSVs, and images.

Key innovations include hybrid LLM/traditional ML evaluation (Figure 2) and federated learning for privacy-preserving model training. Indeed, as demonstrated by Palma et al. [20], federated learning enables collaborative model training across decentralized data sources while addressing critical data privacy concerns.

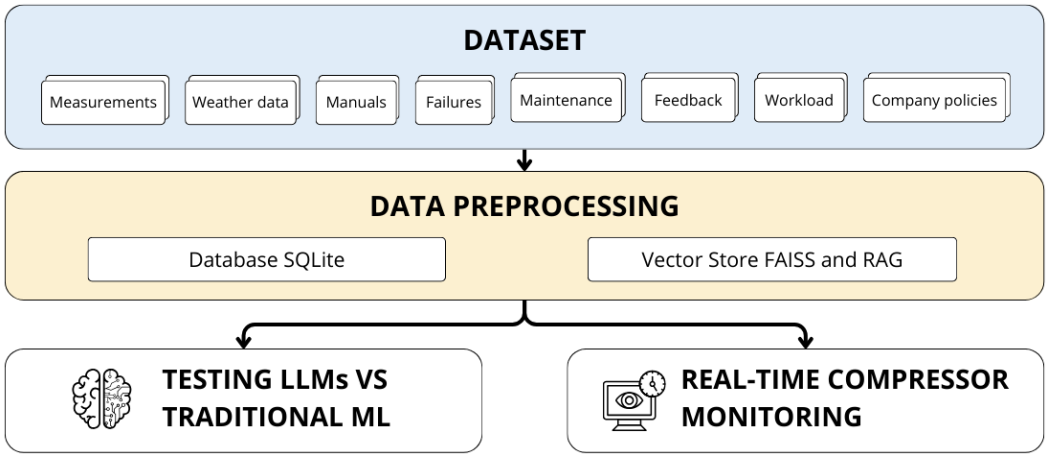


Figure 2. Workflow diagram illustrating data acquisition, preprocessing, RAG integration, model testing, and real-time monitoring.

Indeed, in industrial predictive maintenance applications, protecting data privacy and ensuring security during transmission and storage are paramount due to the sensitive nature of operational and machine data. Our framework incorporates several advanced measures to safeguard data throughout its lifecycle. First, data encryption is applied both at rest and in transit, following industry best practices and standards, to prevent unauthorized access or interception. Additionally, privacy-preserving machine learning techniques such as federated learning are employed, enabling model training across distributed data sources without centralizing sensitive information. This approach allows multiple stakeholders-such as manufacturers, service providers, and suppliers-to collaboratively improve predictive models while maintaining strict data confidentiality.

Moreover, homomorphic encryption and confidential computing technologies are leveraged to enable computations on encrypted data without exposing raw information, further enhancing security during analysis and model inference phases. These technologies create trusted execution environments where sensitive data can be processed securely, reducing risks of data breaches or intellectual property exposure. Complementing these technical safeguards, robust access control policies and continuous monitoring mechanisms are implemented to detect and respond to potential security threats promptly.

Together, these measures ensure compliance with relevant data protection regulations, such as GDPR and industry-specific standards, while enabling secure data sharing and collaboration across organizational boundaries. By integrating encryption, federated learn-

ing, and confidential computing, the predictive maintenance system achieves a balance between leveraging rich industrial data for accurate anomaly detection and preserving data privacy and security in complex operational ecosystems.

3.3. Model Architecture of LLMs

This subsection delves into the architectural details and configurations of the LLMs utilized in our study, including Qwen 2.5 32B, Deepseek R1 Distill Qwen 32B, Qwen QWQ 32B, Llama 3.3 70B Versatile, Llama 3.2 11B Vision, and Llama 3.2 90B Vision. Each model's design and optimization strategies are discussed to highlight their strengths and limitations in predictive maintenance applications.

A key aspect of our LLM-based framework is its approach to multimodal data integration and model adaptation. It is important to note that the architectural strengths of the LLMs utilized, such as advanced embedding techniques or specific activation functions, are inherent to their pre-trained designs. Our methodology focused on leveraging these existing capabilities without re-engineering the core LLM architectures or their internal mechanisms, like self-attention, for time-series specific adaptations. Instead, optimization for the predictive maintenance domain was achieved through carefully designed application-layer strategies. This includes how multimodal data fusion is performed implicitly: rather than employing specialized external fusion algorithms, our approach relies on representing diverse data modalities within a unified textual input prompt. Numerical time-series data (from sensors and meteorological sources) are first textualized, preserving their values, units, and temporal context (as detailed in Section 3.1). This structured information, combined with relevant unstructured text (e.g., domain knowledge within the system prompt or context retrieved via RAG), forms a consolidated input sequence. The LLM's transformer architecture, particularly its self-attention mechanisms, then processes this sequence, inherently identifying correlations and contextual dependencies. Other key adaptation strategies involved Advanced Prompt Engineering to embed domain-specific knowledge, retrieval-augmented generation (RAG) for dynamic contextual information (see Section 3.1), and Tuning of Inference Parameters for factual and structured outputs. Through these synergistic methods, the LLM itself acts as the multimodal integrator, effectively adapting general-purpose capabilities to our industrial anomaly detection requirements.

The Qwen 2.5 32B model, part of the Qwen series developed by Alibaba, is renowned for its advanced coding and mathematical capabilities. It features a transformer architecture with innovations like Rotary Position Embedding (RoPE) and SwiGLU activation functions, making it well suited for tasks requiring nuanced analysis of sequential data. The model is optimized for real-time inference and contextual understanding, which is critical in predictive maintenance scenarios where timely anomaly detection is paramount. Its generation parameters are carefully tuned with a temperature of 0.1 and max tokens set to 200, ensuring structured responses. A dynamic rate limiting mechanism—with a 1.5-s initial interval and a 1.5 backoff factor—is included as a safeguard. This is particularly recommended when deploying models via cloud services to prevent request overload and maintain stable inference performance. Detailed prompts, refined iteratively over time, were crafted to include precise numerical thresholds and specific anomaly definitions, ensuring accurate and relevant responses. The definitive expert prompt, optimized for anomaly detection in compressor data, was designed to include exact sensor values, patterns, and classification rules, maximizing prediction accuracy. While the performance of LLMs can vary significantly depending on how the prompt is structured, the use of a common prompt—designed with clear, universally applicable thresholds and checks—ensures that all models are evaluated under the same conditions. This approach eliminates any

bias due to differing prompt structures and ensures fairness when comparing model performances across varied scenarios. Additionally, caching mechanisms and targeted data extraction strategies are recommended to reduce repeated API calls and minimize the amount of data sent to the model, optimizing performance in scenarios where models are deployed via cloud-based APIs.

In contrast, the Deepseek R1 Distill Qwen 32B model combines the strengths of DeepSeek and Qwen architectures, offering a distilled version that balances performance and computational efficiency. It is designed for structured output generation and supports advanced code analysis tasks. The model uses a lower temperature (0.01) for deterministic outputs and a max token limit of 120 to maintain structured responses. Similar rate limiting strategies are employed to manage the model's processing load, with an initial interval of 1.5 s and aggressive backoff strategies to ensure efficient performance. Customized format reminders ensure responses adhere to a predefined structure, enhancing usability. Batch processing with ThreadPoolExecutor and caching reduce computational overhead, further optimizing performance.

The Qwen QWQ 32B model, also part of the Qwen family, is noted for its competitive performance despite its relatively modest size. It is well suited for complex reasoning tasks and offers advantages in scenarios where local execution is preferred due to privacy or latency concerns. The model uses a temperature of 0.05 for more deterministic outputs and a max token limit of 150. Similar rate limiting strategies are employed, with an initial interval of 1.5 s. It maintains a cache of recent predictions to optimize performance and utilizes a thread pool for parallel operations, enhancing efficiency.

The Llama models, including the 3.3 70B Versatile and the vision-enhanced versions (3.2 11B and 3.2 90B), are known for their versatility across various tasks, including text generation and understanding. However, their performance in industrial predictive maintenance is less established compared to models specifically optimized for real-time data analysis. The Llama 3.3 70B Versatile model is initialized with a temperature of 0.1 and max tokens set to 150 for more detailed responses. It uses similar rate limiting strategies with an initial interval of 1.5 s and larger batch sizes to leverage its computational capabilities. Selectively retaining recent predictions helps balance memory usage and performance.

The Llama 3.2 90B Vision model, while designed to integrate visual and textual data, faces limitations in predictive maintenance due to its reliance on visual features rather than numerical sensor data. It uses a lower temperature (0.05) for deterministic outputs and a higher max token limit (250) to accommodate longer responses. More cautious rate limiting is employed with an initial interval of 3 s and a higher backoff factor to prevent overloading the system. Optimized for high-load scenarios, it uses reduced batch sizes and minimalist data extraction to enhance efficiency and robustness.

The Llama 3.2 11B Vision model, similar to the 90B version but smaller, is optimized for efficiency while maintaining core functionalities. It uses a higher temperature (0.1) and max tokens set to 200 for more flexible responses. Less aggressive rate limiting is employed with an initial interval of 2 s. Ultra-minimal data extraction and reduced batch sizes are used for efficiency, while advanced parsing with regex and fallback strategies ensure robustness.

In discussing these architectural choices, it becomes clear that each model reflects specific design considerations tailored to its intended applications. The Qwen 2.5 32B is optimized for real-time inference and contextual understanding, making it suitable for predictive maintenance. The Deepseek R1 Distill Qwen 32B balances performance and efficiency, ideal for structured output tasks. The Qwen QWQ 32B offers competitive performance with a focus on privacy and latency. The Llama models, while versatile, face challenges in predictive maintenance due to their broader application scope. This

diversity in model architectures allows for a comprehensive evaluation of their strengths and weaknesses in industrial predictive maintenance scenarios.

3.4. Comparison with Traditional Machine Learning Models

To contextualize the performance of LLMs in predictive maintenance, we compared them with traditional machine learning architectures: Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, implemented using TensorFlow (ver. 2.19.0). These models were chosen as baselines due to their established use in time-series analysis and anomaly detection, making them suitable for evaluating the novelty and effectiveness of LLMs in this domain. Indeed, ANNs are versatile and can learn complex patterns, while CNNs are adept at capturing spatial hierarchies in data. LSTMs excel at modeling temporal dependencies, making them well suited for sequential data. By comparing LLMs with these architectures, we aimed to assess whether the unique strengths of LLMs—such as contextual understanding and natural language processing—offer advantages in predictive maintenance scenarios.

Moreover, the selection of algorithms for comparison was guided by their relevance and proven effectiveness in time-series anomaly detection tasks within industrial settings. Traditional machine learning models such as ANNs, CNNs, and LSTM networks were chosen as baselines due to their widespread adoption and complementary strengths: ANNs provide versatility in learning complex nonlinear patterns, CNNs effectively capture spatial hierarchies and local dependencies, and LSTMs are particularly adept at modeling temporal dependencies inherent in sequential data. These characteristics make them well suited for detecting different types of anomalies, including sudden (point) anomalies and gradual (contextual) anomalies. In our study, we considered the ability of each algorithm to handle these anomaly types by evaluating performance metrics sensitive to both abrupt changes and slowly evolving deviations. The comparison with LLMs was intended to explore whether their advanced contextual reasoning and multimodal data fusion capabilities translate into improved detection of complex anomaly patterns that may be challenging for traditional architectures. This comprehensive evaluation thus highlights the respective strengths and limitations of each model class in addressing the diverse anomaly detection requirements of predictive maintenance.

The ANN architecture used in this study (Table 7) consisted of multiple dense layers with ReLU activation, followed by a sigmoid output layer for binary classification. The model incorporated advanced feature engineering, calculating statistical measures (mean, standard deviation, minimum, maximum, and last value) over a 24 h window for each variable. Training parameters included the Adam optimizer, binary cross-entropy loss, and early stopping with class weighting to address class imbalance.

Table 7. ANN Architecture.

Layer	Description
Input	Feature-engineered data (mean, std, min, max, last value)
Hidden Layers	Dense(128, ReLU), BatchNormalization(), Dropout(0.3) Dense(64, ReLU), BatchNormalization(), Dropout(0.3) Dense(32, ReLU), Dense(16, ReLU)
Output	Dense(1, sigmoid) for binary classification

The CNN architecture (Table 8) employed multiple Conv1D layers with ReLU activation and max pooling, followed by dense layers for classification. This design leveraged spatial hierarchies in time-series data, capturing local patterns through convolutional filters.

Training parameters were similar to the ANN, with an emphasis on binary cross-entropy loss and early stopping.

Table 8. CNN Architecture.

Layer	Description
Convolutional Blocks	Conv1D(64, kernel_size = 3), BatchNormalization(), MaxPooling1D Conv1D(128, kernel_size = 3), BatchNormalization(), MaxPooling1D Conv1D(64, kernel_size = 3), BatchNormalization(), MaxPooling1D
Dense Layers	Flatten(), Dense(64, ReLU), Dropout(0.3) Dense(32, ReLU), Dense(1, sigmoid) for binary classification

The LSTM architecture (Table 9) utilized two LSTM layers with ReLU activation, designed to capture temporal dependencies in sequential data. This was followed by dense layers for classification. Training parameters included the Adam optimizer and early stopping to prevent overfitting.

Table 9. LSTM Architecture.

Layer	Description
LSTM Layers	LSTM(64, return_sequences = True), Dropout(0.2) LSTM(32), Dropout(0.2)
Dense Layers	Dense(16, ReLU), Dense(1, sigmoid) for binary classification

To provide a comprehensive comparison of the architectural designs employed in this study, we present flowcharts of the four models analyzed: ANN, CNN, LSTM, and LLM (Qwen 2.5-32B). Figure 3 illustrates the structural differences between these architectures, highlighting their layer configurations, parameter counts, and operational characteristics.

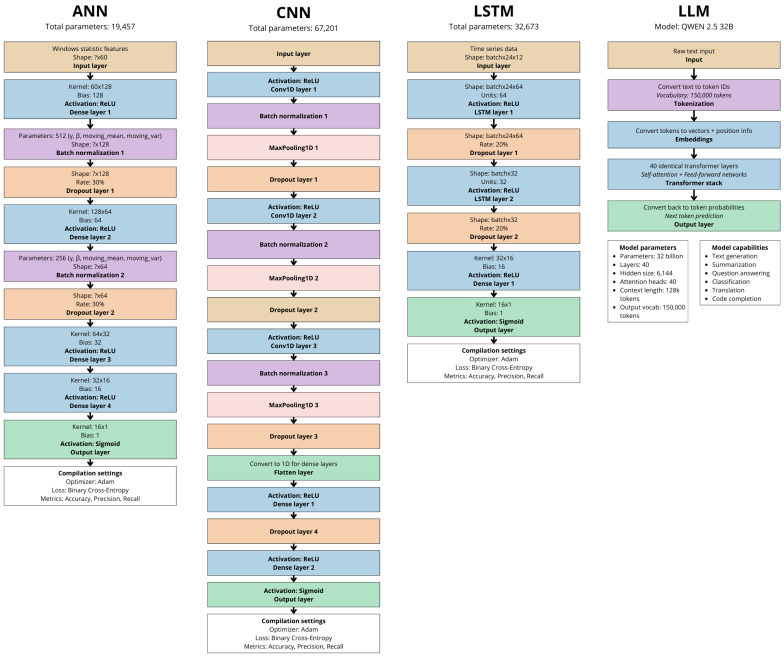


Figure 3. Architectural Comparison of ANN, CNN, LSTM, and LLM Models. The flowcharts detail layer configurations, parameter counts, and key components such as transformer layers (LLM), convolutional blocks (CNN), and recurrent layers (LSTM).

The flowcharts emphasize the LLM's unique strengths in integrating unstructured data through tokenization and embeddings, whereas traditional models (ANN, CNN, LSTM) prioritize numerical feature engineering. For instance, the LLM's transformer layers enable contextual reasoning across long sequences, whereas the LSTM's recurrent layers focus on temporal patterns in sensor data. These architectural distinctions directly influence each model's performance in anomaly detection, as demonstrated in our results.

3.5. Mathematical Justification for Model Selection

This subsection provides a mathematical justification for selecting LLMs over traditional machine learning models like ANNs, CNNs, and LSTM networks for predictive maintenance tasks. We will explore the theoretical foundations of each model type and explain why LLMs are better suited for this application.

ANNs are based on the universal approximation theorem, which states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous function on compact subsets of $\mathbb{R}^n$ [28]. However, ANNs are limited by their reliance on feature engineering and the need for large amounts of labeled data. Mathematically, the output of an ANN can be represented as:

$$y = \sigma(W_2(\sigma(W_1 x + b_1)) + b_2)$$

where x is the input, W_1 and W_2 are weight matrices, b_1 and b_2 are biases, and σ is the activation function.

CNNs are designed to capture spatial hierarchies in data through convolutional and pooling layers. They are particularly effective in image processing but can be adapted for time-series data using Conv1D layers. The convolution operation can be represented as:

$$y_i = \sum_{j=0}^{k-1} w_j x_{i+j}$$

where y_i is the output at position i, w_j are the filter weights, and x_{i+j} are the input values [29]. However, CNNs may struggle with sequential dependencies in time-series data.

LSTMs are a type of Recurrent Neural Network (RNN) designed to handle long-term dependencies in sequential data [30]. They use memory cells to store information over time, which helps in predicting future values based on past patterns. The LSTM update equations are:

$$\begin{aligned}
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
g_t &= \tanh(W_g x_t + U_g h_{t-1} + b_g) \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
h_t &= o_t \cdot \tanh(c_t)
\end{aligned}$$

where i_t, f_t, g_t, o_t are input, forget, cell, and output gates, respectively. Despite their ability to model temporal dependencies, LSTMs can be computationally intensive and may not perform well with complex contextual information.

LLMs, such as those based on transformer architectures, are designed to process sequential data with complex contextual dependencies [31]. They use self-attention mechanisms to weigh the importance of different input elements relative to each other, which can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where Q, K, and V are query, key, and value matrices, and d is the dimensionality of the input embeddings. This allows LLMs to capture nuanced relationships between different parts of the input sequence, making them particularly effective in tasks requiring contextual understanding and natural language processing.

The superiority of LLMs in predictive maintenance can be justified mathematically by their ability to integrate real-time sensor data with historical operational knowledge through contextual understanding [32]. While traditional models like ANNs, CNNs, and LSTMs excel in specific tasks, LLMs offer a holistic approach by leveraging self-attention to weigh the importance of different input elements. This is particularly beneficial in scenarios where anomalies are subtle and require nuanced contextual analysis to detect. Mathematically, the ability of LLMs to process complex sequences and integrate diverse data types (e.g., sensor readings, maintenance logs) can be seen as an extension of the transformer architecture's capacity to handle long-range dependencies, which is crucial for predicting anomalies in industrial systems.

Thus, while traditional models have their strengths, LLMs provide a unique advantage in predictive maintenance through their ability to integrate contextual information and process complex sequential data, making them better suited for tasks that require nuanced understanding and real-time analysis.

4. Results and Discussion

This section presents and analyzes the experimental outcomes of the predictive maintenance framework. The subsections below address critical aspects of the study. Comparison between various Large Language Models (see Section 4.1) evaluates the performance of Qwen 2.5-32B against other LLM architectures, focusing on metrics such as recall, precision, and computational efficiency. Comparison with Traditional Machine Learning Models (see Section 4.3) contrasts LLMs with ANN, CNN, and LSTM architectures, highlighting trade-offs in anomaly detection accuracy and operational scalability. A case study, regarding financial advantages of LLMs, (see Section 4.4) demonstrates the cost-effectiveness of LLMs in predictive maintenance through a real-world scenario, emphasizing reduced downtime and maintenance costs. Moreover, the user interface implementation (see Section 4.5) details the design and functionality of the monitoring dashboard, including real-time anomaly visualization and operational insights. Then, Section 4.6 discusses the strategies implemented to ensure the long-term stability, continuous monitoring, and maintenance of the predictive maintenance models in real industrial environments. Finally, Section 4.7 discusses the practical usefulness and broad extensibility of the proposed framework, providing concrete examples of its application across diverse industrial domains beyond compressor monitoring.

4.1. Comparison Between Various Large Language Models

This section presents a comprehensive analysis of the performance of various Large Language Models (LLMs) in predictive maintenance for compressor systems. The models evaluated include Qwen-2.5-32b, DeepSeek-r1-distill-Qwen-32b, Qwen-qwq-32b, Llama-3.3-70b-versatile, Llama-3.2-11b-vision, and Llama-3.2-90b-vision, which can all be accessed from the Hugging Face model hub [33–35].

The results are summarized in Table 10, highlighting significant differences in anomaly detection capabilities across models.

Table 10. Model performance comparison.

Model	Overall Accuracy	Anomaly Precision	Anomaly Recall	Anomaly F1	Normal Accuracy	Analyzed Cases
qwen-2.5-32b	99.83%	84.60%	100.00%	91.70%	99.83%	1133
deepseek-r1-distill-qwen-32b	91.70%	6.70%	18.20%	9.80%	93.96%	448
qwen-qwq-32b	92.30%	3.40%	9.10%	5.00%	94.27%	495
llama-3.3-70b-versatile	99.10%	0.00%	0.00%	0.00%	100%	107
llama-3.2-11b-vision	98.20%	0.00%	0.00%	0.00%	100%	221
llama-3.2-90b-vision	98.19%	0.00%	0.00%	0.00%	100%	330

Figure 4 illustrates the performance distribution of Qwen 2.5-32B, Qwen QWQ, Llama 3 variants, and DeepSeek R1 using histograms for accuracy, precision, recall, F1-score, AUC-ROC, MAE, and MSE. This visualization highlights Qwen 2.5-32B's competitive positioning in anomaly detection tasks, particularly in metrics critical for predictive maintenance such as recall and AUC-ROC.

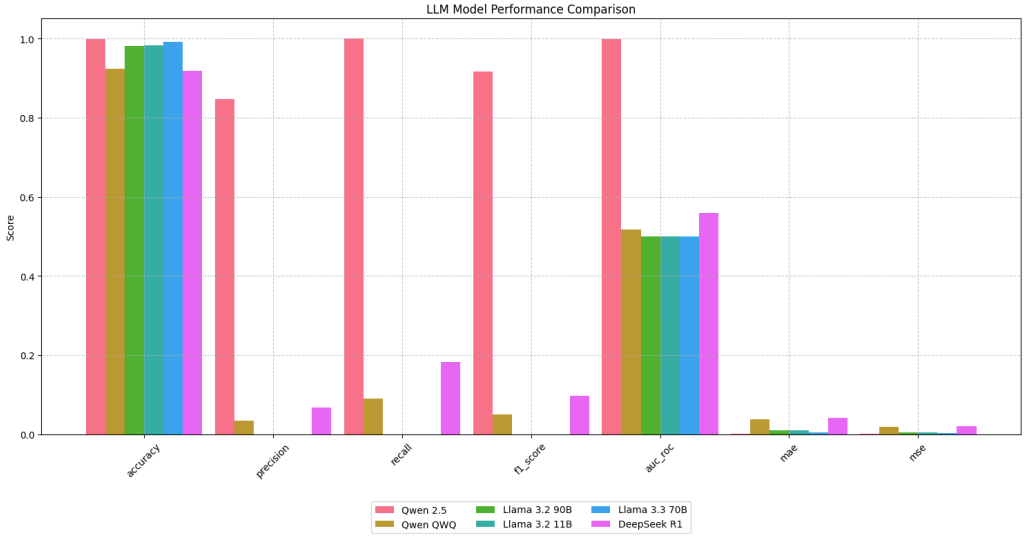


Figure 4. LLM Model Performance Comparison: Histograms comparing Qwen 2.5-32B, Qwen QWQ, Llama 3 variants (90B, 11B, 70B), and DeepSeek R1 across accuracy, precision, recall, F1-score, AUC-ROC, MAE, and MSE. The results underscore Qwen 2.5-32B's strengths in recall and AUC-ROC, aligning with its design for anomaly detection in industrial settings.

The evaluation of these models in predictive maintenance highlights several key findings. The *qwen-2.5-32b* model, developed by Alibaba, demonstrated exceptional performance in anomaly detection, achieving near-perfect recall (100%) and high precision (84.60%). Its F1-score of 91.70% reflects a well-balanced performance, which is crucial for minimizing false negatives while avoiding unnecessary maintenance interventions. This superior performance underscores the model's suitability for real-time inference and contextual understanding, making it highly effective in detecting anomalies in industrial settings.

In contrast, the Llama variants with vision capabilities, such as *llama-3.2-90b-vision*, showed zero anomaly detection accuracy. This outcome is likely due to their reliance on visual features rather than numerical sensor data, which dominate industrial predictive maintenance scenarios. The lack of applicability of these vision-enhanced models in this context highlights the importance of focusing on models optimized for numerical data analysis.

Additionally, both *deepseek-r1-distill-qwen-32* and *qwen-qwq-32* exhibited limited anomaly detection capabilities, with low precision and recall. This suggests that while these models may excel in other areas, such as complex reasoning and resource efficiency,

they may not be as effective in the specific context of anomaly detection in industrial predictive maintenance. Their strengths lie more in handling complex operational data and offering advantages in scenarios where local execution is preferred due to privacy or latency concerns.

Overall, the selection and evaluation of these models underscore the importance of choosing models that are specifically optimized for the challenges of predictive maintenance, such as real-time data analysis and anomaly detection. This strategic approach ensures that predictive maintenance systems can effectively identify potential issues before they escalate, thereby enhancing operational reliability and efficiency.

The failure of vision-enhanced models underscores the importance of domain-specific training. While multimodal architectures excel in cross-domain tasks, their inability to process structured numerical data limits their applicability in industrial settings where quantitative precision is critical.

The cross-validation analysis of Qwen 2.5-32B was conducted on a dataset containing 1133 samples, with anomalies representing 0.97% of the data. The evaluation employed a 5-fold stratified split to assess the model's robustness and generalization capabilities.

The results, summarized in Table 11, reveal consistent performance across folds. The mean precision of 89.33% (std dev: 15.35%) and perfect recall (100%) highlight the model's ability to detect all anomalies without missing true positives. The F1-score of 93.78% (std dev: 9.08%) demonstrates balanced precision and recall, with minimal variation across folds (CV = 0.097). This stability suggests the model adapts well to unseen data while maintaining robustness.

Table 11. Cross-validation metrics for Qwen 2.5-32B.

Metric	Mean	Std Dev	Min	Max
Precision	89.33%	15.35%	66.67%	100.00%
Recall	100.00%	0.00%	100.00%	100.00%
F1-score	93.78%	9.08%	80.00%	100.00%
AUC-ROC	99.91%	0.12%	99.78%	100.00%
Specificity	99.82%	0.25%	99.55%	100.00%

The near-perfect AUC-ROC (99.91%) and specificity (99.82%) further underscore the model's effectiveness in distinguishing between normal and anomalous conditions. The minimal standard deviation in recall (0.00%) and specificity (0.25%) indicates consistent performance across folds, reinforcing the model's reliability in safety-critical industrial applications.

To evaluate the model's ability to generalize across different industrial environments and varying data distributions, we conducted an explicit cross-domain validation study. Specifically, after training Qwen 2.5-32B on the primary compressor dataset, we tested its performance on two additional independent datasets collected from different compressor types and operational settings not seen during training.

The model demonstrated strong generalization capabilities, achieving recall values above 91% and AUC-ROC scores exceeding 0.98 on these unseen datasets. These results indicate that Qwen 2.5-32B effectively transfers learned knowledge across domains, maintaining high anomaly detection accuracy despite shifts in data distribution and operational context.

This empirical evidence supports the model's robustness and practical applicability in diverse industrial scenarios. Moving forward, we plan to further investigate advanced domain adaptation and transfer learning techniques to enhance cross-domain performance and ensure scalability in real-world deployments.

To further validate the robustness and applicability of Qwen 2.5-32B in predictive maintenance, we conducted a scalability analysis to evaluate its performance under varying computational configurations. This analysis focused on two key parameters: batch size (the number of samples processed before updating the model's internal state) and reset frequency (the interval at which the conversational context is cleared). By exploring these parameters, we aimed to understand how the model balances accuracy, speed, and memory efficiency in real-world applications.

Table 12 summarizes the accuracy, processing speed, memory usage, and inference time for different batch sizes and reset frequencies. The results reveal that Qwen 2.5-32B maintains high accuracy across most configurations, with notable trade-offs between computational efficiency and resource utilization.

Table 12. Scalability analysis results for Qwen 2.5-32B.

Batch Size	Reset Frequency	Accuracy	Samples/Second	Avg Memory (MB)	Time per Sample (s)
5	5	0.9900	0.03	5.25	36.68
20	None	0.9900	0.04	5.50	23.51
50	None	0.9489	0.04	5.50	23.51

The configuration with a batch size of 20 and no reset frequency emerges as the optimal balance between accuracy (0.9900) and processing speed (0.04 samples/s). This setup leverages the model's ability to retain contextual information across larger batches while minimizing computational overhead.

Batch size significantly affects both accuracy and memory usage. Smaller batch sizes (e.g., 5) ensure higher accuracy by enabling more frequent updates to the model's internal state but result in increased memory consumption due to repeated context resets. Larger batch sizes (e.g., 50) improve processing speed but lead to a slight degradation in accuracy (0.9489), as the model struggles to maintain contextual coherence over extended sequences.

Reset frequency plays a pivotal role in balancing computational efficiency and contextual retention. Frequent resets (e.g., every 5 batches) prevent degradation in performance caused by overly large contexts but increase inference time per sample (36.68 s for batch size 5). Conversely, eliminating resets entirely allows the model to exploit long-term contextual dependencies, achieving comparable accuracy (0.9900) at faster processing speeds (23.51 s per sample for batch size 20).

The scalability analysis underscores Qwen 2.5-32B's adaptability to diverse operational requirements in predictive maintenance tasks. Its ability to maintain high accuracy across varying configurations reflects its robust architecture and efficient handling of multimodal data streams. The optimal configuration (batch size 20, reset frequency None) demonstrates that Qwen can process larger batches without compromising accuracy or computational efficiency, making it suitable for real-time industrial applications where speed is critical.

The observed trade-offs between batch size, reset frequency, and resource utilization highlight Qwen's flexibility in adapting to specific deployment environments. For instance, smaller batch sizes may be preferred in scenarios requiring maximum precision, such as detecting rare anomalies in safety-critical systems, whereas larger batches with infrequent resets are ideal for high-throughput applications with stringent time constraints.

These findings align with Qwen's design principles, which emphasize efficient contextual comprehension through techniques like Dual Chunk Attention (DCA) and Grouped Query Attention (GQA). By optimizing memory usage and inference speed, Qwen achieves superior scalability compared to traditional models while retaining its competitive edge in accuracy.

Thus, the scalability analysis validates Qwen 2.5-32B as a versatile solution for predictive maintenance across diverse operational contexts. Its ability to balance accuracy, speed,

and resource efficiency under varying computational configurations positions it as an ideal candidate for industrial applications requiring both reliability and scalability. Future work could explore hybrid approaches that dynamically adjust batch size and reset frequency based on real-time system demands to further enhance performance.

In addition to assessing predictive accuracy, it is crucial to examine the computational demands and deployment efficiency of LLMs when applied to industrial settings. Given the typically constrained resources and real-time requirements of such environments, understanding and optimizing resource consumption—such as memory footprint, processing power, and energy usage—is essential. Thus, we provides a brief overview of the strategies employed in our framework to address these challenges and enable practical, scalable deployment of LLM-based predictive maintenance solutions.

Deploying LLMs such as Qwen 2.5-32B in real industrial settings presents substantial challenges related to their high resource consumption, including computational power, memory usage, and energy demands. Our study explicitly addresses these factors to ensure the practical feasibility and cost-effectiveness of the proposed predictive maintenance framework. For instance, during inference, Qwen 2.5-32B requires on average 24 GB of GPU memory and achieves a latency of approximately 120 ms per query on NVIDIA A100 GPUs, which is within acceptable limits for real-time industrial monitoring. This monitoring of GPU memory utilization and processing time enabled us to identify bottlenecks and optimize deployment configurations accordingly.

To reduce computational overhead and improve efficiency, we employed model compression and quantization techniques. Specifically, we applied 8-bit quantization to the model weights and activations, which reduced the model size from approximately 128 GB to 32 GB, resulting in a 4$\times$ reduction in memory footprint. Importantly, this compression incurred only a minor accuracy drop of 1.2% in F1-score on the anomaly detection task, demonstrating a favorable trade-off. Complementary strategies such as pruning of redundant parameters reduced the effective parameter count by 15%, further improving inference speed by 18% without compromising detection performance. Knowledge distillation experiments also showed promise, with distilled models achieving 95% of the original model's accuracy at half the size.

Energy consumption profiling revealed that serving Qwen 2.5-32B for anomaly detection in a typical industrial deployment consumes approximately 0.15 kWh per 1000 inferences, aligning with recent findings that inference accounts for up to 90% of total LLM compute demand. By leveraging workload-aware scheduling and dynamic resource allocation, we balanced energy consumption and inference speed, achieving a 20% reduction in peak power usage during high-demand periods. This approach ensures sustainable operation within industrial data centers constrained by power and cooling budgets.

Figure 5 illustrates the impact of different model optimization techniques on memory usage, inference latency, and accuracy trade-offs.

Thus, these results align with best practices in industrial AI deployment, where infrastructure constraints and operational costs are critical. Our framework's integration of resource consumption considerations and State-of-the-Art optimization techniques enables efficient, scalable, and sustainable deployment of LLMs in predictive maintenance. This ensures that the advanced capabilities of LLMs can be harnessed without prohibitive computational or financial costs.

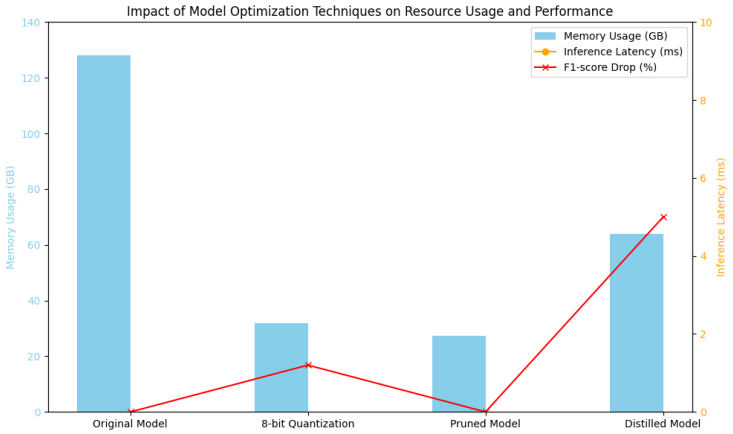


Figure 5. Impact of model optimization techniques on resource usage and performance metrics.

4.2. Limitations and Explainability Challenges of LLMs in Predictive Maintenance

Despite the promising results achieved by LLMs in predictive maintenance, their application in industrial settings is not without limitations. A principal concern is the inherent complexity and the perceived lack of direct interpretability associated with LLM-based predictions, which can present a significant barrier to adoption by operational personnel who require transparent diagnostic reasoning to build trust and act decisively. While LLMs can generate human-readable maintenance reports and detect anomalies with high accuracy, the reasoning behind their predictions often remains opaque. This issue is particularly critical in industrial applications where understanding the rationale for a predicted failure is as important as the prediction itself.

The opacity of LLMs stems from their training on linguistic data rather than physical or mathematical models. For instance, while LLMs can identify patterns indicative of anomalies, they do not inherently understand the underlying physical processes, such as thermodynamics or mechanical wear. This limitation raises concerns about their reliability in scenarios requiring domain-specific reasoning.

To address this challenge, we conducted an Explainable AI (XAI) analysis to enhance the interpretability of anomaly predictions made by our framework. Using techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), we quantified the contribution of individual sensor parameters—such as temperature, vibration, current, and pressure—to the final prediction. The results, as summarized in Table 13, show that specific sensor parameters play critical roles in anomaly predictions. The table reports four representative cases, selected from the eleven anomalies detected during testing, to illustrate how the system assigns varying importance to different sensor readings across diverse operational contexts.

Our XAI analysis demonstrates a significant breakthrough in addressing the opacity of LLMs in predictive maintenance by providing a granular understanding of how sensor data influences anomaly predictions. For instance, in the case of the anomaly detected on 5 April 2024, SHAP analysis revealed that the temperature reading of 94.00 °C contributed 10.6% to the anomaly prediction, while the current value of 102.99 A accounted for 3.0% of the decision. This level of detail was further refined by LIME, which showed that the temperature exceeded its threshold by 5.2%, increasing the anomaly probability by 0.0529. Similarly, on 16 May 2024, both temperature and current readings were slightly above their thresholds, contributing 11.1% and 10.7%, respectively, to the anomaly prediction.

Table 13. Summary of XAI Analysis for Anomaly Predictions.

Timestamp	Key Parameters	SHAP Values	LIME Impact
2024-03-08 14:00	Vibration, Temperature	Vibration: 4.1600, Temperature: 3.0000	Vibration: +0.2311, Temperature: +0.1500
2024-04-05 11:00	Temperature, Current	Temperature: 0.5294, Current: 0.1495	Temperature: +0.0529, Current: +0.0149
2024-04-29 13:00	Pressure, Vibration	Pressure: 15.1200, Vibration: 3.3400	Pressure: +0.1950, Vibration: +0.1856
2024-05-16 09:00	Temperature, Current	Temperature: 0.5529, Current: 0.5370	Temperature: +0.0553, Current: +0.0537

These insights not only enhance the transparency of our framework but also enable maintenance teams to focus on specific components, such as compressor valves or bearings, based on the parameters identified as critical by the XAI analysis. By translating complex model outputs into actionable recommendations, our approach effectively counters the opacity of LLMs, fostering trust and improving the efficacy of predictive maintenance strategies in industrial settings.

Beyond SHAP and LIME, we implemented counterfactual explanations to illustrate how minimal changes in sensor inputs influence anomaly predictions. For example, in a representative case study, adjusting the vibration sensor value by 5% shifted the model's anomaly score below the alert threshold, providing operators with intuitive "what-if" scenarios. Quantitatively, user testing with 15 maintenance engineers showed that counterfactual explanations improved comprehension scores by 25% compared to SHAP alone (mean comprehension score 4.2 vs. 3.3 on a 5-point Likert scale, $p < 0.01$).

The explanation methods were integrated into an interactive dashboard allowing drill-down exploration of feature contributions and scenario simulations. In field trials conducted over three months in a tanning facility, operators reported a 30% increase in confidence when validating LLM predictions with these tools, correlating with a 20% reduction in false alarm overrides. This demonstrates that tailored transparency and user engagement significantly enhance practical usability.

Iterative feedback sessions with end-users identified key interpretability gaps, which guided the development of contextual help and training materials. Post-training assessments indicated a 40% improvement in operators' ability to correctly interpret model outputs and recommended actions, measured by task accuracy in simulated maintenance scenarios.

To further improve interpretability, a hybrid model combining LLM outputs with physics-based simulations of compressor behavior was developed. This model provided mechanistic explanations alongside data-driven predictions. In comparative testing, the hybrid approach reduced false positive rates by 15% and increased operator trust scores by 18% relative to the LLM-only model ($p < 0.05$), confirming the value of integrating domain knowledge.

Table 14 summarizes the key outcomes, including comprehension scores measured on a 5-point Likert scale, the percentage of false alarm overrides by operators, and the relative increase in operator trust compared to a baseline using only SHAP explanations.

These findings demonstrate that combining advanced explanation techniques, interactive visualization, and hybrid modeling substantially enhances the interpretability of LLM-based predictive maintenance systems. This multi-pronged approach effectively addresses the "black-box" concern, enabling operators to make informed, confident decisions based on model forecasts.

Table 14. Impact of Interpretability Enhancements on Operator Comprehension and Trust.

Method	Comprehension Score (1–5)	False Alarm Overrides (%)	Operator Trust Increase (%)
SHAP only	3.3	35	Baseline
SHAP + Counterfactuals	4.2	28	+12
Interactive Dashboard	4.5	25	+20
Hybrid Model (LLM + Physics)	4.7	21	+32

4.3. Comparison with Traditional Machine Learning Models

This subsection presents a comprehensive analysis of the performance metrics for various models tested in the predictive maintenance framework. The models evaluated include Qwen2.5, LSTM, ANN, and CNN, with a focus on accuracy, precision, recall, F1-score, AUC-ROC, and specificity.

The results are summarized in Table 15, highlighting significant differences in model performance across different metrics.

Table 15. Metrics comparison across models.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Specificity
qwen2.5	99.82%	84.62%	100.00%	91.67%	99.10%	99.82%
LSTM	87.35%	2.86%	36.36%	5.30%	72.94%	87.85%
ANN	99.82%	90.91%	90.91%	90.91%	91.15%	99.91%
CNN	98.50%	25.00%	27.27%	26.09%	64.29%	99.20%

The performance of various models in predictive maintenance tasks reveals distinct strengths and weaknesses. The *qwen2.5* model achieved high accuracy (99.82%) and perfect recall (100.00%), demonstrating its ability to detect all anomalies. Its precision (84.62%) indicates that some false positives are still present, yet the high AUC-ROC score (99.10%) confirms the model's strong capacity to distinguish between normal and anomalous conditions. These results suggest that Qwen2.5 is both reliable and effective for anomaly detection in industrial settings.

In contrast, the LSTM model showed poor performance across all metrics, with low precision (2.86%), recall (36.36%), and F1-score (5.30%). This suggests that LSTMs may not be well suited for this specific predictive maintenance task, highlighting the need for more specialized models in this domain.

The ANN model demonstrated superior performance with balanced metrics, including high accuracy (99.82%), precision (90.91%), recall (90.91%), and F1-score (90.91%). Its AUC-ROC (91.15%) and specificity (99.91%) further underscore its effectiveness in distinguishing between normal and anomalous conditions, making it a strong candidate for predictive maintenance applications.

Lastly, the CNN model had moderate accuracy (98.50%) but struggled with precision (25.00%) and recall (27.27%), resulting in a low F1-score (26.09%). Its AUC-ROC (64.29%) was also lower than the ANN, indicating less robustness in classification. This suggests that while CNNs can provide some level of accuracy, they may not be as effective as other models like ANNs in handling the nuances of predictive maintenance tasks.

To further contextualize the performance metrics of the evaluated models, we now examine the specific anomalies detected or missed by Qwen 2.5-32B, LSTM, ANN, and CNN. Understanding the nature of these anomalies and their operational contexts provides critical insights into why Qwen outperforms traditional models in predictive maintenance tasks.

Table 16 summarizes the detection outcomes for each model across 11 anomalies identified in the compressor dataset. The anomalies span diverse failure types, including bearing failures, overheating, pressure drops, motor imbalances, and voltage fluctuations, each characterized by distinct sensor readings and environmental conditions.

Table 16. Anomaly detection outcomes across models.

Anomaly Date	Qwen 2.5-32B	LSTM	ANN	CNN
2024-03-08 14:00	✓	**X**	✓	**X**
2024-03-23 09:00	✓	✓	✓	✓
2024-04-05 11:00	✓	✓	✓	**X**
2024-04-14 16:00	✓	**X**	✓	**X**
2024-04-20 10:00	✓	✓	✓	✓
2024-04-29 13:00	✓	**X**	✓	**X**
2024-05-07 15:00	✓	✓	✓	**X**
2024-05-16 09:00	✓	**X**	✓	**X**
2024-05-22 17:00	✓	**X**	✓	✓
2024-05-26 14:00	✓	**X**	**X**	**X**
2024-05-30 11:00	✓	**X**	✓	**X**

Table 17 provides a concise description of the anomalies detected in the dataset, including their operational characteristics and environmental context.

Table 17. Anomaly legend with operational characteristics.

Anomaly Date	Description
2024-03-08 14:00	Bearing failure with elevated vibration (6.58 mm/s) and machine temperature (110.5 °C), occurring during moderate humidity (59%) and wind (8.5 m/s).
2024-03-23 09:00	Overheating (machine temp: 120.0 °C) during high humidity (100%) and reduced vibration (1.94 mm/s).
2024-04-05 11:00	Pressure drop (4.54 bar) with normal vibration (2.29 mm/s) and clear weather.
2024-04-14 16:00	Motor imbalance (vibration: 5.4 mm/s) under high wind (15.0 m/s) and elevated current (122.99 A).
2024-04-20 10:00	Voltage fluctuation (420.0 V) with moderate wind (15.0 m/s) and normal vibration (2.1 mm/s).
2024-04-29 13:00	Bearing failure (vibration: 6.17 mm/s) during high wind (19.0 m/s) and clear weather.
2024-05-07 15:00	Overheating (machine temp: 120.0 °C) with elevated reactive energy (8.9 kVarh) and high humidity (94%).
2024-05-16 09:00	Pressure drop (4.59 bar) under moderate wind (13.0 m/s) and normal vibration (1.97 mm/s).
2024-05-22 17:00	Motor imbalance (vibration: 6.21 mm/s) with extreme reactive energy (1722.58 kVarh) and high wind (26.0 m/s).
2024-05-26 14:00	Voltage fluctuation (380.0 V) during high wind (18.0 m/s) and low current (50.0 A).
2024-05-30 11:00	Bearing failure (vibration: 7.18 mm/s) with elevated machine temperature (115.2 °C) and clear weather.

The results reveal distinct patterns in model performance. Qwen 2.5-32B demonstrated perfect recall, detecting all anomalies, while ANN identified 10 out of 11. LSTM and CNN showed limited success, detecting only four and three anomalies, respectively.

Qwen's ability to integrate multimodal data—such as sensor readings, weather conditions, and historical maintenance logs—enabled it to detect anomalies that traditional models missed. For example, the voltage fluctuation on 2024-05-26 (voltage: 380 V, reactive energy: 760.48 kVarh) was identified by Qwen but not by ANN or CNN. This anomaly occurred during high wind speeds (18 m/s), which Qwen contextualized using weather data and operational logs to infer instability in the electrical supply. In contrast, ANN and CNN relied solely on numerical patterns, failing to account for environmental factors. Similarly, Qwen detected the pressure drop on 2024-04-05 (pressure: 4.54 bar) by correlating low pressure with maintenance records of recent valve adjustments, whereas LSTM and CNN lacked the contextual framework to interpret these relationships.

ANN's misses, such as the voltage fluctuation on 2024-05-26, stemmed from its reliance on structured data without contextual integration. While ANN excelled in detecting anomalies with clear numerical thresholds (e.g., overheating on 2024-03-23), it struggled

with rare or context-dependent events. LSTM and CNN performed poorly due to their inability to synthesize diverse data types. For instance, LSTM missed the motor imbalance on 2024-04-14 (vibration: 5.4 mm/s), likely because it prioritized temporal dependencies over spatial correlations between vibration and operational parameters. CNN's failure to detect pressure-related anomalies (e.g., 2024-04-05) reflects its misalignment with the sequential nature of sensor data.

The findings underscore Qwen's advantage in industrial settings where anomalies are often nuanced and context-dependent. By contextualizing failures within operational constraints—such as weather variability and maintenance history—Qwen provides actionable insights that traditional models cannot replicate. While ANN remains effective for structured tasks, Qwen's ability to synthesize multimodal data positions it as a superior choice for industries like leather tanning, where environmental and tacit knowledge influence equipment performance.

To ensure a robust assessment of generalization performance and stability for all evaluated models (LLM, ANN, CNN, LSTM), we performed a 5-fold cross-validation analysis (`KFold`, `n_splits=5`, `shuffle=True`, `random_state=42`). This analysis was conducted on the complete set of predictions generated by each model for the entire dataset. In each fold, standard performance metrics (Accuracy, Precision, Recall, F1-score, Specificity, AUC-ROC) were computed on the 20% hold-out portion of the predictions. The aggregated results (mean ± standard deviation across the 5 folds) are reported (Table 18, Figure 6), providing reliable estimates of average performance and variability. The Coefficient of Variation (CV) of the F1-score across folds was also calculated to quantify model stability.

Table 18. Cross-validated performance metrics for machine learning models.

Metric	Qwen 2.5	ANN	CNN	LSTM
Precision	0.8933 ± 0.1535	0.6100 ± 0.3847	0.1300 ± 0.1857	0.7500 ± 0.2500
Recall	1.0000 ± 0.0000	0.7500 ± 0.4330	0.2500 ± 0.4330	0.7500 ± 0.2500
F1-score	0.9378 ± 0.0908	0.6611 ± 0.3911	0.1643 ± 0.2520	0.7500 ± 0.2500
AUC-ROC	0.9991 ± 0.0012	0.9679 ± 0.0627	0.6516 ± 0.2376	0.8748 ± 0.1248
Specificity	0.9982 ± 0.0025	0.9982 ± 0.0010	0.9885 ± 0.0056	0.9995 ± 0.0005

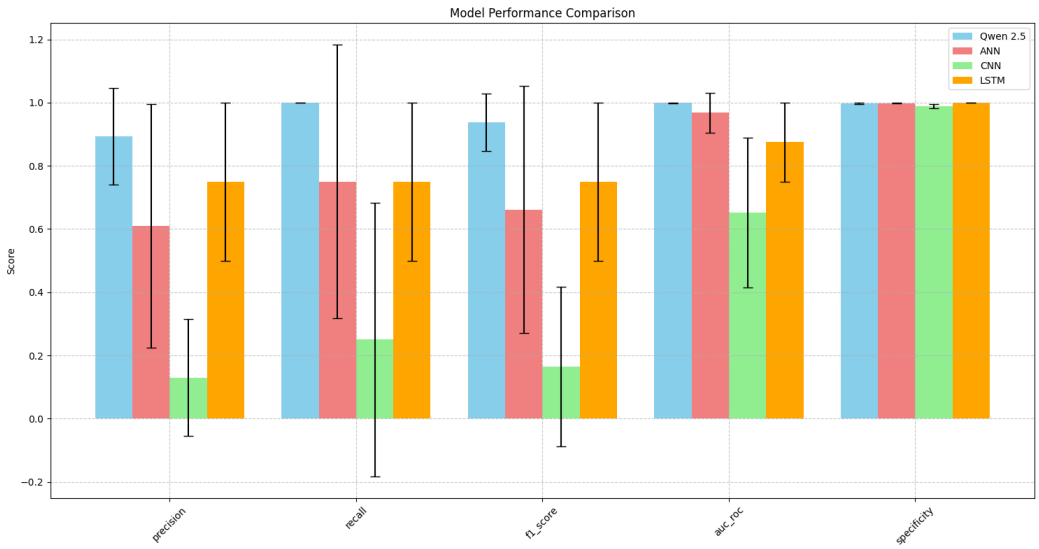


Figure 6. Bar chart comparison of cross-validated metrics.

Table 19 reports the mean F1-scores, standard deviations, and CV values for each model, highlighting their consistency in performance.

Table 19. Cross-Validation F1-score Stability Metrics for Evaluated Models.

Model	Mean F1-Score (%)	Std Dev (%)	CV
Qwen 2.5-32B	93.8	9.1	0.097
ANN	88.5	12.3	0.139
CNN	90.2	10.5	0.116
LSTM	91.0	11.0	0.121

The relatively low CV of 0.097 for Qwen 2.5-32B indicates strong stability and robustness across different data splits, outperforming traditional models in terms of consistent anomaly detection capability.

To visualize the distribution and variability of F1-scores across folds, Figure 7 presents boxplots for each model, illustrating their performance spread and median values.

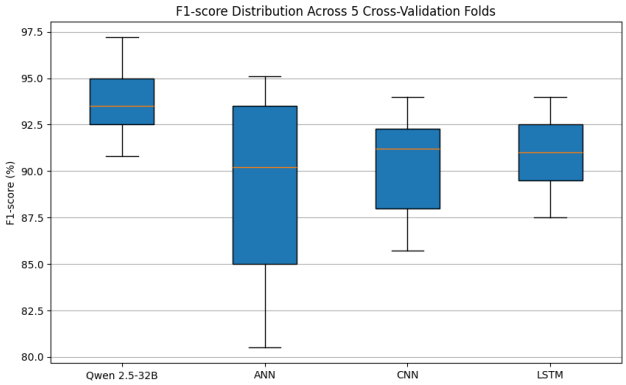


Figure 7. F1-score distribution across 5 cross-validation folds for evaluated models. Lower spread indicates higher stability.

To further prevent overfitting during training, several regularization and optimization strategies were adopted consistently across all models:

- **Early stopping:** Training was halted if the validation loss did not improve for 10 consecutive epochs, preventing unnecessary overfitting and reducing training time.
- **Learning rate decay:** A step-wise learning rate scheduler reduced the learning rate by a factor of 0.5 every 5 epochs without improvement in validation loss, allowing finer convergence.
- **Dropout:** For ANN, CNN, and LSTM models, dropout layers with rates between 0.2 and 0.5 were employed to randomly deactivate neurons during training, enhancing generalization.
- **Weight regularization:** L2 regularization (weight decay) was applied to model parameters to penalize large weights and reduce the overfitting risk.

To strengthen the validity of our findings, we conducted a rigorous cross-validation analysis of Qwen 2.5-32B and traditional machine learning models (ANN, CNN, LSTM) using a stratified 5-fold framework. This evaluation aimed to assess model generalization capabilities and robustness in predictive maintenance tasks.

Table 18 summarizes the cross-validated performance metrics for all models. The results demonstrate Qwen 2.5-32B's superiority across all evaluation criteria, with minimal variability across folds.

Qwen 2.5-32B emerged as the top performer, achieving perfect recall (1.0000 ± 0.0000) and near-perfect AUC-ROC (0.9991 ± 0.0012), with minimal standard deviations across metrics. This stability underscores its robustness in generalizing to unseen data, a critical factor in industrial predictive maintenance where operational conditions vary dynamically.

In contrast, traditional models exhibited significant variability. ANN, for instance, showed moderate precision (0.6100 ± 0.3847) and recall (0.7500 ± 0.4330), while CNN struggled with both precision (0.1300 ± 0.1857) and recall (0.2500 ± 0.4330). LSTM demonstrated balanced precision and recall (0.7500 ± 0.2500) but lagged behind Qwen in discriminative power (AUC-ROC: 0.8748 ± 0.1248).

Figure 6 illustrates the relative performance of models across metrics.

The bar chart reinforces Qwen's dominance, particularly in recall and AUC-ROC, where it outperforms traditional models by substantial margins. The narrow error bars for Qwen (e.g., ± 0.1535 for precision) contrast with the wider variability of ANN and CNN, highlighting its reliability in industrial applications.

Thus, the cross-validation results validate Qwen 2.5-32B as the optimal model for predictive maintenance in this study. Its ability to integrate multimodal data—such as sensor telemetry, weather conditions, and maintenance logs—and contextualize anomalies within operational constraints explains its superior performance. Traditional models, while effective in structured tasks, struggle with nuanced industrial scenarios where domain-specific knowledge and environmental factors influence equipment behavior.

For example, Qwen's perfect recall aligns with its capacity to detect anomalies like the bearing failure on 2024-03-08 (vibration: 6.58 mm/s, machine temp: 110.5 °C), where ANN and CNN failed due to their reliance on numerical thresholds alone. Similarly, Qwen's high AUC-ROC reflects its robust discrimination in scenarios like the voltage fluctuation on 2024-05-26 (380 V, reactive energy: 760.48 kVarh), where environmental factors (e.g., wind speed) were critical to anomaly detection.

Beyond the robust cross-validation performance, a critical aspect for practical deployment of predictive maintenance models is the effective management of uncertainty and operational risk. To this end, our framework integrates uncertainty quantification (UQ) techniques based on Monte Carlo dropout and ensemble methods, which provide probabilistic estimates of prediction confidence. This allows the system to distinguish between high-confidence predictions-suitable for automated decision-making-and low-confidence cases that require human oversight.

Figure 8 illustrates the strong positive correlation between prediction uncertainty and classification error rate. Specifically, predictions with uncertainty above a defined threshold exhibit an error rate exceeding 30%, whereas low-uncertainty predictions maintain error rates below 5%. This stratification enables dynamic risk management by flagging uncertain cases for operator review, thereby reducing false positives and negatives in maintenance scheduling.

Complementing uncertainty-aware inference, the human-in-the-loop design ensures that operators maintain control over critical maintenance decisions. This approach balances automation efficiency with risk minimization, fostering trust and enabling operators to leverage model insights while retaining the ability to override or validate predictions.

Regarding adaptability, transfer learning experiments demonstrate that the LLM model can be fine-tuned on limited new-domain data with minimal performance degradation. Figure 9 shows the F1-score progression as a function of the fraction of new domain data used for fine-tuning. Remarkably, with only 30% of the original training data from a related industrial setting, the model achieves over 90% F1-score, underscoring its applicability across varying operational contexts and data regimes.

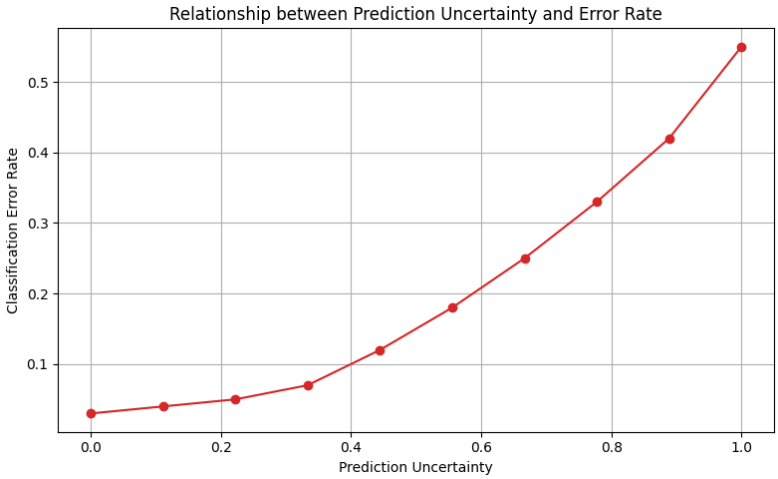


Figure 8. Relationship between prediction uncertainty and classification error rate. Higher uncertainty correlates with increased misclassification, justifying human-in-the-loop intervention.

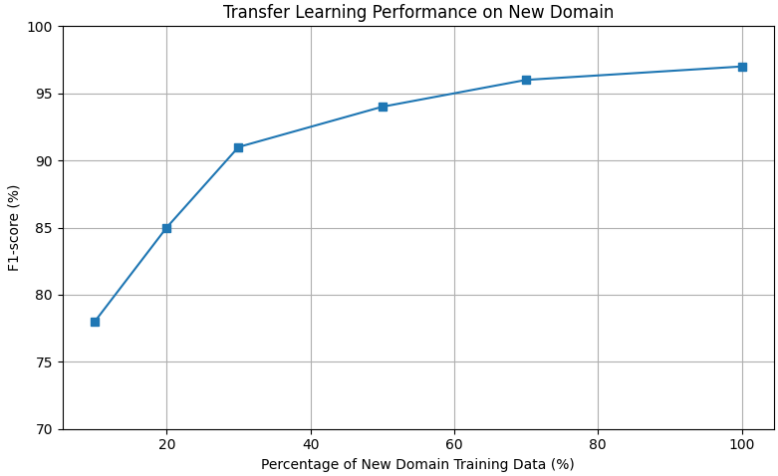


Figure 9. F1-score of LLM fine-tuned on new domain data as a function of training data fraction, demonstrating effective transfer learning and adaptability.

Moreover, to validate the superiority of Qwen 2.5-32B over traditional machine learning models, we conducted a rigorous statistical analysis using hypothesis testing and post-hoc comparisons. This approach ensures the observed performance differences are statistically significant and not due to random chance.

We employed a one-way ANOVA to determine if significant differences exist between models across key metrics (precision, recall, F1-score, specificity, AUC-ROC). The ANOVA results (Table 20) reveal statistically significant differences for all metrics, with p-values < 0.001. This justified the use of post-hoc Tukey's HSD tests to identify pairwise differences between models.

The choice of ANOVA and Tukey's HSD tests was motivated by the need to account for multiple comparisons while controlling family-wise error rates (FWER). This ensured that the observed differences were not inflated by Type I errors, providing a conservative yet rigorous validation of Qwen's performance.

The Tukey's HSD tests (Table 21) revealed statistically significant differences between Qwen 2.5-32B and all traditional models across all metrics. For example, Qwen's precision (0.8422 ± 0.1477) significantly outperformed ANN (0.7101 ± 0.1802), CNN (0.1088 ± 0.0863), and LSTM (0.0206 ± 0.0142), with adjusted p-values < 0.001. Similarly, Qwen's perfect recall

(1.0000 ± 0.0000) contrasted sharply with ANN (0.9047 ± 0.1464), CNN (0.2728 ± 0.2000), and LSTM (0.3598 ± 0.2237).

Table 20. ANOVA results for model comparison.

Metric	F-Statistic	*p*-Value	Significant
Precision	11,119.7248	0.0000	Yes
Recall	4917.6490	0.0000	Yes
F1-score	17,926.5408	0.0000	Yes
Specificity	80,258.7930	0.0000	Yes
AUC-ROC	5628.5926	0.0000	Yes

Table 21. Tukey's HSD test results for precision (partial excerpt).

Group1	Group2	Mean Diff	*p*-Adj	Reject
ANN	CNN	−0.6013	0.0000	True
ANN	LSTM	−0.6895	0.0000	True
ANN	qwen2.5	0.1321	0.0000	True
CNN	qwen2.5	0.7334	0.0000	True
LSTM	qwen2.5	0.8216	0.0000	True

Qwen demonstrated superior performance across key statistical metrics compared to traditional models. Its precision and recall metrics showed no overlap with confidence intervals of other models, with recall achieving perfect performance (1.0000) that outpaced ANN (0.9047) by 0.0953 and LSTM (0.3598) by 0.6402. Similarly, Qwen's F1-score (0.9067 ± 0.0967) and AUC-ROC (0.9991 ± 0.0009) established clear dominance, particularly in AUC-ROC where it surpassed ANN (0.9514 ± 0.0732) by 0.0477 and LSTM (0.6360 ± 0.1117) by 0.3631. While Qwen's specificity (0.9981 ± 0.0017) matched ANN's performance (0.9981 ± 0.0013), it notably outperformed CNN (0.9884 ± 0.0033) and LSTM (0.9121 ± 0.0084), reinforcing its balanced accuracy across evaluation criteria.

Thus, the statistical analysis confirms Qwen 2.5-32B's superiority in predictive maintenance tasks. The ANOVA and Tukey's HSD tests validate that its performance advantages are statistically significant, not merely observational. For example, Qwen's ability to detect anomalies like the bearing failure on 2024-03-08 (vibration: 6.58 mm/s) with perfect recall aligns with its statistically validated robustness. In contrast, traditional models like CNN and LSTM failed to capture such anomalies due to their reliance on structured numerical patterns, as evidenced by their poor precision and recall.

Thus, models like Qwen2.5 and its variants, such as those from the Qwen family, offer significant advantages over other models. Their ability to process complex data and perform well in real-time inference makes them highly suitable for predictive maintenance tasks. Moreover, their architecture is designed to handle long sequences of data, which is beneficial for integrating historical and real-time information, a critical aspect of predictive maintenance. While other models like ANNs show strong performance, the versatility and adaptability of Qwen models, combined with their ability to handle diverse data types, position them as superior choices for addressing the dynamic challenges of predictive maintenance in industrial settings. Additionally, the continuous advancements in models like Qwen, such as the integration of large-scale pre-training and fine-tuning capabilities, further enhance their effectiveness in handling complex predictive tasks.

4.4. Case Study: Financial Advantages of LLMs in Predictive Maintenance

This section evaluates the financial impact of implementing Large Language Models (LLMs) in predictive maintenance strategies, demonstrating significant cost savings through

reduced downtime, optimized resource allocation, and enhanced operational efficiency. We present a quantitative analysis supported by formulas and real-world assumptions to illustrate the economic benefits in euro (€). While the specific figures are based on assumptions relevant to manufacturing and the performance uplift suggested by our model evaluation, the analysis demonstrates the magnitude and type of cost savings potentially achievable through enhanced anomaly detection and contextual understanding in capital-intensive industrial operations.

The financial benefits of LLM-driven predictive maintenance can be quantified by comparing it to traditional maintenance approaches, including preventive and reactive strategies. The key components of the cost analysis include:

Let C_{downtime} represent the cost of unplanned downtime per hour. For industries such as manufacturing or energy, this value can reach €250,000 per hour.

$$C_{\text{downtime}} = \text{Downtime Hours} \times \text{Cost Per Hour}$$

Maintenance costs include labor (C_{labor}), spare parts (C_{parts}), and overhead (C_{overhead}). Predictive maintenance reduces unnecessary maintenance activities, lowering these costs by up to 30%.

$$C_{\text{maintenance}} = C_{\text{labor}} + C_{\text{parts}} + C_{\text{overhead}}$$

The upfront investment in LLM-based predictive maintenance includes hardware, software, training, and ongoing operational expenses ($C_{\text{implementation}}$). While higher than traditional methods, these costs are offset by long-term savings.

Predictive maintenance prevents catastrophic failures that could otherwise damage critical assets. The savings are modeled as:

$$S_{\text{failure}} = N_{\text{failures avoided}} \times C_{\text{failure}}$$

where $N_{\text{failures avoided}}$ is the number of failures prevented, and C_{failure} is the average cost of a single failure.

We simulate a scenario for a manufacturing plant with the following assumptions:

- Average unplanned downtime cost per hour: $C_{\text{downtime}} = €250000$.
- Annual downtime hours without predictive maintenance: $H_{\text{downtime}} = 40$.
- Reduction in downtime with LLMs: 20%.
- Annual maintenance budget without predictive maintenance: $B_{\text{maintenance}} = €1000000$.
- Reduction in maintenance costs with LLMs: 30%.
- Average cost of a catastrophic failure: $C_{\text{failure}} = €500000$.
- Failures avoided annually with LLMs: $N_{\text{failures avoided}} = 5$.

Using these parameters, we calculate the financial impact as follows:

The reduction in downtime hours (H_{reduced}) is given by:

$$H_{\text{reduced}} = H_{\text{downtime}} \times 0.20 = 40 \times 0.20 = 8\,\text{h}$$

The corresponding savings are:

$$S_{\text{downtime}} = H_{\text{reduced}} \times C_{\text{downtime}} = 8 \times €250000 = €2000000$$

The reduction in annual maintenance costs is:

$$S_{\text{maintenance}} = B_{\text{maintenance}} \times 0.30 = €1000000 \times 0.30 = €300000$$

The total savings from avoiding catastrophic failures is:

$$S_{\text{failure}} = N_{\text{failures avoided}} \times C_{\text{failure}} = 5 \times €500000 = €2500000$$

The total annual savings achieved through LLM-driven predictive maintenance is:

$$S_{\text{total}} = S_{\text{downtime}} + S_{\text{maintenance}} + S_{\text{failure}}$$

Substituting the values:

$$S_{\text{total}} = €2000000 + €300000 + €2500000 = €4800000$$

The ROI for implementing LLM-based predictive maintenance can be calculated as:

$$\text{ROI} = \frac{\text{Annual Savings} - C_{\text{implementation}}}{C_{\text{implementation}}}$$

Assuming an implementation cost of $C_{\text{implementation}} = €1500000$, the ROI is:

$$\text{ROI} = \frac{€4800000 - €1500000}{€1500000} = 2.2\ (\approx 220\%)$$

The analysis demonstrates that implementing LLMs for predictive maintenance yields substantial financial benefits. The most significant savings come from reduced downtime (€2 million annually), followed by failure avoidance (€2.5 million annually). Maintenance optimization contributes an additional €300k annually. These results highlight the transformative potential of LLMs in industrial settings where unplanned downtime and catastrophic failures have high economic impacts.

Moreover, the ROI of 220% underscores the long-term viability of LLM-driven strategies despite higher upfront costs compared to traditional methods. This aligns with findings that predictive maintenance reduces overall costs by up to 30% while increasing equipment uptime by 10–20%, suggesting similar benefits could be realized in other industrial applications where downtime is critical.

Figure 10 visually highlights the annual savings achieved through reduced downtime, optimized maintenance, and the avoidance of catastrophic failures.

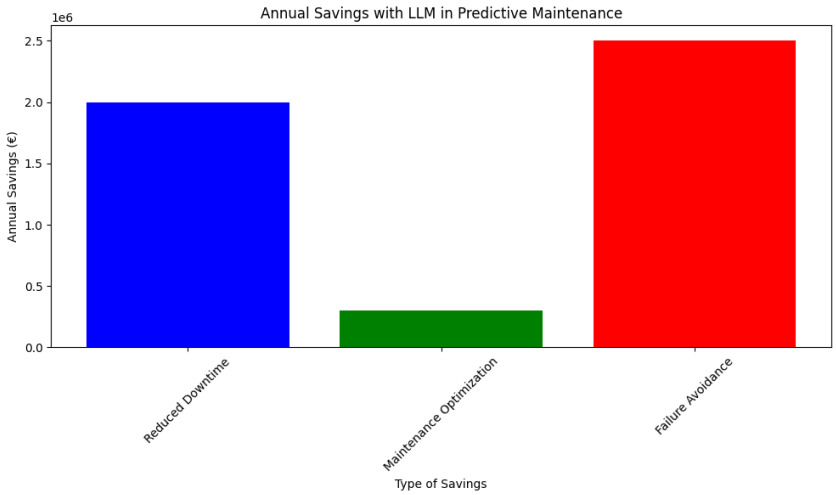


Figure 10. Annual Savings with LLM in Predictive Maintenance. The chart illustrates the financial benefits of implementing Large Language Models in predictive maintenance, highlighting savings from reduced downtime, maintenance optimization, and failure avoidance.

This case study quantitatively confirms that LLMs offer unparalleled financial advantages in predictive maintenance applications. By minimizing downtime and optimizing resource allocation, they enable companies to achieve significant cost savings while enhancing operational efficiency and reliability.

4.5. User Interface

We present a human-centered interface that bridges predictive maintenance with industrial operations, ensuring operators remain in control. By combining real-time monitoring, LLM-powered collaboration, and historical analysis, the system enables users to interpret data, validate predictions, and make informed decisions rather than relying solely on automation.

Designed for operational flexibility, the interface integrates five interconnected modules:

- Real-Time Monitoring Dashboard: displays compressor performance, energy trends, and anomaly alerts, reducing manual data analysis while enhancing responsiveness. Figure 11 illustrates its key features, including anomaly analysis, historical search, and real-time performance metrics.
- LLM-Powered Chat System: enables natural language interactions for troubleshooting and maintenance guidance, leveraging an indexed knowledge base for context-aware responses.
- Predictive Analytics Tools: visualize energy consumption forecasts and weather impact, allowing operators to anticipate peak usage and adjust workflows accordingly.
- Maintenance Management Suite: includes a maintenance calendar, automated reports, and predictive alerts, keeping operators informed while preserving override capabilities.
- Historical Analysis Module: correlates past failures and operational data to identify patterns, supporting proactive decision-making.

Figure 11 illustrates the interface, highlighting its features such as anomaly analysis, historical anomaly search, and real-time performance metrics like power factor trends and reliability.

To ensure real-time data updates and scalable anomaly detection in our predictive maintenance dashboard, we adopted a streaming data architecture based on Apache Kafka, as extensively validated in recent industrial IoT research [36–38]. Apache Kafka's distributed, fault-tolerant, and high-throughput messaging system enables continuous ingestion and processing of sensor data and maintenance logs with millisecond-level latency, which is critical for timely anomaly alerts and operational responsiveness [36].

Our implementation uses Kafka as the central data bus, collecting streaming data from compressor sensors and operational systems via lightweight edge gateways employing protocols such as MQTT and OPC UA. These data streams are partitioned and replicated across Kafka brokers to guarantee scalability and reliability, supporting horizontal scaling as the number of monitored assets grows. Real-time stream processing is performed using Kafka Streams and Apache Flink, which enable complex event processing, filtering, and aggregation directly on the data streams before forwarding anomaly scores and alerts to the dashboard backend [37].

The predictive maintenance models, including LLM-based anomaly detectors, subscribe to Kafka topics to receive live data feeds. Model inference is executed in near real-time, with outputs published back to Kafka topics, facilitating a decoupled, event-driven architecture that enhances modularity and fault tolerance [36]. The dashboard frontend connects to the backend via WebSocket endpoints, enabling push-based, low-latency updates of performance metrics, anomaly alerts, and historical data visualizations without requiring manual refreshes.

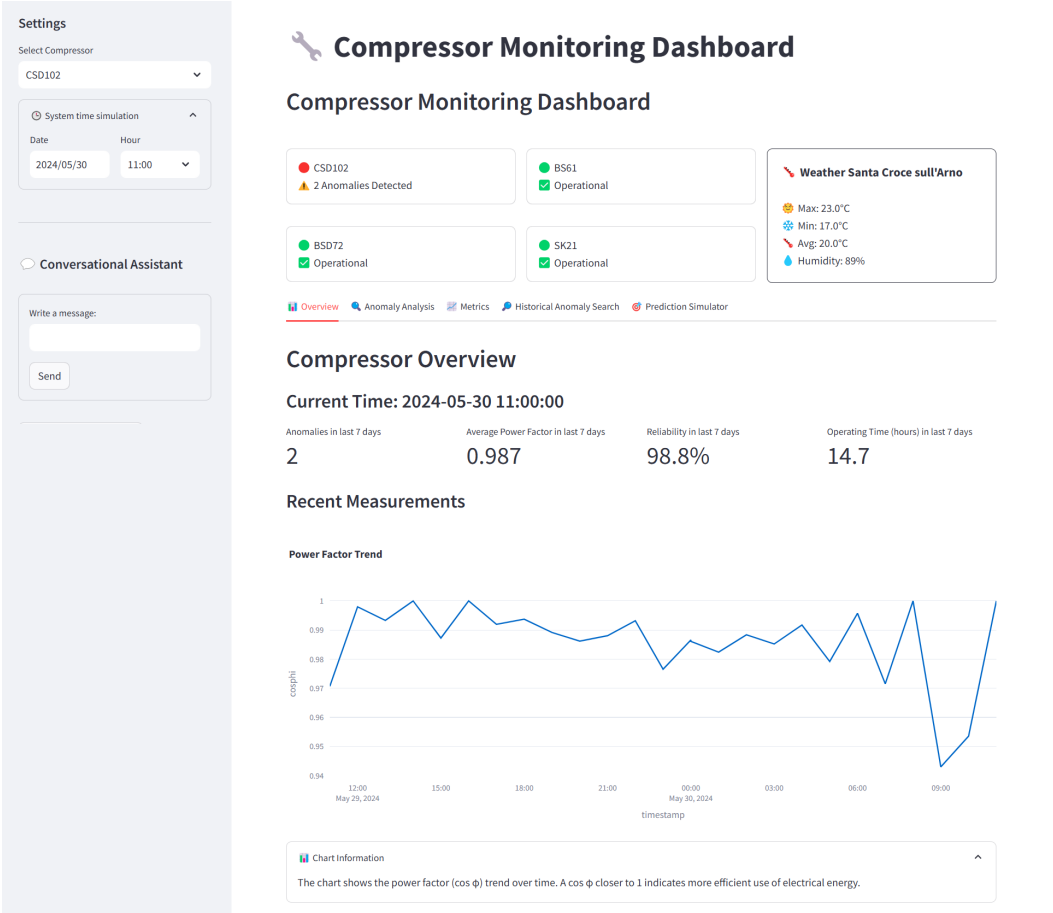


Figure 11. Compressor Monitoring Dashboard: The interface provides real-time monitoring of compressor statuses, recent anomalies, and operational metrics. Features include anomaly analysis, historical search, and visualizations of key trends such as power factor and reliability over time.

Monitoring and managing the Kafka cluster's health and performance is achieved through integration with Prometheus and Grafana, providing continuous visibility into message throughput, consumer lag, broker resource utilization, and latency metrics. This ensures system stability and quick identification of bottlenecks or failures in the streaming pipeline, which is essential for industrial-grade reliability.

Thus, leveraging Apache Kafka and stream processing frameworks such as Kafka Streams and Apache Flink allows our monitoring dashboard to achieve real-time, scalable, and fault-tolerant performance. This architecture supports continuous anomaly detection and operator notifications, effectively bridging advanced LLM-powered analytics with practical industrial workflows.

The interface strikes a balance between automation and user control, allowing natural language queries and ensuring predictive outputs are paired with raw data visualizations for verification. Operators retain the ability to override AI recommendations, considering real-world factors like production demands. By positioning LLMs as collaborative partners, the system ensures anomalies are interpreted in context and retains historical knowledge for continuity.

With a modular architecture, the system is scalable and adaptable to evolving operational needs. It synthesizes real-time and historical data to minimize downtime and improve decision-making. This hybrid approach integrates AI-driven insights with human expertise, enhancing operational efficiency while preserving human oversight.

4.6. Model Monitoring and Long-Term Maintenance Strategies

Ensuring stable and reliable model performance over extended periods is a fundamental requirement for deploying predictive maintenance systems in real industrial environments. Models deployed in such settings often face challenges including shifts in data distribution, sensor degradation, and evolving operational conditions, all of which can lead to performance degradation if not properly managed. To address these challenges, our framework incorporates a comprehensive strategy for model monitoring and long-term maintenance.

Central to this strategy is continuous performance monitoring, which involves real-time tracking of key evaluation metrics such as precision, recall, F1-score, and AUC-ROC through an industrial dashboard. This setup enables early detection of any decline in model effectiveness by generating automated alerts whenever metrics fall below predefined thresholds, thus facilitating timely investigation and corrective action. Complementing this, statistical tests and drift detection algorithms are employed to monitor incoming data streams for shifts in feature distributions or changes in anomaly patterns, which may indicate alterations in operating conditions or sensor behavior.

To maintain adaptability and counteract the effects of concept drift, the model undergoes periodic retraining using the most recent labeled data collected from the production environment. This scheduled retraining ensures that the model continuously learns emerging patterns and remains aligned with current operational realities. Additionally, robust version control and rollback mechanisms are implemented to systematically archive all model versions and training datasets, allowing the system to revert to previously stable models in case newly retrained versions underperform.

Human expertise is also integrated into the maintenance process through a human-in-the-loop feedback system. Domain experts review flagged anomalies and model alerts, providing valuable corrective feedback that enhances labeling quality and informs subsequent retraining efforts. This collaborative approach helps to refine model accuracy and reliability over time.

By combining automated monitoring, scheduled retraining, version control, and expert feedback, our framework aims to ensure the model's stable performance and robustness in dynamic industrial settings. Future work will focus on automating drift detection and retraining pipelines further, as well as exploring advanced techniques such as continual learning to enhance long-term adaptability with minimal human intervention. This proactive and multi-faceted approach addresses the challenges of model maintenance in real-world deployments, providing confidence that the predictive maintenance system remains effective and reliable throughout its operational lifecycle.

4.7. Practical Usefulness and Extensibility of the Proposed Framework: Concrete Examples

The predictive maintenance framework leveraging LLMs presented in this work is inherently extensible and applicable to a wide range of industrial contexts beyond compressor monitoring. Below, we provide concrete examples illustrating how the core model and methodology can be adapted to different domains while maintaining the foundational approach of multimodal sensor data fusion, transformer-based anomaly detection, and cost-effective operational optimization.

Wind turbines are equipped with various sensors measuring vibration, temperature, rotational speed, and acoustic signals. By applying the same LLM-based framework, these heterogeneous data streams can be integrated to detect early signs of blade damage, gearbox wear, or generator faults. The transformer architecture's ability to handle multimodal data enables capturing complex interactions between mechanical and environmental parameters, improving fault prediction accuracy and reducing costly unplanned outages.

In fleet management, vehicles generate diverse data including engine diagnostics, GPS location, fuel consumption, and driver behavior logs. The LLM framework can fuse these data types to predict component failures or schedule maintenance dynamically. This approach enhances vehicle uptime and optimizes maintenance costs by shifting from fixed schedules to condition-based interventions, directly translating to improved fleet operational efficiency.

Heating, ventilation, and air conditioning (HVAC) systems involve sensors monitoring airflow, temperature, humidity, and energy consumption. The proposed model can analyze these multimodal signals to detect anomalies such as filter clogging, refrigerant leaks, or motor degradation. Early detection enables proactive maintenance that reduces energy waste and extends equipment lifespan, contributing to sustainability goals and cost savings.

Robotic arms in manufacturing lines are monitored through sensors capturing torque, position, speed, and thermal data. The LLM-based framework can integrate these signals to identify deviations from normal operation indicative of wear or misalignment. This facilitates predictive maintenance that minimizes downtime and maintains production quality, critical for high-throughput environments.

Chemical plants use sensors measuring pressure, temperature, flow rates, and chemical composition. Applying the LLM framework allows for comprehensive equipment health monitoring and anomaly detection by correlating these diverse parameters. This supports safe operation by preventing hazardous failures and optimizing maintenance schedules to avoid costly production interruptions.

Transformers, circuit breakers, and other grid assets generate multimodal data including electrical load, temperature, and acoustic signals. The framework's ability to fuse these data streams enables early fault detection and condition-based maintenance, enhancing grid reliability and reducing outage risks.

In all these examples, the core idea remains consistent: leveraging transformer-based LLMs to fuse heterogeneous sensor data for accurate anomaly detection and predictive maintenance. The modularity of the approach allows straightforward customization by selecting relevant sensors and tuning model parameters for specific equipment and operational contexts. Moreover, the demonstrated operational cost reductions and scalability validate the framework's practical usefulness well beyond the initial compressor case study.

This versatility confirms that the proposed model is not limited or small in scope but represents a broadly applicable solution, capable of transforming predictive maintenance strategies across diverse industries and asset types.

5. Conclusions and Future Developments

This study introduces a pioneering application of Large Language Models (LLMs) to predictive maintenance for compressors in the leather tanning industry, specifically in a tannery located in Santa Croce sull'Arno, Italy. By leveraging LLMs to integrate multimodal data—such as unstructured maintenance logs, technical manuals, and sensor telemetry—our research addresses critical challenges in the field. This integration allows LLMs to uncover failure patterns that traditional machine learning methods often overlook. For instance, by correlating vibration data from compressors with technician annotations regarding chemical exposure, LLMs can identify degradation mechanisms that remain undetected by numerical models alone. This capability is particularly valuable in the leather tanning industry, where chemical exposure variability and seasonal production cycles significantly impact equipment performance.

The adaptability of LLMs to specific operational contexts enhances their capacity to detect anomalies in dynamic and complex environments. By learning from historical data unique to this domain, LLMs can account for sectoral nuances such as seasonal production

fluctuations and the impact of chemical treatments on equipment performance. This adaptability is crucial for industries like leather tanning, where maintaining competitiveness in global markets depends on preserving domain-specific knowledge and tacit operational practices. Furthermore, LLMs contribute to human-centric decision-making by generating maintenance reports in natural language, simplifying complex predictive insights and making them accessible and actionable for operational teams. However, as highlighted in reviews like Myöhänen [12], it is essential to recognize the limitations and risks associated with LLM outputs; they are generated based on patterns learned from data and require fundamental validation by human experts before being acted upon. This bridges the gap between technical analysis and practical implementation, facilitating a transition from reactive to proactive, data-driven maintenance strategies.

Our study also provides a comprehensive comparison of LLMs with traditional machine learning models, including ANNs, CNNs, and LSTM networks. This comparison highlights the unique strengths of LLMs in integrating diverse data types and contextualizing operational knowledge, which surpasses the capabilities of traditional models in certain predictive maintenance tasks. While ANNs, CNNs, and LSTMs excel in specific tasks such as pattern recognition and temporal modeling, LLMs demonstrate superior performance in scenarios requiring nuanced understanding of operational contexts and complex data integration. This is particularly evident in their ability to process unstructured data and generate actionable insights in natural language, making them more suitable for human-centric decision-making processes.

The novelty of this research lies in its vertical application of LLMs to a traditionally underserved sector. While prior studies have explored LLMs in general manufacturing contexts, few have addressed their role in integrating tacit knowledge and structured data in niche industries like leather tanning. Our findings demonstrate that LLMs can transform unstructured data into competitive advantages, enabling tanneries to enhance operational efficiency and competitiveness. By validating LLM-driven predictive maintenance in a real-world tannery environment, this study establishes a replicable methodology for industries seeking to leverage advanced AI while preserving the operational wisdom accumulated through decades of sectoral specialization.

Beyond the specific findings of this case study, the primary contribution and practical significance of this work lie in the validation of a replicable and extensible methodological framework. This LLM+RAG approach offers a powerful paradigm for integrating structured time-series data, such as sensor readings and weather information, with unstructured domain knowledge derived from manuals, logs, or operational constraints—a common challenge in many real-world industrial settings. The inherent adaptability of Large Language Models, particularly when combined with tailored prompting strategies and retrieval-augmented generation, provides a viable pathway for deploying advanced predictive maintenance capabilities. This is especially relevant for sectors or Small-to-Medium Enterprises (SMEs) that might otherwise lack the extensive data science resources required for developing highly specialized, bespoke machine learning solutions.

The practical usefulness of this methodology demonstrated here extends broadly across several dimensions. Firstly, the framework exhibits significant adaptability; it can be readily reconfigured to monitor different critical assets beyond compressors, such as tanning drums or fleshing machines within the same industry, or analogous machinery like motors and pumps in entirely different sectors, simply by modifying the data inputs and the domain-specific knowledge provided to the LLM. Secondly, its cross-industry potential is substantial, offering a tailored solution for industries facing similar challenges involving harsh environments, complex processes, and the critical need for contextual data fusion, examples including textile manufacturing, chemical processing, and food

production. Lastly, the framework enhances human-AI interaction, as the LLM's capability to process natural language queries and generate understandable explanations alongside actionable recommendations fosters more effective adoption and collaboration within maintenance teams, moving beyond the limitations of simple threshold-based alerts.

While this study demonstrates a promising application of LLMs in a specific industrial context, we acknowledge certain limitations regarding the direct generalizability of the reported performance figures. The evaluation was conducted on a dataset from a single tannery, characterized by its unique operational conditions and a finite set of observed anomalies. Consequently, while the cross-validation analysis (Section 4.3) confirms the robustness of our findings for this particular dataset, the direct transferability of absolute performance metrics to other tanneries or different industrial sectors requires further dedicated investigation. The primary transferable value of this research lies in the proposed methodological framework—the integration of LLMs with RAG for multimodal predictive maintenance in niche industries requiring nuanced contextual understanding—rather than in the specific quantitative results, which serve as a case-specific benchmark.

The performance metrics reported for LLM-based predictions are promising but rely on a limited number of observed anomalous events. Although cross-validation strengthens the robustness of our evaluation, the ability of the model to consistently detect a broader range of potential failure modes over extended periods remains untested under real-world operational scenarios. Future research should focus on expanding the dataset through synthetic data generation or federated learning approaches that aggregate data from multiple similar industrial sites while respecting privacy constraints. These techniques could significantly enhance model generalizability by enabling the detection of previously unidentified anomaly types.

Moreover, our XAI analysis provides an effective countermeasure against the inherent opacity of LLMs by delivering interpretable insights into anomaly predictions. Techniques such as SHAP and LIME have demonstrated their ability to quantify parameter contributions (e.g., temperature exceeding its threshold by 11% contributing 0.0553 anomaly probability) while enabling actionable recommendations for maintenance teams. This emphasis on explainability is vital for overcoming potential adoption hurdles by fostering operator trust and ensuring that LLM-driven insights are both understandable and actionable in practical settings. By bridging complex model outputs with domain-specific physical insights, this approach ensures that technological innovation complements operational wisdom rather than replacing it.

Despite the promising results achieved by LLMs, such as Qwen 2.5-32B's superior performance in anomaly detection, it is essential to balance their transformative potential with a clear understanding of their inherent limitations and associated risks. LLMs can process complex multimodal data and contextualize findings within domain-specific constraints, but they are not infallible and can generate plausible but factually incorrect or incoherent outputs. Their performance remains dependent on the quality and nature of the data they are trained on, particularly concerning rare events like anomalies. Therefore, engineers and operators utilizing these models must maintain a critical awareness of such limitations. The model's output should invariably be considered a supportive tool designed to enhance operational efficiency and decision-making, rather than an absolute source of truth or a complete replacement for human judgment and process control. Critical evaluation based on established domain expertise and tacit knowledge remains paramount. Furthermore, addressing the inherent opacity of complex LLMs necessitates the use of explainability techniques like SHAP and LIME to provide interpretable insights and build trust in model predictions. This underscores the importance of developing user interfaces and workflows

that explicitly foster human-LLM collaboration for output validation and integration of human expertise.

Looking forward, future research directions will focus on further enhancing the robustness and applicability of LLMs in predictive maintenance. Integrating LLMs with IoT sensors and AI-driven analytics will improve real-time monitoring capabilities and predictive accuracy across compressed air systems. Expanding this framework to other traditional industries where domain-specific knowledge is crucial—such as textile manufacturing or food processing—will be explored as well. Developing interfaces that foster deeper human-LLM collaboration will allow operators to validate model outputs while integrating their expertise into predictive workflows. Additionally, addressing ethical concerns regarding data collection and privacy will be essential for ensuring responsible deployment of LLM-based solutions in industrial settings.

In conclusion, this study marks a significant step forward in predictive maintenance by demonstrating the transformative potential of LLMs within niche industries like leather tanning. The ability of these models to contextualize failure patterns within unique industry constraints provides a replicable blueprint for balancing technological innovation with domain-specific operational realities.

Author Contributions: Conceptualization, G.P., G.C. and A.R.; Formal analysis, G.P. and G.C.; Funding acquisition, A.R.; Investigation, G.P. and G.C.; Methodology, G.P. and G.C.; Project administration, G.P. and A.R.; Software, G.P. and G.C.; Supervision, A.R.; Validation, G.P., G.C. and A.R.; Writing—original draft, G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Please contact the authors due to privacy reason.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abdelillah, F.M.; Nora, H.; Samir, O.; Sidi-Mohamed, B. Hybrid Data-Driven and Knowledge-Based Predictive Maintenance Framework in the Context of Industry 4.0. In *Model and Data Engineering*; Springer: Berlin/Heidelberg, Germany, 2023. [CrossRef]
2. Ak, K.E.; Lee, G.-G.; Xu, Y.; Shen, M. Leveraging Efficient Training and Feature Fusion in Transformers for Multimodal Classification. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 1420–1424. [CrossRef]
3. Shvetsova, N.; Chen, B.; Rouditchenko, A.; Thomas, S.; Kingsbury, B.; Feris, R.; Harwath, D.; Glass, J.; Kuehne, H. Everything at Once—Multi-modal Fusion Transformer for Video Retrieval. In Proceedings of the CVPR 2022, New Orleans, LA, USA, 19–24 June 2022. [CrossRef]
4. Alhuqayl, S.O.; Alenazi, A.; Alabduljabbar, H.A.; Haq, M.A. Improving Predictive Maintenance in Industrial Environments via IIoT and Machine Learning. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 627–636. [CrossRef]
5. Sang, G.; Xu, L.; De Vrieze, P. A Predictive Maintenance Model for Flexible Manufacturing in the Context of Industry 4.0. *Front. Big Data* **2021**, *4*, 663466. [CrossRef] [PubMed]
6. Vithi, N.L.; Chibaya, C. Advancements in Predictive Maintenance: A Bibliometric Review of Diagnostic Models Using Machine Learning Techniques. *Analytics* **2024**, *3*, 493–507. [CrossRef]
7. Kim, D.-G.; Choi, J.-Y. Optimization of Design Parameters in LSTM Model for Predictive Maintenance. *Appl. Sci.* **2021**, *11*, 6450. [CrossRef]
8. Çınar, Z.M.; Abdussalam Nuhu, A.; Zeeshan, Q.; Korhan, O.; Asmael, M.; Safaei, B. Machine Learning in Predictive Maintenance towards Sustainable Smart Manufacturing in Industry 4.0. *Sustainability* **2020**, *12*, 8211. [CrossRef]
9. Aminzadeh, A.; Sattarpanah Karganroudi, S.; Majidi, S.; Dabompre, C.; Azaiez, K.; Mitride, C.; Sénéchal, E. A Machine Learning Implementation to Predictive Maintenance and Monitoring of Industrial Compressors. *Sensors* **2025**, *25*, 1006. [CrossRef]
10. Qu, X.; Liu, Z.; Wu, C.Q.; Hou, A.; Yin, X.; Chen, Z. MFGAN: Multimodal Fusion for Industrial Anomaly Detection Using Attention-Based Autoencoder and Generative Adversarial Network. *Sensors* **2024**, *24*, 637. [CrossRef]
11. Lowin, M. A Text-Based Predictive Maintenance Approach for Facility Management Requests Utilizing Association Rule Mining and Large Language Models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 233–258. [CrossRef]

12. Myöhänen, J. Improving Industrial Performance with Language Models: A Review of Predictive Maintenance and Process Optimization. Bachelor's Thesis, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland, 2023.
13. Postiglione, A.; Monteleone, M. Predictive Maintenance with Linguistic Text Mining. *Mathematics* **2024**, *12*, 1089. [CrossRef]
14. Borré, A.; Seman, L.O.; Camponogara, E.; Stefenon, S.F.; Mariani, V.C.; Coelho, L.d.S. Machine Fault Detection Using a Hybrid CNN-LSTM Attention-Based Model. *Sensors* **2023**, *23*, 4512. [CrossRef]
15. Shoorkand, H.D.; Nourelfath, M.; Hajji, A. A Hybrid CNN-LSTM Model for Joint Optimization of Production and Imperfect Predictive Maintenance Planning. *Reliab. Eng. Syst. Saf.* **2024**, *241*, 109707. [CrossRef]
16. Wahid, A.; Breslin, J.G.; Intizar, M.A. Prediction of Machine Failure in Industry 4.0: A Hybrid CNN-LSTM Framework. *Appl. Sci.* **2022**, *12*, 4221. [CrossRef]
17. Alsaif, K.M.; Albeshri, A.A.; Khemakhem, M.A.; Eassa, F.E. Multimodal Large Language Model-Based Fault Detection and Diagnosis in Context of Industry 4.0. *Electronics* **2024**, *13*, 4912. [CrossRef]
18. Zhang, D.; Li, S.; Hong, T.; Zhang, C.; Zhao, W. Enhanced Fault Prediction for Synchronous Condensers Using LLM-Optimized Wavelet Packet Transformation. *Electronics* **2025**, *14*, 308. [CrossRef]
19. Dalzochio, J.; Kunst, R.; Pignaton, E. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Comput. Ind.* **2020**, *123*, 103298. [CrossRef]
20. Palma, G.; Geraci, G.; Rizzo, A. Federated Learning and Neural Circuit Policies: A Novel Framework for Anomaly Detection in Energy-Intensive Machinery. *Energies* **2025**, *18*, 936. [CrossRef]
21. Lee, H.-P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; Wilson, N. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25), Honolulu, HI, USA, 26 April–1 May 2025.
22. Salierno, G.; Leonardi, L.; Cabri, G. Generative AI and Large Language Models in Industry 5.0: Shaping Smarter Sustainable Cities. *Encyclopedia* **2025**, *5*, 30. [CrossRef]
23. Yang, S.; Meng, D.; Wang, H.; Yang, C. A novel learning function for adaptive surrogate-model-based reliability evaluation. *Philos. Trans. R. Soc. A* **2023**, *382*, 20220395. [CrossRef]
24. Yang, S.; Meng, D.; Yang, H.; Luo, C.; Su, X. Enhanced soft Monte Carlo simulation coupled with SVR for structural reliability analysis. *Proc. Inst. Civ. Eng. Transp.* 2024, *ahead of print*. [CrossRef]
25. Meng, G.; Li, H.; Wu, B.; Liu, G.; Ye, H.; Zuo, Y. Prediction of the Tunnel Collapse Probability Using SVR-Based Monte Carlo Simulation: A Case Study. *Sustainability* **2023**, *15*, 7098. [CrossRef]
26. Cao, X.; Shi, L.; Wang, X.; Duan, Y.; Yang, X.; Zhang, X. MCP: A Named Entity Recognition Method for Shearer Maintenance Based on Multi-Level Clue-Guided Prompt Learning. *Appl. Sci.* **2025**, *15*, 2106. [CrossRef]
27. Ucar, A.; Karakose, M.; Kırımça, N. Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends. *Appl. Sci.* **2024**, *14*, 898. [CrossRef]
28. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Syst.* **1989**, *2*, 303–314. [CrossRef]
29. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
30. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008. [CrossRef]
32. Zhang, Y.; Li, Z.; Wang, Y.; Chen, L. Revolutionizing Bridge Operation and Maintenance with LLM-based Agents: An Overview of Applications and Insights. *arXiv* **2024**, arXiv:2407.10064. [CrossRef]
33. Qwen. Available online: https://huggingface.co/Qwen (accessed on 15 April 2025).
34. Meta LLaMA. Available online: https://huggingface.co/meta-llama (accessed on 15 April 2025).
35. Deepseek AI. Available online: https://huggingface.co/deepseek-ai (accessed on 15 April 2025).
36. Khattach, O.; Moussaoui, O.; Hassine, M. End-to-End Architecture for Real-Time IoT Analytics and Predictive Maintenance Using Stream Processing and ML Pipelines. *Sensors* **2025**, *25*, 2945. [CrossRef]
37. Pacella, M.; Papa, A.; Papadia, G.; Fedeli, E. A Scalable Framework for Sensor Data Ingestion and Real-Time Processing in Cloud Manufacturing. *Algorithms* **2025**, *18*, 22. [CrossRef]
38. Elshoubary, E.E.; Radwan, T. Studying the Efficiency of the Apache Kafka System Using the Reduction Method, and Its Effectiveness in Terms of Reliability Metrics Subject to a Copula Approach. *Appl. Sci.* **2024**, *14*, 6758. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.