Exploring the Impact of Dataset Accuracy on Machinery Functional Safety: Insights from an AI-Based Predictive Maintenance System

Padma Iyenghar[1,2][a]
[1] innotec GmbH, Hornbergstrasse 45, 70794 Filderstadt, Germany
[2] Faculty of Engineering and Computer Science,
University of Applied Sciences Osnabrueck, 49009 Osnabruck, Germany

Keywords: Data Quality, Data Accuracy, Functional Safety, Artificial Intelligence (AI), Machine Learning (ML), Predictive Maintenance, Reliability, Availability.

Abstract: This paper focuses on the critical role of dataset accuracy in the context of machinery functional safety within an AI-based predictive maintenance system in a manufacturing setting. Through experiments introducing perturbations simulating real-world challenges, a decrease in performance metrics was observed—factors such as sensor noise, labeling errors, missing data, and outliers were identified as contributors to the compromise of the AI model's accuracy. Implications for reliability and availability were discussed, emphasizing the need for high-quality datasets to minimize the risk of unplanned downtime. Recommendations include the implementation of robust data quality assurance processes and improved outlier detection mechanisms to ensure the reliability and availability of machinery in high-risk environments.

1 INTRODUCTION

AI-enabled machinery functional safety involves the application of Artificial Intelligence (AI) techniques to enhance safety functions within machinery in adherence to established standards such as ISO 13849 (ISO13849, 2023) and ISO 62061 (IEC62061, 2022). One prominent example is predictive maintenance, where AI algorithms analyse data from sensors and equipment to predict potential failures before they occur. By employing Machine Learning (ML) models, such systems can identify patterns, anomalies, and degradation in machinery performance, enabling timely maintenance interventions to prevent unexpected breakdowns. This approach aligns with the broader principles of machinery functional safety, aiming to minimize risks and ensure safe operation.

However, recognizing the pivotal role of data quality is essential for the effective implementation of AI-driven enhancements in a machinery functional safety. Multiple definitions of data quality exist in the literature. A commonly adopted definition is *fitness for use for a specific purpose*. In the context of functional safety, this entails ensuring that the data is suitable for attaining technological objectives, facilitating informed and effective decision-making, and optimizing functional safety processes for AI-based software and systems. Poor data quality can lead to unreliable models and may pose significant risks. Thus, ensuring data quality is crucial when developing AI/ML models for high-risk applications in the context of functional safety.

In the aforesaid context, this paper conducts a comprehensive investigation on the impact of dataset accuracy in machinery functional safety example. The study provides novel insights into the implications of dataset inaccuracies on the reliability and availability of machinery in high-risk environments. Beyond analysis, the paper offers practical recommendations for practitioners, suggesting measures to enhance data quality assurance processes and improve outlier detection mechanisms. The findings and recommendations serve as valuable guidance for practitioners involved in the development of AI systems for machinery functional safety and beyond, emphasizing practical considerations for achieving reliable and available machinery.

Inline with the aforementioned contributions, the following represent the novelties:

- Real-world use case focus: The paper centres on a practical use case of predictive maintenance in a manufacturing setting, providing insights into the

[a] https://orcid.org/0000-0002-1765-3695

Iyenghar, P.
Exploring the Impact of Dataset Accuracy on Machinery Functional Safety: Insights from an AI-Based Predictive Maintenance System.
DOI: 10.5220/0012683600003687

In *Proceedings of the 19th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2024)*, pages 484-497
ISBN: 978-989-758-696-5; ISSN: 2184-4895

implications of dataset inaccuracies.

- Systematic experimentation: The study employs systematic experiments introducing perturbations (i.e., intentional disruptions or alternations), simulating common challenges faced by real-world datasets used for training AI models.
- Identification of contributing factors: Specific factors, including sensor noise, labelling errors, missing data, and outliers, are identified and analysed for their contribution to the compromise of the AI model's accuracy.

The remainder of the paper is organized as follows. Following this introduction section, related work and a summary of key insights on data quality assurance for AI/ML model in the context of functional safety is presented in section 2. The real-world practical use case and experimental scenarios are detailed in section 3. Results and analysis are presented in section 4. Discussion and conclusion is presented in section 5.

2 RELATED WORK AND SUMMARY OF KEY INSIGHTS

Data quality is a fundamental consideration in the development of AI/ML models, because it directly impacts the performance, accuracy, and reliability of AI models[1]. The critical importance of high-quality data for AI applications is empirically investigated in (Budach, 2022). The impact of six traditional data quality dimensions such as *consistent representation, completeness, feature accuracy, target accuracy, uniqueness and target class balance* on the performance of fifteen popular ML algorithms across various tasks, highlighting the relationship between data quality dimensions and algorithm effectiveness, is presented in (Budach, 2022). This study underscores the critical role of accurate labelling, feature accuracy, and data quality dimensions in different ML scenarios, providing insights on algorithm performance robustness. But, a specific use case pertaining to a specific domain like functional safety is missing in (Budach, 2022).

More recently, in (Priestley et al., 2023) a survey of data quality requirements that matter in ML development pipelines is presented. The proposed framework categorizes data quality criteria based on both the stage of the ML lifecycle and the main dimensions of data quality, providing a practical guide for data practitioners and organizations to enhance their data management routines in preparation for ML applications. While (Sessions and Valtorta, 2009) introduces a data accuracy assessment method employing a Bayesian Network learning algorithm, empirical studies examining the influence of dataset accuracy on machinery functional safety, particularly in real-world scenarios, are currently lacking.

On the other hand, several approaches have emerged to validate data fed to ML pipelines. For example, the validation system implemented in (Polyzotis et al., 2019) and (Schelter et al., 2020) focus on validating data given a classification pipeline as a black box. In (Schelter et al., 2021), an experimental library JENGA is introduced for testing ML-model's robustness under data errors. While the authors use the concept of polluters or data corruptions, an extensive experimental study is missing in (Schelter et al., 2021). Automatic and precise data validation for ML is discussed in (Shankar et al., 2023).

A most recent survey on AI for safety-critical systems in industrial and transportation domains is presented in (Perez-Cerrolaza, 2023). However, it lacks an in-depth experimental evaluation on a specific aspect such as dataset quality in an industrial use case. Further, experiences of adopting automated data validation in an industrial ML project is presented in (Lwakatare et al., 2021). The results in (Lwakatare et al., 2021) indicate that adopting a data validation process and tool in ML projects is an effective approach of testing ML-enabled software systems.

The literature review, state-of-the-art analysis, and related work collectively reveal that various dimensions of data quality have been scrutinized in relation to the performance of ML algorithms. Nevertheless, there is a noticeable gap in specific research focusing on data quality assurance within the functional safety domain. Moreover, there is a lack of exploration into crucial aspects like dataset accuracy in a machinery setting, particularly in real-world use cases such as AI-based predictive maintenance systems.

2.1 Data Quality Assurance Facets

In this section, some important facets of data quality assurance within AI-based machinery functional safety software/systems are summarized. General insights are drawn from the literature review, nevertheless, an effort is made to adapt these insights to the specific requirements of the functional safety domain wherever applicable.

1. Safety critical scenario identification (data relevance): In the context of machinery functional safety, it's crucial to identify safety-critical scenarios. This involves collaboration with domain

[1] https://research.aimultiple.com/data-quality-ai/

experts to understand potential hazards and risks associated with machinery operations. Simulation tools like Simulink[2] and LabVIEW [3] can be employed to analyse and simulate critical scenarios to ensure the system's response meets safety requirements.

2. Safety validation procedures (data accuracy): Ensuring data accuracy includes implementing validation procedures during data collection and cross-referencing with authoritative sources. Implementing validation procedures specific to safety-critical data involves rigorous testing and verification processes during data collection. Tools like Great Expectations [4] or custom scripts are useful for data integrity checks.
3. Safety standard compliance (data consistency): Machinery functional safety often requires adherence to specific safety standards (e.g., ISO 13849 for machinery safety). Data consistency in terms of formats, units, and representations becomes crucial for compliance. Libraries such as Pandas[5] or TensorFlow Data Validation[6] can be applied for standardization and cleaning.
4. Safety gap analysis (data completeness): Addressing data completeness involves performing gap analysis specifically tailored to safety requirement. Identifying missing data points that are critical for safety assessments and augmenting datasets with relevant safety-related information is essential. Tools like Apache NiFi [7] or custom scripts are valuable for data augmentation.
5. Real-time safety monitoring (temporal relevance): Here, temporal relevance extends to real-time monitoring of safety-critical data. Ensuring that the safety data is continuously updated to reflect changes in the operational environment is crucial. Automated data update scripts, alongside version control systems like Git [8], facilitate tracking changes over time.
6. Safety critical bias mitigation (data collection bias): Similar to general AI systems, machinery functional safety must address bias, especially if it could impact safety-critical decisions. Conducting bias analysis specific to safety-related data and employing tools like TensorFlow's Fairness Indicators [9] and AI Fairness 360[10] becomes essential for mitigating biases in safety-critical applications.
7. Functional safety security measures (data security and privacy): Data security and privacy in machinery functional safety involve additional considerations due to the potential impact on human safety. Encryption and anonymization techniques must be robustly implemented. Privacy-preserving libraries like PySyft[11] are valuable for safeguarding safety-critical data in the context of machinery operations.
8. Safety critical event labeling (data labelling q)uality: In the realm of machinery functional safety, labelling quality extends beyond general data labelling. It involves accurately labelling safety-critical events and standardizing procedures for doing so. Labelling tools like Labelbox[12], along with tools like Dataturks[13] for reliability testing, are instrumental in accurate data labelling. They can be adapted to ensure accurate labelling of safety-critical data/events.
9. Safety outlier detection and handling: Outliers in safety-critical data may indicate abnormal behaviour that could lead to hazardous situations. Specialized outlier detection methods using statistical techniques or ML algorithms Tools like Scikit-learn (Pedregosa, 2011) for Python provide various algorithms for outlier detection, and custom preprocessing scripts can be employed for identifying and handling safety-critical outliers.
10. Safety documentation protocols: Documentation in machinery functional safety domain involves not only general data documentation but also specific safety documentation. Comprehensive data dictionaries, safety protocols, and preprocessing steps must be documented. Documentation systems like Confluence[14] or version-controlled repositories like GitHub[15] are effective for managing data documentation.
11. Validation and Testing: Validation and testing in the context of machinery functional safety require a heightened focus on ensuring that safety requirements are met. Using separate datasets for training, validation, and testing becomes critical, and

[2] https://www.mathworks.com/help/simulink/
[3] https://www.ni.com/documentation/en/labview/
[4] https://greatexpectations.io/
[5] https://pandas.pydata.org/
[6] https://www.tensorflow.org/tfx/guide/tfdv
[7] https://nifi.apache.org/docs.html
[8] https://git-scm.com/doc
[9] https://www.tensorflow.org/tfx/guide/tfma
[10] https://github.com/Trusted-AI/AIF360
[11] https://github.com/OpenMined/PySyft
[12] https://www.labelbox.com/
[13] https://www.dataturks.com/
[14] https://www.atlassian.com/software/confluence
[15] https://github.com/

automated testing frameworks like TensorFlow Extended (TFX)[16] or pytest[17] should be adapted for safety-critical model validation and testing.

12. Continuous Monitoring: Continuous monitoring in the context of machinery functional safety involves real-time monitoring of safety-critical data and model performance. Establishing alerting systems for drift detection like Seldon Alibi Detect[18] and utilizing monitoring platforms contribute to effective continuous monitoring of safety-related aspects in machinery operations.

Thus, to ensure reliability and integrity of data in functional safety-critical applications, a comprehensive data quality assurance framework is essential. The crux of the such a framework involves ensuring accuracy, consistency, completeness, and relevance of data. This includes measures such as domain analysis, accurate data handling, standardization, addressing bias, security measures, high-quality labelling, outlier detection, thorough documentation, and continuous monitoring. The selection of tools is crucial, emphasizing the need for careful evaluation based on project requirements and technology stack in functional safety domains.

While various aspects have been discussed above, the work presented in this paper primarily focuses on assessing the influence of dataset accuracy on functional safety. The evaluation is conducted through the analysis of an AI-based predictive maintenance system, extracting valuable insights on dataset accuracy from a systematic experimentation.

3 SYSTEMATIC EXPERIMENTATION

Let us consider a real-world scenario in the context of machinery functional safety, specifically for a manufacturing process where an AI-based system is used for predictive maintenance. The AI model is trained to predict potential failures in a critical component of a machine. The machine has a key component, a motor, and the aforementioned AI model has been deployed to predict potential failures in the motor based on (historical) sensor data.

As outlined in section 2.1, *data accuracy* stands out as one of the key aspects in ensuring data quality assurance. The objective of our systematic experimentation is to illustrate the impact of input data accuracy on the performance of the model, subsequently influencing predictive maintenance outcomes and implications for functional safety aspects, such as reliability and availability.

3.1 Dataset Description

The sensor values are used to simulate input data for the aforesaid predictive maintenance AI model. In the real-world scenario, this involves training a model to predict whether a machine is likely to experience a failure or malfunction based on historical sensor data.

The AI model, is trained to learn patterns in the input data (features) to predict the target variable (labels). The synthetic dataset is generated with five features for each data point. They are, *temperature*: which is the operating temperature of the motor, *vibration:* the vibration levels experienced by the motor, *current draw:* the amount of current the motor is drawing during operation, *voltage fluctuations:* fluctuations in the voltage supplied to the motor and *running hours:* the total number of hours the motor has been in operation. These values are intended to simulate readings from sensors on machinery. The labels are binary (0 or 1) and represent the target variable. In the context of predictive maintenance, these labels indicate whether a machine is operating normally (0) or if there is a potential issue (1).

In the experimental setup, a range of datasets each comprising different sample sizes namely, 1000, 10000, 20000, 50000, 100000, 150000, and 300000 is utilized. The investigation aims to analyse the impact of varying data quality and also incorporates a sensitivity analysis by simulating issues within the datasets. This approach allows to comprehensively assess how challenges in data accuracy affect the performance of AI/ML models, especially concerning accuracy and generalization across different dataset sizes.

Please note that, while the paper presents a real-life case study of predictive maintenance, the utilization of synthetic datasets is a common practice in ML research for several reasons. Firstly, synthetic datasets offer control over various parameters, enabling systematic experimentation and analysis of model behaviour under different conditions. Secondly, real-world datasets for high-risk environments like predictive maintenance can be limited or inaccessible due to privacy concerns or proprietary reasons. Therefore, synthetic data allows to simulate realistic scenarios while maintaining data privacy and accessibility. Additionally, using synthetic data facilitates the reproducibility of the study, as the dataset generation process can be fully documented and shared, ensuring transparency and enabling other researchers

[16] https://www.tensorflow.org/tfx

[17] https://docs.pytest.org/en/7.4.x/

[18] https://github.com/SeldonIO/alibi-detect

to validate and build upon the findings. Thus, while the data may be synthetic, the study's focus remains on addressing real-world challenges and providing insights applicable to practical predictive maintenance scenarios.

3.2 Issues in Dataset

The experimental setup considers a specific set of issues with the AI dataset, and the subsequent text briefly each identified issue. A remark on the solutions to avoid each of the issue is briefly mentioned. Note that the selection of these specific issues is based on their critical impact on the overall quality and reliability of the AI model for predictive maintenance. This decision is based on a combination of domain knowledge, best practices in AI/ML and literature reviews (Katsuki and Osogami, 2023), (Teh et al., 2020), (Northcutt et al., 2021).

3.2.1 Sensor Precision and Calibration

- Issue: Sensor measurements may have inherent inaccuracies or imprecisions.
- Solution: Implement sensor calibration procedures to enhance the precision of measurements. Regularly validate and adjust sensor accuracy to minimize errors in the dataset.

3.2.2 Labelling Errors

- Issue: Errors in labelling data instances can lead to misinterpretations by the AI model.
- Solution: Conduct a thorough review and validation of labelled data. Implement a process for cross-checking and verifying labelled instances to ensure accuracy.

3.2.3 Incomplete or Missing Data

- Issue: Incomplete or missing data points may result in an imprecise understanding of the machine's health.
- Solution: Perform a gap analysis to identify and address missing data. Augment the dataset with additional instances to ensure completeness and accuracy.

3.2.4 Outliers in Training Data

- Issue: An outlier refers to anomalies or extreme values present in the training dataset and can negatively impact model accuracy.
- Solution: Utilize outlier detection techniques to identify and handle anomalies in the dataset. Ensure that the model is not biased by rare or extreme cases.

3.3 Experiments on Dataset Accuracy Impact

To empirically illustrate the importance of AI dataset accuracy in the context of machinery functional safety, experiments are designed that highlight the impact of inaccurate or incomplete datasets on the performance of the predictive maintenance AI model. The AI model is trained using a dataset with properly labelled instances and minimal noise. This is referred to as the baseline model for the experiments.

3.3.1 Data Perturbation

Perturbations are introduced (in the baseline model) to simulate common issues in real-world datasets, as listed below

- Sensor Noise: Random noise is added to sensor measurements to simulate inaccuracies.
- Labelling Errors: Labelling errors are introduced to simulate inaccuracies in the ground truth (i.e., accurate and reliable data)
- Missing Data: Random data points are removed to simulate incomplete datasets.
- Outliers: Outliers are introduced to test the model's robustness to extreme cases.

The perturbation strength ranges from 0.1 to 0.9 indicating the proportion or percentage of the dataset that will be perturbed. For example, perturbation_strength = 0.1 implies that 10% of the dataset will be perturbed. The specific behaviour of the perturbation depends on the perturbation type, and each type of perturbation (sensor noise, labelling errors, missing data, outliers) has its own mechanism for introducing changes to the dataset.

For the experiments in this paper, perturbations introduced are: *sensor noise*: adds random noise to the features, *labelling errors*: flips a certain percentage of labels, *missing data*: introduces missing values (NaN) in a certain percentage of data and *outliers*: replaces a certain percentage of data with outliers. The combination of perturbation type and perturbation strength allows simulating different scenarios of data imperfections and evaluate how well the AI/ML model performs under such conditions.

3.4 Performance Metrics and Evaluation Scenarios

To demonstrate the impact of data accuracy and noisy data on model performance, various evaluation metrics such as accuracy, precision, recall, and F1 score are used to compare the model's performance for various evaluation scenarios. In the following, the performance metrics are defined, and the evaluation scenarios are briefly described.

3.4.1 Performance Metrics

Accuracy metric represents the overall correctness of the model's predictions. It's calculated as the ratio of correctly predicted instances to the total instances. *Precision* is the proportion of true positive predictions among all positive predictions. It indicates the model's ability to avoid false positives. *Recall (sensitivity)* is the proportion of true positive predictions among all actual positive instances. It measures the model's ability to capture all relevant instances. *F1-Score* is the harmonic mean of precision and recall. It provides a balanced measure of precision and recall, especially when there is an imbalance in the class distribution.

Apart from these performance metrics, sensitivity analysis is conducted to assess the model's robustness and performance under varying conditions or parameter changes, providing insights into how the model's predictions may be influenced by alterations in input parameters or external factors. Thus, sensitivity analysis contributes to quality assurance for data quality by revealing the impact of data variations on model performance. It facilitates the identification of potential data quality issues and guides strategies for improving both data quality and model robustness.

3.4.2 Evaluation Scenarios

The evaluation scenarios in the empirical study are outlined below.

- Baseline Evaluation: The baseline model is evaluated on a clean, accurate dataset to establish a reference performance level.
- Impact of Sensor Noise: The model's performance when trained and tested with datasets containing varying degrees of sensor noise is evaluated.
- Effect of Labelling Errors: The model's robustness to labelling errors by training it on datasets with introduced label inaccuracies is evaluated.
- Handling Missing Data: The model's performance when faced with missing data points, assessing the impact of data incompleteness is investigated.
- Robustness to Outliers: The model's ability to handle outliers by training and testing on datasets containing extreme values is tested.

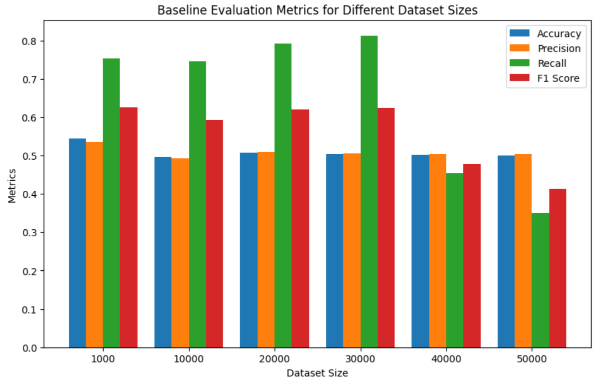

Figure 1: Baseline evaluation metrics for different dataset sizes.

3.5 Implementation

A custom-defined script is implemented in Python which uses libraries such as NumPy, scikit-learn and Matplotlib to generate synthetic datasets, perturb the data, train and evaluate logistic regression models, perform baseline evaluation, detailed sensitivity analyses varying perturbation strength and type, and visualize the results. The experiments are run on an Intel Core i7-8550U-1.8 GHz CPU, X64-based PC system running Windows 10.

4 RESULTS AND ANALYSIS

Results and analysis of the results is presented in this section, primarily based on the performance metrics and evaluation scenarios outlined in section 3.4.1 and section 3.4.2 respectively.

4.1 Baseline Evaluation

The evaluation of the model on clean and accurate datasets of varying sizes reveals interesting insights into its performance metrics, as seen in Fig. 1. The accuracy of our model, representing the proportion of correctly classified instances, ranges between 49% and 55%.

Precision, which measures the accuracy of positive predictions, demonstrates moderate values ranging from approximately 48% to 53%. This implies that when the model predicts a positive instance, it is reasonably accurate, but there is room for improvement.

Recall (sensitivity) values, ranging from 26% to 81%, indicate the model's ability to capture a substantial proportion of actual positive instances. Despite the variability, the model excels in identifying positive instances.

F1 score, a harmonic mean of precision and recall, provides a balanced assessment. The F1 scores in range from 34% to 62%, indicating a reasonable compromise between precision and recall. The high recall values suggest that the model is adept at identifying positive instances, but the lower overall accuracy implies challenges in correctly classifying negative instances. This discrepancy may stem from the model's struggle with certain negative cases.

Thus, the model used here model exhibits strengths in capturing positive instances but faces challenges in achieving higher overall accuracy. This is left intentionally unoptimized, as the primary focus of this work is on investigating the effects of perturbations on data quality. Therefore, maintaining the baseline performance as is, the performance on perturbed datasets was analysed.

4.2 Sensor Noise Perturbation

The sensitivity analysis for sensor noise perturbation for each perturbation strength across different datasets is shown in Fig.2. For a dataset size of 1000 samples, the accuracy decreases as perturbation strength increases. Precision and F1 Score show a similar trend. High recall, at lower perturbation strengths, suggests a trade-off between precision and recall. For a dataset size of 10000, similar trends are seen as the 1000-size dataset. The impact of sensor noise on accuracy is consistent across dataset sizes. For dataset sizes 20000 and 30000 the general trend of decreasing accuracy with increased perturbation strength persists. In summary, from Fig. 2, it is evident that the model's performance is consistently affected by sensor noise across different dataset sizes.

4.3 Labelling Errors Perturbation

The sensitivity analysis for labelling errors perturbation for various dataset sizes is shown in Fig. 3. For a dataset size of 1000, similar trends to sensor noise are seen, but labelling errors have a slightly more pronounced impact on accuracy. On the other hand, precision and recall are affected, showing a trade-off. For dataset size of 10000, 20000 and 30000 consistent patterns with a decrease in accuracy, precision, and recall are seen as perturbation strength increases. In summary, labelling errors have a more significant impact on performance compared to sensor noise.

4.4 Missing Data Perturbation

The sensitivity analysis for missing data perturbation for various dataset sizes is shown in Fig. 4. For dataset size of 1000, a similar pattern of decreasing accuracy with higher perturbation strength is observed. The impact is more pronounced in recall, suggesting missing data features affects model sensitivity. For data set size of 10000, 20000 and 30000 consistent trends across dataset sizes is observed. The model is sensitive to missing data, and the impact increases with dataset size.

4.5 Outliers Perturbation

The sensitivity analysis for outliers perturbation for various dataset sizes is shown in Fig. 5. For dataset size of 1000 samples, the accuracy shows a fluctuating pattern with perturbation strength. Precision and recall also show a similar pattern. For dataset size of 10000, 20000 and 30000, accuracy, precision, and recall exhibit similar trends. Outliers have a noticeable impact on the model's performance, and the effect remains consistent across different dataset sizes.

4.6 Overall Inference on Impact of Data Accuracy

The baseline model performance serves as a reference point or control and provides a basis for understanding how the model responds to different data conditions. Hence, the baseline model with about 50% accuracy is justified in the context of demonstrating the impact of perturbations on data accuracy.

The sensitivity analysis clearly demonstrates that the model's performance is sensitive to various types of perturbations. The decrease in accuracy, precision, and recall indicates that noisy data significantly affects model performance. In general, as the dataset size increases, the impact of perturbations on model performance becomes more pronounced. Labelling errors and missing data have a more substantial impact on performance compared to sensor noise and outliers. There's a noticeable trade-off between precision and recall in many scenarios, indicating the sensitivity of the model to noisy data.

The experiment underscores the importance of data accuracy in training robust ML. Noisy data, introduced through different perturbation types, consistently degrades model performance across various dataset sizes. Larger datasets provide some resilience

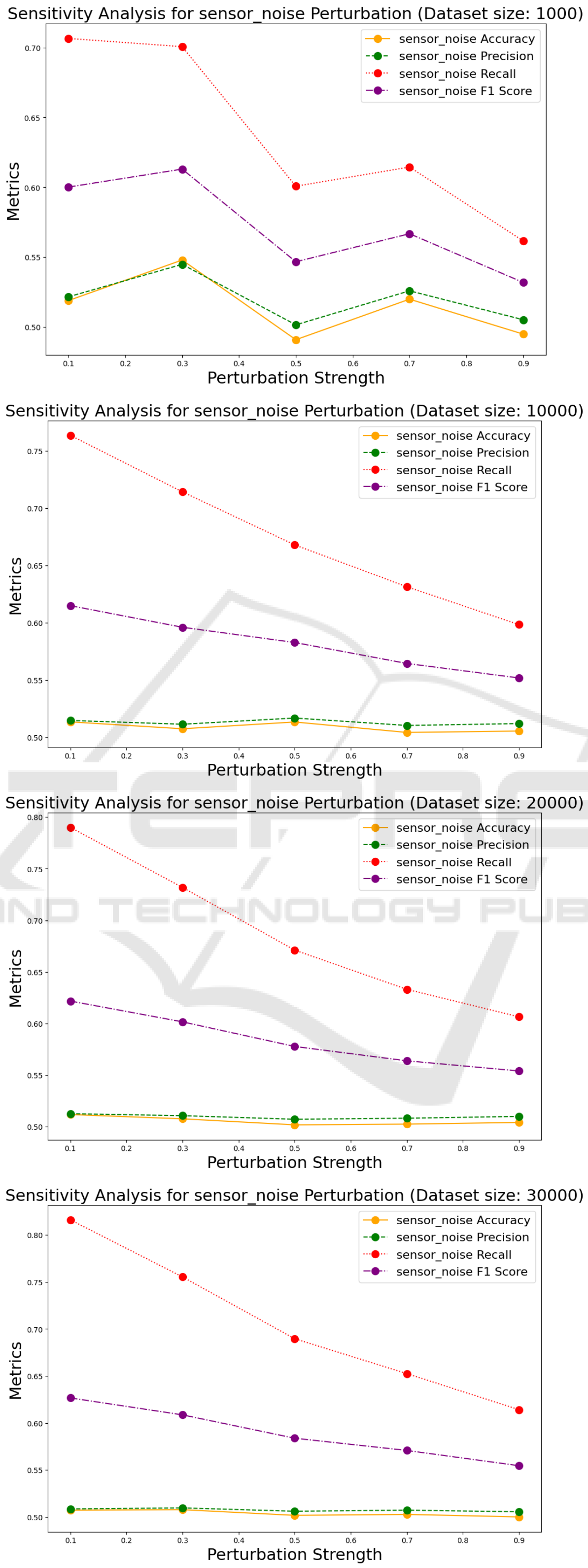

Figure 2: Sensitivity analysis for sensor noise perturbation for various dataset sizes.

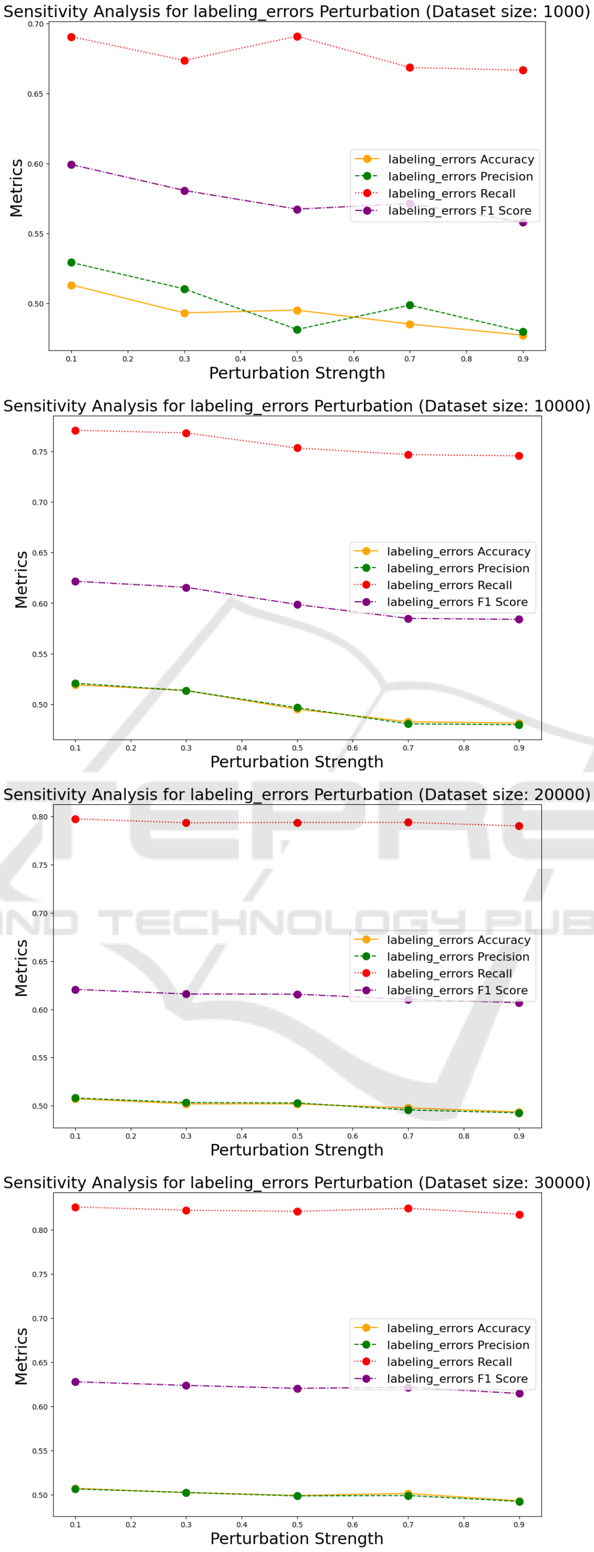

Figure 3: Sensitivity analysis for labelling errors perturbation for various dataset sizes.

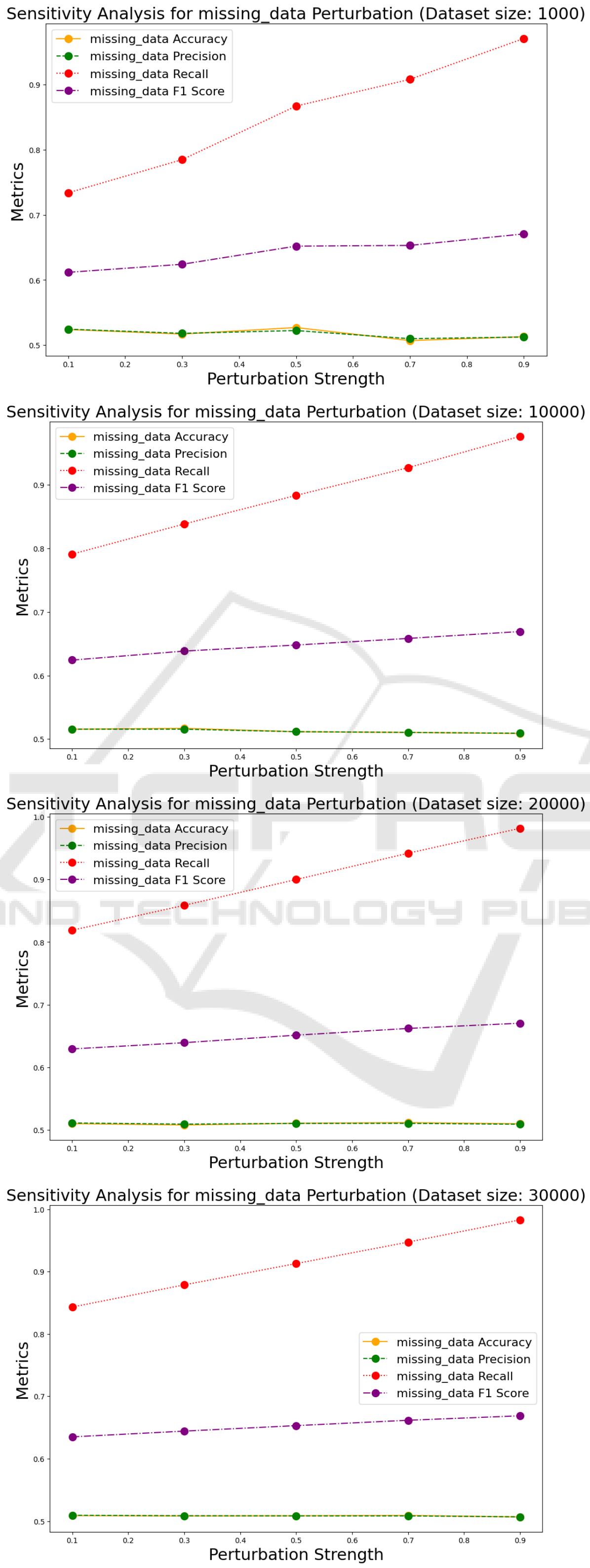

Figure 4: Sensitivity analysis for Missing data perturbation for various dataset sizes.

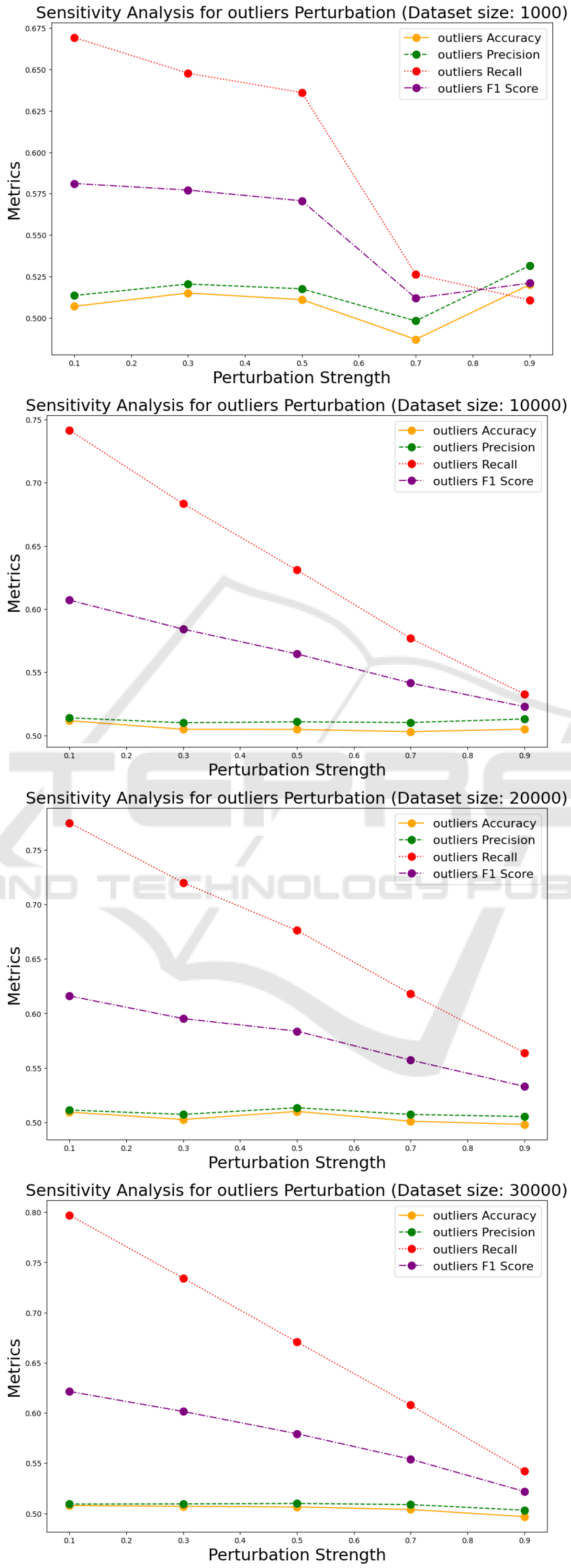

Figure 5: Sensitivity analysis for Outliers perturbation for various dataset sizes.

to perturbations, but maintaining data accuracy is still critical.

4.7 Implications on Functional Safety

The observed impact of dataset inaccuracies on the performance of the AI model has significant implications for the reliability and availability of machinery. These are crucial aspects in the context of functional safety.

4.7.1 Reliability

Reliability, in this context, refers to the ability of the AI model to consistently provide accurate predictions regarding the health and maintenance needs of machinery. The decrease in performance metrics under perturbations indicates that inaccuracies in the dataset can compromise the reliability of the AI model. Further, inaccuracies in the dataset can lead to false alarms or missed predictions, affecting the overall reliability of the predictive maintenance system. An unreliable AI model may result in unexpected breakdowns, leading to unplanned downtime and potential safety risks. Thus, to enhance model reliability, it's crucial to address dataset inaccuracies through improved data quality assurance processes. Implementing a more robust outlier detection mechanisms to handle extreme cases and to ensure the model's predictions are trustworthy could help improve model reliability.

4.7.2 Availability

Availability, in the context of machinery functional safety, refers to the system's ability to operate as expected without disruptions, downtime, or failures. Inaccurate predictions from the AI model can impact the availability of machinery, as maintenance actions may be taken when unnecessary or delayed when needed. Unreliable predictions may lead to unnecessary maintenance interventions, reducing the availability of machinery due to time spent on non-critical tasks. Conversely, if the model fails to detect actual issues due to inaccuracies, it may lead to unexpected failures, increasing downtime and impacting availability.

Thus, it is imperative to ensure that the AI model is trained on a high-quality dataset to improve the accuracy of predictions and reduce the likelihood of unnecessary maintenance actions. Further, implementing a well-defined feedback loop to continuously improve the model's accuracy over time, contributing to increased availability by minimizing false predictions and identifying actual issues promptly, is recommended.

4.7.3 Overall Considerations

Rigorous data quality assurance processes serve as a critical risk mitigation strategy. By addressing dataset inaccuracies, the overall risk of unreliable predictions and potential safety hazards is reduced. Striking a balance between accurate predictions and minimizing downtime is essential. The AI model should provide reliable insights without causing unnecessary disruptions to machinery operations. The findings underscore the importance of continuous monitoring, feedback loops, and adaptation to maintain and improve the reliability and availability of the AI system over time. The AI system should be integrated with existing safety measures and protocols to ensure that predictions align with the overall safety goals of the machinery.

Thus, addressing dataset inaccuracies is pivotal for ensuring the reliability and availability of machinery in functional safety applications. Accurate predictions from AI models contribute to efficient maintenance practices, reduce the risk of unplanned downtime, and enhance overall safety in high-risk environments.

4.8 Discussion

Factors such as sensor noise, labelling errors, missing data, and outliers were identified as contributors to the compromise of the AI model's accuracy through systematic experimentation. Perturbations simulating these factors were introduced into the datasets used for training and testing the AI model. By observing the changes in performance metrics such as accuracy, precision, recall, and F1 score under different perturbation scenarios, the impact of each factor on the model's accuracy was quantified and analysed. This empirical approach allowed for a comprehensive understanding of how inaccuracies in the dataset affect the model's predictive capabilities.

4.8.1 Importance of Dataset Quality

In the experiments, perturbations were systematically introduced to simulate common real-world challenges in datasets used for training AI models in machinery functional safety. The consistent trend of decreasing performance metrics across all scenarios underscores the paramount importance of high-quality, accurate datasets in ensuring the reliability and effectiveness of AI models.

- Sensor noise impact: The decline in performance under increased sensor noise emphasizes the vulnerability of the AI model to inaccuracies in sensor measurements. This underscores the need for precise sensor calibration and robust data preprocessing techniques to mitigate noise-induced errors.
- Effect of labelling errors: The decrease in performance when labelling errors were introduced highlights the sensitivity of the AI model to inaccuracies in ground truth information (i.e., accurate or reliable data). It underscores the necessity for rigorous data labelling validation processes to prevent misinterpretations and subsequent model inaccuracies.
- Handling missing data: The deterioration in performance due to missing data emphasizes the significance of data completeness. Incomplete datasets compromise the model's ability to understand and predict machine health accurately. Strategies for data augmentation or careful handling of missing data are crucial in maintaining dataset completeness.
- Robustness to outliers: The observed decline in performance in the presence of outliers underscores the importance of robust model architecture. Models must be designed to handle extreme cases without significant degradation in performance. Additionally, outlier detection and preprocessing steps are essential to ensure model stability.

4.8.2 Implications of Compromised Dataset Accuracy

The implications of compromised dataset accuracy are profound, particularly in safety-critical environments such as machinery operations. Inaccurate or noisy data can lead to a cascade of issues that undermine the reliability and availability of machinery, ultimately impacting operational efficiency and safety.

Firstly, compromised dataset accuracy can result in unreliable predictions from AI models. For instance, sensor noise or labelling errors may cause false alarms or missed detections of potential machinery failures, leading to unnecessary maintenance interventions or, conversely, delayed responses to critical issues. This not only affects the reliability of predictive maintenance systems but also jeopardizes the availability of machinery by increasing the risk of unplanned downtime. Moreover, inaccurate predictions may erode trust in AI-driven systems, potentially leading to scepticism among operators and maintenance personnel, further exacerbating safety risks. Additionally, compromised dataset accuracy can hinder the effectiveness of decision-making processes, as unreliable insights may prompt inappropriate actions or overlook genuine threats. These implications underscore the need for stringent data quality assurance measures to ensure the integrity and reliability of datasets used in AI-driven applications

4.8.3 Recommendations for Data Quality Assurance

The findings in this paper have direct implications for practitioners and organizations involved in the development of AI models not only for machinery functional safety but in high-risk environments across industries. The recommendations presented in this paper, although lacking extensive real-world empirical validation, are based on fundamental principles of data quality assurance and ML. These principles, such as sensor calibration, labeling validation, data completeness strategies, and outlier handling, transcend specific industries and can be adapted to diverse domains beyond manufacturing. The challenges posed by sensor noise, labeling errors, missing data, and outliers are ubiquitous in AI-driven systems, making the strategies proposed here applicable across various applications. While empirical validation would strengthen the evidence, the transferability of these strategies lies in their robustness and resilience to address common data quality issues across different contexts.

Thus, to enhance the robustness and reliability of AI systems in high-risk applications, the following recommendations are proposed:

- Investment in sensor calibration: Prioritize regular calibration of sensors to ensure precise and accurate measurements. This reduces the impact of sensor noise on the AI model's predictions and improves overall system reliability.
- Rigorous labelling validation: Establish robust procedures for validating and cross-verifying labelled data. This includes periodic audits and consistency checks to minimize the introduction of labelling errors during model training.
- Data completeness strategies: Implement strategies to address missing data, such as data augmentation or imputation techniques. Ensuring a complete dataset enhances the model's ability to accurately capture the underlying patterns in machinery health.
- Enhanced outlier handling: Develop models with enhanced outlier detection mechanisms. This includes preprocessing steps to identify and handle

outliers effectively, preventing them from negatively impacting model performance.

- Continuous monitoring and improvement: Establish continuous monitoring mechanisms to track the performance of the AI model over time. Implement feedback loops that allow the model to adapt and improve based on corrected data, maintaining accuracy in dynamic operational environments.

In summary, the experiments provide empirical evidence supporting the critical role of accurate datasets in machinery functional safety setting. By adhering to these recommendations, practitioners can build AI systems that are more resilient, accurate, and reliable, ultimately contributing to enhanced safety outcomes in high-risk applications.

5 CONCLUSION

The empirical experiments conducted to investigate the impact of dataset accuracy on AI model performance in the realm of machinery functional safety have yielded valuable insights into the critical nature of high-quality data in ensuring the reliability and effectiveness of AI systems. It is evident that inaccuracies in the training dataset lead to diminished predictive capabilities, potentially compromising the safety and reliability of machinery in industrial settings. The findings in this work underscore two key points, namely the *importance of dataset quality* and *recommendations for data quality assurance.*

Future work in this domain could explore advanced techniques for enhancing dataset quality, such as the integration of anomaly detection algorithms and robust preprocessing methods. Investigating the adaptability of the AI model to dynamic operational environments and evolving machinery conditions would be valuable. Another avenue for research involves experimentation with a combination of real-world datasets and synthetic datasets. This approach would allow for a more comprehensive evaluation of model performance and generalizability by incorporating the complexities and nuances present in real-world data, while still maintaining the benefits of controlled experimentation offered by synthetic datasets.

REFERENCES

Budach, Lukas, e. (2022). The effects of data quality on machine learning performance. *arXiv:2207.14529*. arXiv preprint, https://arxiv.org/abs/2207.14529.

IEC62061 (2022). *Safety of machinery—Functional safety of safety-related electrical, electronic and programmable electronic control systems*. IEC.

ISO13849 (2023). *Safety of machinery—Safety-related parts of control systems—Part 1: General principles for design*. ISO.

Katsuki, T. and Osogami, T. (2023). Regression with sensor data containing incomplete observations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, Honolulu, Hawaii, USA. PMLR. Copyright 2023 by the author(s).

Lwakatare, L. E., Rånge, E., Crnkovic, I., and Bosch, J. (2021). On the experiences of adopting automated data validation in an industrial machine learning project. *CoRR*, abs/2103.04095.

Northcutt, C. G., Athalye, A., and Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks.

Pedregosa, F. e. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perez-Cerrolaza, e. (2023). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Comput. Surv.* Just Accepted.

Polyzotis, N., Zinkevich, M., Roy, S., Breck, E., and Whang, S. (2019). Data validation for machine learning. In Talwalkar, A., Smith, V., and Zaharia, M., editors, *Proceedings of Machine Learning and Systems*, volume 1, pages 334–347.

Priestley, M., O'donnell, F., and Simperl, E. (2023). A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data and Information Quality*, 15(2).

Schelter, S., Rukat, T., and Biessmann, F. (2020). Learning to validate the predictions of black box classifiers on unseen data. In *2020 ACM SIGMOD International Conference on Management of Data, SIGMOD '20*, pages 1289–1299, New York, NY. Association for Computing Machinery.

Schelter, S., Rukat, T., and Biessmann, F. (2021). JENGA - A framework to study the impact of data errors on the predictions of machine learning models. In *EDBT*, pages 529–534. OpenProceedings.org.

Sessions, V. and Valtorta, M. (2009). Towards a method for data accuracy assessment utilizing a bayesian network learning algorithm. *J. Data and Information Quality*, 1(3).

Shankar, S., Fawaz, L., Gyllstrom, K., and Parameswaran, A. (2023). Automatic and precise data validation for machine learning. page 2198–2207, New York, NY, USA. Association for Computing Machinery.

Teh, H., Kempa-Liehr, A., and Wang, K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, 7:11.