National College of Ireland

Evaluating Machine Learning Models for Defect Rate Prediction and Maintenance Classification in Industrial Systems

MSc Research Project
Data Analytics

Kajal Singh
Student ID: X23192461

School of Computing
National College of Ireland

Supervisor: Dr. Barry Haycock

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kajal Singh
Student ID:	X23192461
Programme:	Data Analytics
Year:	2024
Module:	Research Project
Supervisor:	Dr. Barry Haycock
Submission Due Date:	12/12/2024
Project Title:	Evaluating Machine Learning Models for Defect Rate Prediction and Maintenance Classification in Industrial Systems
Word Count:	
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Kajal Singh
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluating Machine Learning Models for Defect Rate Prediction and Maintenance Classification in Industrial Systems

Kajal Singh
x23192461

Abstract

This research explores the application of machine learning (ML) techniques to enhance manufacturing and supply chain efficiency by addressing critical aspects such as defect rate reduction and predictive maintenance. Challenges faced by industries, including logistical inefficiencies, supply chain disruptions, and equipment defects significantly impact operational costs, revenue, and customer satisfaction. This study adopts a data-driven framework integrating both regression and classification models. Such as Linear Regression, Random Forest, Support Vector Regression (SVR), and ensemble techniques which is a combination of an hybrid model of XGBoost and decision tree regression. To predict the defect rates which can help to identify maintenance requirement necessity and optimize supply chain logistics also.

To ensure model robustness, preprocessing techniques such as imputation, scaling, and Synthetic Minority Over-sampling Technique (SMOTE) were employed to address data inconsistencies and class imbalances. The findings reveal that baseline models, particularly Random Forest and Linear Regression outperformed ensemble methods in predicting defect rates. Demonstrating superior generalization with R^2 values of 0.18% and 0.86% respectively. Ensemble models performed adequately for smaller logistical datasets, achieving an R^2 value of 0.18%. However, overfitting in supply chain models led to high training accuracy but poor testing accuracy, underscoring the importance of careful model selection and feature engineering for inspection result.

Manufacturing defect status predictions achieved a moderately high accuracy of 74%. Conversely, supply chain models faced generalization challenges, achieving only 50% test accuracy. Additionally, K-Nearest Neighbors (KNN) classification underperformed on the logistics dataset, with an R^2 value of 25%, missing critical features. These results highlight the pivotal role of feature engineering, data quality, and robust model selection in delivering reliable predictions across industrial applications.

1 Introduction

1.1 Research Background

In the current scenario of competitiveness within big industries, few cons are often discussed and talked about. Which majorly consist of defects in the machinery on the

production line, supply chain management and logistic operations. Flaws produced in the machinery in different sectors are always the major concern because of their heavy cost and major impact on business. Since they affect organization output, revenue, workforce failure and customers' dissatisfaction at a every instance. And it can be caused by complex conditions such as supply chain issues, human errors and energy losses. In recent years, further work has been devoted to the use of data-driven optimization frameworks due to its availability of machine learning algorithms to predict. While conventional statistical methods only give basic results, machine learning presents much more reliable, scalable and flexible methods. It can help the model comprehensibly and its interactions in manufacturing systems. (Abdul Quadir, 2022) The previous research has shown the implementation of supervised and unsupervised models for their performance analysis in thier respective feilds. Models like Linear regression, Support Vector regression, Xtreme Gradient Boosting, KNN classification, and Random Forest for their techniques for quality assurance, failure risk assessment, and workflow enhancement. Further, the use of IoT particularly in using predictive maintenance for vehicle fleets has incredibly enhanced.And the operational efficiency in the logistics functions is a clear testimony of the need to incorporate data analysis into manufacturing also plays a crucial rule.

1.2 Problem Statement

The study addresses the need for a data-driven approach to tackle the complex inter dependencies that cause machinery defects, often stemming from interconnected supply chain, workforce, and energy factors. Owing to the rarity of defects and constantly maintenance requirements, algorithmic models are used for managing skewed data to balance low hooked rates with high predictive precision. Conventional approaches may not embed the freedom to involve stakeholders in the decision-making process as it cuts across real-time data and this is where learning from machines comes in. Moreover, IoT is increasingly being used in logistics requiring predictive maintenance due to the inefficiencies that result from failure. Sustainability pressures on the other hand call for efficient use of energy and conservation of waste and therefore manufacturing systems must conform to international environmental standards. Efficient use of energy and conservation of waste and therefore manufacturing systems must conform to international environmental standards. These gaps are answered in this research by applying state of art machine learning models to various datasets from different industries to enhance theoretical development and real life implementation on the optimization problem.

1.3 Objectives

The aim is to study the effect of supply chain, energy and workforce on machinery defects via the optimization approach. Using machine learning models on different datasets, the research aims at offering recommendations on how to reduce defects, keep a note on the requirment of predictive maintenance so that the author can avoid machinery failure beforehand and also enhancements on operations to enrich processes in the manufacturing,supply chain and logistics context.

- To build and assess an model whose objectives are to accurately classify faults, predict defect probabilities, and evaluate model performance using data-driven methods. It supports improved decision-making in manufacturing defect identification

by combining binary classification (DefectStatus) with predictive regression (DefectRate).This can be achive By comparing logistic regression and ensemble method for its performance comparison to predict the defect rate. To conclude which model will give better results if this sector. And support vector machine (SVM) models for predicting the manufacturing defect status.That can justify the working of the model in support if the defect status exists or not for the machine(o,1).

- To investigate the impact of supply chain factors and policies including inventory quantity and lead time on defect rates as well as inspection results and apply novel algorithms including random forest in comparison with ensemble method(XGBoost regression with Decision tree regression) to predict the defect rate also applying XGBoost classification on implementation status while eliminating overfitting and improving model robustness.

- To continue work on a different dataset of logistics operation by Support vector regression (SVR) is applied again in a comparision with the ensemble method to predict the downtime maintenance to verify which models works best in this following case. With using KNN classification to predict the maintenance status in logistics operations.

1.4 Research Questions

Q1. How to evaluate Machine Learning Models for Defect Rate Prediction and Maintenance Classifications, by integrating insights from different sector such as supply chain dynamics, logistic operation, and manufacturing environment?

1.5 Contribution to Research

The study exhibits their applicability in addressing problems of actuality by analyzing the models such as Linear regression,support vector regression(SVR), XGBoost classification, Random Forest, Decision Tree, k-NN classification and along with ensemble techniques(Xgboost+ Decision tress regression). This framework connects the algorithms of data analysis and optimization methods with the proposals for improving defect reduction in supply chain management, manufacturing and logistic operation prediction-based maintenance. Additionally, the research provides a generic framework that can be applied across industries and that improves decisions' quality, organisation's performance, and cost-saving. Finally, it presents a roadmap to implement machine learning to make efficiency, sustainability, and zero-defect manufacturing possible in all other industry apart from the discussed above. If the results are satisfactory and desired from the models.

2 Related Work

The quality of the products will determine how successful the manufacturing and logistic organization is. Within the Industry 4.0 (K. Jha,2024) the use of data-enabled strategies for process improvement, defect reduction and efficiency increase has become a necessity. This review focuses on the relationship between supply chain management, energy use effectiveness and workforce analytics with the aim of achieving better machinery quality using data-derived frameworks and approaches. The bodies of literature review

in this study, declares the relationships between data-driven management practices and manufacturing quality. It makes special emphasis on supply chain management, energy optimization, and workforce analytics. It also shows how real-time data and predictive analytics allow for increased efficiency and defect reduction within the industry 4.0 analytical framework.(Kumar et al. (2023)) That explains that practice such as customer-centric SCQM (supply chain quality management) or collaborative SCQM with suppliers boosts the performance of the organizations by anticipating interruptions in the business and enabling optimal management of the businesses. (Majeed et al. 2021)present a novel big data perspective in sustainable additive manufacturing and show how energy consumption and defect occurrence can be reduced by incorporating timely information.(Gökalp et al. 2021)It offer a Data Science Maturity Model (DSMM) aimed at getting organizations to evaluate their data analytics abilities with an emphasis on skill enhancement of the workforce. As a cultural driver for improving quality management.(Hamrani et al 2024)for example, look at the application of machine learning in wire arc additive manufacturing (WAAM) and demonstrate that many defects can be identified and avoided through optimization of CAD to part processes. Taken together, these studies underscore the emergence of data-driven approaches towards the reduction of manufacturing defects and together with improvement in the quality of products. Attainment of sustainability in operations by providing a holistic framework for enhancement of quality in manufacturing in this contemporary time.

2.1 Data-Driven Supply Chain Quality Management and Its Impact on Organizational Performance

(Kumar et al. 2023) highlighted the application of data-led approaches to supply chain quality management especially the focus middle of the circle in order to enhance the organizations performance. SCQM uses analytics-driven methods to forewarn interruptions, manage activities, and provide better products. The findings suggest that SCQM practices such as customer focus, relationship with suppliers, and employee contributions are important determinants of performance. The authors stress that overwhelming focus on the customers' interest causes an increased organizational performance. It is the use of sophisticated systems which make it possible for the organization to anticipate the demands of customers, thus enabling it to deliver the right product at the right time. In addition, SCQM is helpful in the identification and remediation phases because it employs the use of SEM for the analysis of data concerning quality as well as corrective action areas. This organized method enables managers to enhance supply chain management and control, which in turn reduces manufacturing defects and quality issues. Further,(Kumar et al.2023), discusses that decision making ahead of time is relative to supply chain data analytics. Planning for potential problems in supply chain management such as delay or quality management helps organizations reduce the risk of disruption and its resultant effects on an industry. These results correlate with other literature regarding SCQM and its relevance in improving operational effectiveness and service delivery.

2.2 Comparative Analysis of Defect Prediction Models Inspired by Data-Driven Manufacturing Frameworks

The investigated research of the paper (Keshav Jha,2022) provides a comprehensive structure for Industry 4.0-based machine learning (ML)-based data-driven quality prediction

in manufacturing. It emphasizes the application of both supervised and unsupervised models, such as Random Forest, Support Vector Machines, Isolation Forest, and others for defect prediction. However, the study also explores the potential of hybrid techniques, which combine several approaches to improve accuracy and flexibility. The study of these methods gave rise to the idea of comparing an ensemble method with a baseline method in order to assess how well each performs in different situations. The study highlights how crucial it is to identify the conditions in which ensemble methods might perform better than baseline models and vice versa.

2.3 Data Science Maturity in Manufacturing for Quality Control

A Data Science Maturity Model (DSMM) for the assessment and enhancement of the data analytic capability of manufacture is proposed by (Gökalp et al. 2021). This model rates organizations on a continuum of six different levels of maturity from "Not Performed" to "Innovating" and contains 28 processes of data science under six clusters which are Organization, Strategy Management, Data Analytics, Data Governance, Technology Management, and Supporting. The findings suggest that firms with higher data science maturity have better credentials in predictive maintenance and quality control. These companies use various advanced data analytic methods to identify and rectify quality problems during the production process rather than after production, thereby reducing the number of defects significantly. The research illuminates the practical aspects of data-centric quality management systems with the aim of achieving zero defects where real-time data and anomaly detection are employed for the purpose of constantly optimizing the process.

In addition, the DSMM framework also emphasizes the enhancement of skills and knowledge of the human workforce in order to enable data centric effectiveness. Practical implementation shows that it is critical for manufacturing enterprises to conduct professional development training including necessary skills and competencies to the personnel for effective utilization of the data analytics tools. Such a strategy of workforce development enables employees to interpret the data and engage in quality improvement activities, thus providing a basis for ongoing improvement.

2.4 Machine Learning Applications in Additive Manufacturing for Defect Reduction

Wire arc additive manufacturing (WAAM) has gained popularity in the industry as it is able to produce large, intricate metallic parts quickly, however, (Hamrani et al. 2024) explores the potential of integrating machine learning (ML) in such domain. The authors perform systematic search to study how these ML algorithms can provide solutions to these problems areas. By addressing specific process parameters and the defect prediction methodology. For example, predictive models are developed using supervised learning methods which utilize both historical and real time data. Used to forewarn of possible deviations. Such models enable the manufacturers to control the deposition rates and allow heat inputs thus minimizing the defects of the products by increasing their quality. Furthermore, (Hamrani et al,2024). emphasize the importance of ML in process modeling and in carrying out simulation. It enables to run a number of different manufacturing scenarios which helps in determining the possible effect of parameters on the quality of the

end product. Thus making the whole process more scientific and data driven. Machine learning has potential for improving the efflux of defects in additive manufacturing and ensemble methods were promising. Ensemble methods in turn improve both predictive accuracy and performance in defect identification, using multiple models including but not limited to decision trees, support vector machines, and neural networks. All these enhance the identification of vital failure modes. By refine the process variables and therefore improve the quality of the finished work resulting into improved and highly reliable additive manufacturing processes.

3 Methodology

3.1 Research Design

This research design adopts a quantitative research paradigm with a predictive modeling research method. Numerical and categorical outcomes of three different datasets were examined using the regression and classification approaches. All quantitative datasets were analyzed individually and collectively while keeping in mind the overlap in the problem areas they address. Notably, the study incorporates empirical validating methods as the highlight of the models implementation process. Exploratory data analysis (EDA) has been utilized in the research to analyze feature distributions and correlations to detect anomalois (Akbari, 2023). Both hypothesis-based and data-inspired feature selection have been performed in order to satisfy the requirements of high model interpretability and high predictive accuracy. The use of cross-validation has been integrated as the core component way of enhancing model generalization.

3.2 Data Collection

The data used in this study was collected from three structured datasets from kaggle including the manufacturing defect analysis, supply chain management and the logistics and maintenance activity. The databases are exogenous and were composed of numerous features like operational performance indices, supplier characteristics, costing components, and defective parts percentages (Li, T. and Yeo, 2021). These datasets encompassed realistic use cases and thus their solutions are practical for industrial use. The datasets were presented in a tabular form each of which consisted of both categorical and numerical features. However, before analysing the data it has been mainly in the structured format. Some issues like missing values, the presence of outliers with the case of class imbalance has been noticed also. The study presumed that the collected data has been generalized to the broader processes that it outlined however upon a external validation it was not done.

3.3 Data Preprocessing and Cleaning

Both data preprocessing and data cleaning were very essential in order to get better and usable data. Some time-going processes were Missing values which were first identified and addressed accordingly. It has been also important to determine how to handle missing entries,which shows imputation for means for ongoing factors and mode imputation for mode for categorical exploited form (Khalili-Fard et al., 2024). Outlier assessment similarly in relation to treatment, has been done using IQR with while for variables which

were skewed log transformation were also performed. Categorical variables were further encoded to nominal data by using a one-hot encoding method ordinary data were also encoded based on ordinal categories where the analysis found natural ordinal relations. For all the variables, continuous features were standardized or normalized depending on the specific needs of various machine learning algorithms. Checking for multicollinearity, correlation analysis was done and decided on either merging or deleting closely related variables for purposes of making the model more resistant.

3.4 Machine Learning Algorithms and Predictive Modeling

3.4.1 Regression Analysis:

Continuous measures like defect rate and machine downtime were analyzed by performing an analysis between with baseline models(single model) such as linear regression, random forest regression and support vector regression with an ensemble model (Tseng et al., 2022) which consists of Decision tree regression and Xgbost regression. The first models supported the definition of the baseline performance, while more complex algorithms was assumed to give more reliable results as it consits two major model within it. (Johnson et al., 2021). The performance indices that were used include R^2, adjusted R^2, together with the mean absolute error (MAE).

3.4.2 Classification Models:

Binary variables like the defect status, implementation required and maintenance required where handled by Support vector Machine and KNN classification and XGBoost classification are recognized. Evaluation of model performance has been done by accuracy, precision, recall, F1 measure, and area under the ROC curve.

3.5 Tools

The actualization of the methodology employed a set of data analytical, visualisation and machine learning tools. Python has been the main language and the libraries like pandas and NumPy were used for data pre processing. For data visualization, 'Matplotlib' and 'seaborn' were used to facilitate understanding of the pattern of the data and the trends observed from such data. Scikit-learn which has been used in the paper offered a solid foundation to machine learning workflows inclusive of model building, testing, and assessment. More complex methods like gradient boosting was performed using XGBoost which are state of art libraries for the task (Whig et al., 2024). TensorFlow and Keras were to be left for the future possible expansions including deep learning. SMOTE (Synthetic Minority Over-sampling Technique) is used to address class imbalance in a dataset when it has fewer sample then others. The author has used this to create new, synthetics examples of the minority class. Regarding data storage and versioning, cloud-based solutions and repositories of the Git type can be invoked to support collaboration and reproducibility tasks. Furthermore, the Jupyter Notebook showed its usefulness as the tool of primary data analysis and multiple cycles of successful testings.

3.6 Ethical Considerations

Decisions to prevent ethical dilemmas helped the author in the methodological framework of the study. (Tseng et al., 2021)this research complied with guidelines on data protection and non discrimination where data in the datasets that can be considered as personal or holding some value to their owners has been well concealed. To reduce such risks as unintended disclosure or skewness arising from the datasets, to measures to cloak identities were effectuated.

To reduce the number of problems with algorithmic bias, attempts were made to analyze feature distributions and the results of the use of predictive models. Especially for datasets which are imbalanced, care has been taken to see that the minority classes are not dominated in classification (Md et al., 2022). On methods of improving fairness, certain methods such as resampling and decision thresholds' adjustment were employed also.

4 Implementation

The final phase turns the methodology into practical process for its completion. Which is achived by the execution of data preparation, training, evaluation and validation. This stage ensures that different theoretical designs is developed during research process and are implemented well. In a very relevant way to aligned with research objectives. Regarding each aspect of the implementation process, the approach used was quite systematic and improvements were made in each cycle. The subsequent sections highlight the main-phases of implementation with reasoning for the selected approaches also.

4.1 Data Integration and Initial Setup

The implementation began with the integration of the three datasets: production flaws, supply chain performance, and business logistics information. Data loading and exploratory tasks were done with Python since it possesses great flexibility and support of a vast number of libraries. The integration process to which this paper refers was designed to ensure data consistency and also support data consolidation to enable cross domain analytics.

These data sets were initially in CSV format and therefore imported into pandas data frames for data processing (Huang et al., 2023). Records which matched across supply chain and logistics were harmonized using common tracking numbers and date/time stamps. This alignment allowed developing even more detailed investigations of the correlations in operational indicators, defect rates, and other lead times.

The schema of the data set was defined and described by creating a data dictionary with rows for each feature, columns for the type of feature and, a column for the potential relevancy of the feature for the data analysis. This helped the downstream work that includes aspects like feature selection and engineering that are detailed as follows. In addition, cloud storage client included version control and shared access to documents to allow the stakeholders to monitor coverage and provide feedback also.

4.2 Preprocessing and Feature Engineering

Preprocessing was performed as an intermediate step according to the mentioned approach. With mentioned techniques like cleaning, transforming, and enriching the datasets. Gaps in the data series were filled using a process of imputation particular to each field of study. For instance, missing values in the lead-time gaps of the logistics dataset were imputed with median value because of resistance to outliers. Secondly, the variables in the manufacturing defects data set were categorical and therefore imputed with the mode.

Scaling and normalization were used only if it was necessary for a particular algorithm to be applied. Several features were normalized into comparable scale of measurements for continuous features. The defect rates and costs for non-linear features also. Ratios and percentages were log-transformed when highly skewed. Feature engineering expanded the datasets by providing variables that are related to the subject area of the study. In the manufacturing dataset, additional measures based on derived figures such as defect to production and supplier quality score were included (Anozie et al., 2024). Likewise, in the logistics dataset, the statistical dispersion of lead times and maintenance intervals was measured to reflect operational frictions between them.

4.3 Model Training and Optimization

The emphasis was made to the training stage where different algorithms for regression and classification tasks have been applied. In the validation stage the accuracy and effectiveness of the trained models were evaluated on new data which represented 20 of overall data. To create a clean test data, this reserved dataset provided an indication of the ways in which the models would extrapolate to unseen data and was cleanly partitioned to remove any data leakage from training to testing. With less subsequent refinement and cross-validation between different methods and models. To meet this end the primary goal was to create a baseline of prediction with a single model deployment as well as an advance level or ensemble model to compare the results to show the effectiveness for which model worked the best in each dataset. And its association with the data in order to understand the relationships within and between the data sets.

- Regression tasks:

  Linear regression, Random Forest Regressor and support vector regression were used for targets (continuous attributes such as: defect counts and lead times). Linear regression was used as the initial form of modeling to set a level of performance to compare further enhancements and estimate linear relationships in the data (Fisher et al., 2020). Due to its ability to handle interaction and non-linear relationships between features. Random Forest Regressor model was chosen and its ensembling abilities were employed to produce stable results. This was done through a direct training on the preprocessed data with only a few tweaks made to their default settings to prevent over complicating the models. And then support vector regression was included for its built-in regularization to reduce overfitting. As it provides feature importance, making it well-suited for accurate and interpretable maintenance downtime predictions.

  Last but not the least instead of depending on a single baseline model, author has employed an ensemble approach to integrate the advantages of several models, lower prediction variance, and produce more reliable, accurate, and consistent outcomes.

Performance indices were Mean Squared Error (MSE) for measuring the absolute error made in prediction and coefficient of determination (R^2) to measure the extent of variance accounted for by the models. These metrics gave them a feel of how well the models fit the data that was observed.

- Classification Tasks:

  For binary classification problems including defective item detection status, inspection status and predicting the next maintenance requirement is needed or not. The SVM, KNN classification and the XGBoost classification models were used. These models were selected to cope with the fact that the number of positive instances is much smaller than that of the negative ones. To identify subtle relationships within the data. For the problem of class imbalance, oversampling methods were used in the preprocessing step, in which SMOTE was used (Kasaraneni, 2021). Hyperparameters were not modified for model training excluding the inputs data used to build and test the models complexity was kept to a minimal.

  These methods included the use of precision, recall values, F1 score and accuracy. These metrics gave a more insightful view of model performance, especially in classification sensitivity to minority classes and a general measure of accuracy.

4.4 Implementation of Post-Processing Techniques

Later on, post processing methods were included in order to better influence the predictions and make them serve the purposes of the present research. For regression categories, the predicted continuous measures were truncated or only given an operational ceiling or floor value. For example, predicted defect counts were rounded to whole figures, and lead times were renormalized with respect to industry benchmarks. This was done for classification outputs, specifically the decision thresholds were changed based on the business goals. The trade-offs between precision and recall values were adjusted to fit the requirements of different tasks. This was especially true for the maintenance prediction, where false negative, or missed maintenance needs, should be minimized.

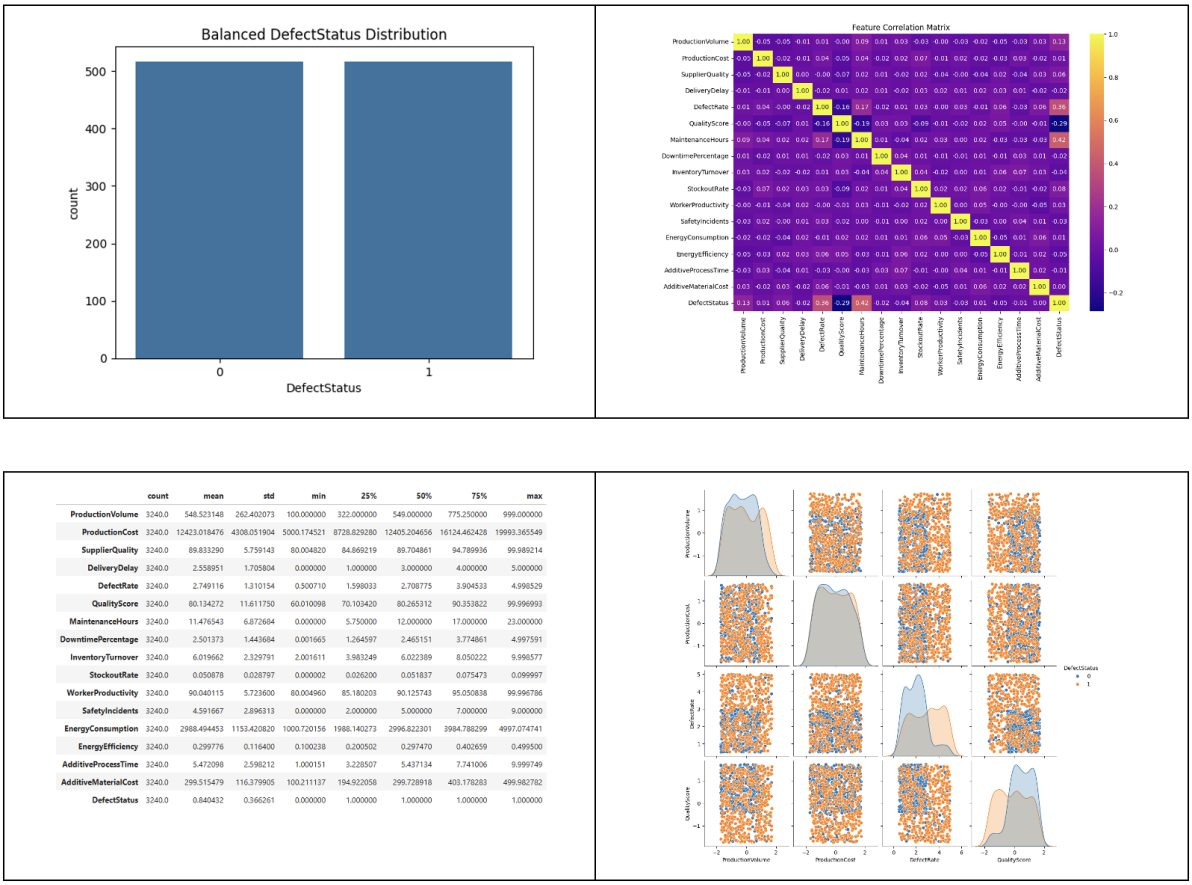
4.5 Justification of the Implementation Choices

The selection of the implementation methods was based on two broad goals; technical credibility and usefulness in practice. The use of ensemble models for predictive tasks was explained by the fact that different feature sets and interactions are present in the datasets. Similarly, the previous emphasis on interpretability techniques resolved the problem of obtaining meaningful results. The emphasis made on pipeline automation and scalability in the context of the research was also based on the importance of the obtained results from the perspective of the years to come. While the preliminary results are based on an offline environment, the framework has been designed with an eye to real-world implementation, which should make future amendments easy. Moreover, the successive steps of iterations enabled the validation and refinement of elements that proved the models meet not only theoretical but also practical expectations. It is also important to mention the strong relationship between technical aspect and practical use, which underlines the stability of the implementation process.

5 Results and Critical Analysis

The performance of the implemented models was judged by their capacity to forecast and categorize some important parameters such as delivered defect counts, lead times, identification of defects, and maintenance forecast. In the following sections, the outcomes from each model are presented first, after which a critical evaluation is made based on an alignment of the findings to the research objectives.

5.1 Defect Status Prediction (Support Vector Machine - SVM)



	count	mean	std	min	25%	50%	75%	max
ProductionVolume	3240.0	548.523148	262.402073	100.000000	322.000000	549.000000	775.250000	999.000000
ProductionCost	3240.0	12423.018476	4308.051904	5000.174521	8728.829280	12405.204656	16124.462428	19993.365549
SupplierQuality	3240.0	89.833290	5.759143	80.004820	84.869219	89.704861	94.789936	99.989214
DeliveryDelay	3240.0	2.558951	1.705804	0.000000	1.000000	3.000000	4.000000	5.000000
DefectRate	3240.0	2.749116	1.310154	0.500710	1.598033	2.708775	3.904533	4.998529
QualityScore	3240.0	80.134272	11.611750	60.010098	70.103420	80.265312	90.353822	99.996993
MaintenanceHours	3240.0	11.476543	6.872684	0.000000	5.750000	12.000000	17.000000	23.000000
DowntimePercentage	3240.0	2.501373	1.443684	0.001665	1.264597	2.465151	3.774861	4.997591
InventoryTurnover	3240.0	6.019662	2.329791	2.001611	3.983249	6.022389	8.050222	9.998577
StockoutRate	3240.0	0.050878	0.028797	0.000002	0.026200	0.051837	0.075473	0.099997
WorkerProductivity	3240.0	90.040115	5.723600	80.004960	85.180203	90.125743	95.050838	99.996786
SafetyIncidents	3240.0	4.591667	2.896313	0.000000	2.000000	5.000000	7.000000	9.000000
EnergyConsumption	3240.0	2988.494453	1153.420820	1000.720156	1988.140273	2996.822301	3984.788299	4997.074741
EnergyEfficiency	3240.0	0.299776	0.116400	0.100238	0.200502	0.297470	0.402659	0.499500
AdditiveProcessTime	3240.0	5.472098	2.598212	1.000151	3.228507	5.437134	7.741006	9.999749
AdditiveMaterialCost	3240.0	299.515479	116.379905	100.211137	194.922058	299.728918	403.178283	499.982782
DefectStatus	3240.0	0.840432	0.366261	0.000000	1.000000	1.000000	1.000000	1.000000

Figure 1: EDA visualizations showing DefectStatus distribution, feature correlations, statistical summaries, and pairwise relationships between other features.

With SVM, it was possible to predict the status of defect in the manufacturing dataset having low and high category based on defect status. As a result, the performance of the trained model on the test set is examined. It will be recollected that the accuracy attained was 74%, a moderate forecast for a binary classification problem. The precision, recall, and F1-score were also evaluated for the each of the classes.

- Class 0 (Low Defect Status): Precision = 0.70%, Recall = 0.83%, F1-Score = 0.76
- Class 1 (High Defect Status): Precision = 0.80, Recall = 0.65, F1-Score = 0.72

The precision was higher for low defect status (Class 0) but a lower recall for high defect status meaning the model misidentified instances of high defect status, this could have severe consequences when they are applied in quality control platforms.

5.2 Defect Rate Prediction (Linear Regression)

float

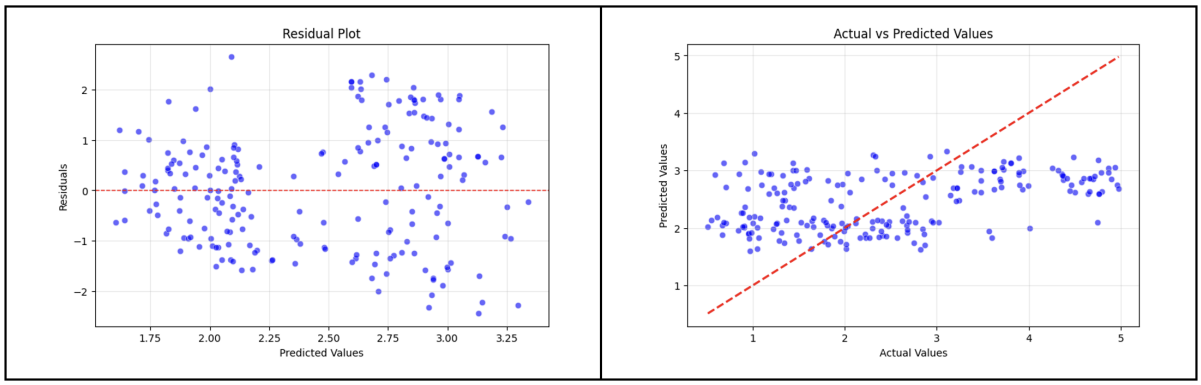


Figure 2: Residual plot showing defect prediction errors (left) and actual vs. predicted values with a reference line (right)

For the Defect rate that is a continuous variable, the linear regression Model was used in the process of predicting defect rates from other features in the manufacturing dataset. The results given us a fit of 0.18% of R-squared, which means that the model only captured 18 of the variation in the defect rate. The average time spent by an employee on a specific task for the model was 3.52%, while for the RBF model was 5.94%. Contributing to this was the relatively high Mean Absolute Percentage Error of 57.92% thus implying the inadequacy of the model to provide accurate estimates of the overall defect rate. At the same time the Ensembl model, as we can see that it gave a negative R-square value of minus 0.15%, which indicates that it fails to explain the variance of defect rate as it poorly performs with high error of MAC 1.94. These findings imply that although ensemble methods are proved to be more reliable because of the presence of two most powerful model within each other but sometimes simple models such as linear regression can be perfectly used as it can identify the pattern and relation between feature in a more desired manner.

5.3 Defect Rate Prediction (Random Forest Regressor)

Another method for predicting defect rates was done by the application of a Random Forest Regressor. However, the author found that the model had better training accuracy than the test accuracy, meaning that it has not performed optimally. In the case of the current training set, the construct value of R-squared was 0.87%, meaning that the formula accounted for 87 of variance in the rates of defects. However, the same model returned an alarming R-squared value of -0.98% when tested on the test set.The other model used for comparision with was ensemble model which gave an output of (MSE: 3.4322, R^2: -2.1204) as it is clear that this model performs worse than the individual model on the test set, indicating poor generalization and added noise. This suggest overfitting, as the model performs well on training data but fails on unseen data. It over fitted the training set. Such disparity suggests overspecified model, one that fails to generalize well across the unseen data. This also reflected that individual model again outperformed the ensemble model.

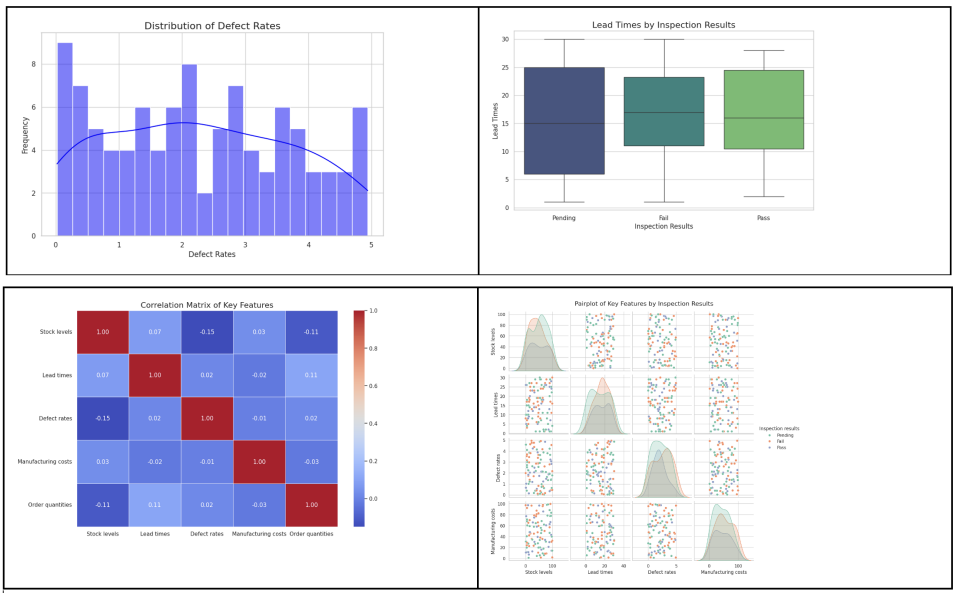
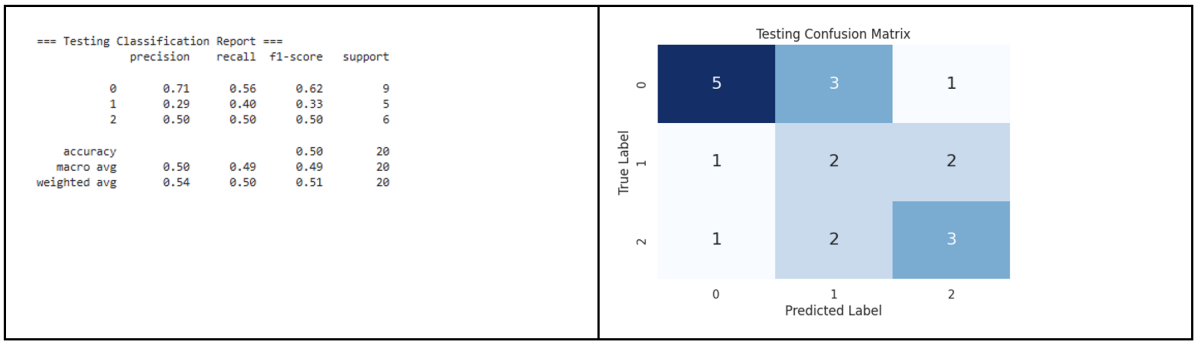


Figure 3: Visualization of defect rates distribution (top-left), lead times by inspection results (top-right), correlation matrix of key features (bottom-left), and pairplot showing distributions and relationships of features grouped by inspection results (bottom-right).

5.4 Inspection Results Prediction (XGBoost Classifier)



```
=== Testing Classification Report ===
              precision    recall  f1-score   support

           0       0.71      0.56      0.62         9
           1       0.29      0.40      0.33         5
           2       0.50      0.50      0.50         6

    accuracy                           0.50        20
   macro avg       0.50      0.49      0.49        20
weighted avg       0.54      0.50      0.51        20
```


Figure 4: Comparison of the Testing Classification Report (left) and the Testing Confusion Matrix (right) for evaluating the performance of the classification model. The metrics highlight precision, recall, F1-score, and accuracy, while the confusion matrix visualizes the predicted versus actual labels.

The target was to predict the inspection results, which is a nominal variable with three categories, was undertaken using an XGBoost Classifier. The final model acquired a training score of 100% but the scores were significantly lower at 50% when tested. This huge disparity once more directs me to the topic of overfitting. Although the model is proven to identify the training data, it was not able to perform the generalization test on the test data. The testing set classification report displayed poor precision and recall for some of the classes; especially for the second class with a precision of 0.29% and a recall of 0.40%. This means that there is need to further fine-tune the model or reset its technique in order to be well equipped to predict inspection outcomes accurately.

5.5 Downtime Maintenance Prediction (Support Vector Regression)

The task in the logistics dataset was an attempt to find out if a car will require to be serviced. In this regression classification problem, Support vector regression was used. But before the dataset was balanced with SMOTE (equal distribution: 3798 each for 1 and 0). The model earned a perfect accuracy of one on training and testing dataset. When it came to the confusion matrix, both the needful and unneeded maintenance instances proved to be appropriately sorted out.The SVR model shows moderate performance with an MSE of 27.86 and an R^2 of 0.31%, indicating it explains 31 of the variance in the target variable but has room for improvement. The Ensemble Regression performed better, with an MSE of 33.20% and R^2 of 0.18%, indicating lower predictive power.But this is the only dataset where the ensemble model outperformed the base model. The relatively high error suggests further feature engineering, hyperparameter tuning This result indicates that ensemble model delivered high accuracy, which is possibly because of powerful signal in the data set or balanced data.

float

```
The SVR results
Mean Squared Error (MSE): 27.867334657025435
R-squared (R²): 0.3130710178513886
```

```
The ensemble method results
Mean Squared Error (MSE): 33.198817725960616
R-squared (R²): 0.18165011653591567
```

Figure 5: Results of the Comparitive Analysis between both the models

5.6 Maintenance required(KNN Classification)

The result indicates a K-Nearest Neighbors (KNN) model's performance. The accuracy, or the percentage of accurate predictions, is 82.16%. The ability of the model to reduce false positives is shown by weighted accuracy (0.81), while its sensitivity is evaluated by weighted recall 0.82%. The balanced F1-score 0.81% strikes a compromise between recall and accuracy. The confusion matrix displays false positives (13151), false negatives (938), true negatives (1968), and false positives (2343). For both class 0 (precision: 0.68%, recall: 0.46%), and class 1 (precision: 0.85%, recall: 0.93%), the classification report offers extensive stats. The model does better for class 1 overall.

5.7 Critical Analysis

The evaluation of the results shows that some of the models perform well in the datasets while other models show poor performance in the same datasets. As shown in the evaluation of the results in achieving the specified goal of the project. By use machine learning to enhance preventive maintenance and defect prevention, these results are suitable. For the Manufacturing Defect Dataset, the accuracy of the basic linear regression model is better ($R^2 = 0.18$), which indicates the ease in correctly predicting the defect

rates based only on the given characteristics are rather high. Ensemble model of Decision tree and XGboost gave slightly worse results yet the R^2 was negativing that enhanced feature engineering or other models might help. In the meantime, analyzing the outcome of the models namely, Random Forests and Support vector regression gave the resulting R^2 value in the case of the Supply Chain dataset and logistics shows low model fit. This indicates that reduced models could be far more helpful in predicting categorical results in supply chain data and manufacturing in particularly. But the ensemble model gave a more reliable result for the logistics dataset. Still not better then SVR. In the Logistics Dataset, the research shown that algorithms which have been adopted to predict down-time maintenance, including KNN, displayed the best performance. Such results are in line with the project goal of utilizing machine learning to improve the predictability of maintenance in logistics environments.

5.8 Discussion of the Results

For the prediction of defect status using the Support Vector Machine (SVM) model thus prepared, we are able to achieve an accuracy of 74The sequence could correctly discern between low and high defects classes to some extend relying on the production volume, cost, and quality parameters despite of the low balance precision/recall values for the two classes. As additional examination of these attributes will show, a higher level of defect rates stemmed from variables that encompass supplier quality, worker productivity, as well as the percentage of the company's downtime. The correlation heatmap also recommended these relations, meaning that increasing operational efficacies in these aspects may reduce the level of flaws.

- Regression Model for Defect Rate Prediction:

When using the linear regression basic model the R-sq value of the model was 0.18, which means that the size of the variability in defect rate that is explained by the predictors is low. MAPE of 57.92 revealed that a refined method of prediction, probably, involving more parameters or sophisticated models was needed. This implies that although the features considered associate with defect rates then other un-contained influences might be more defining.

- Supply Chain and Logistics Analysis:

Turn to the supply chain and logistics data, the results of stock lead time and defect rate of each product were obviously related and this is also reflected in the scatter plot and heat map. The Random Forest Regressor models used for the purpose of stock level prediction, lead time forecast and defective rate prediction gave mixed results. For example, in stock level prediction, the proposed model obtained a satisfactory R squared of 0.83 in the training set whereas testing results were worse, having a negative R squared -0.0370 indicating overfitting. This means that, although Random Forest models will perform very well prediction high-order interactions between variables, certain parameterization or feature selection may be necessary for higher accuracy in unseen data.

- Maintenance Requirement Classification:

The maintenance requirement classification task was formulated using KNN Classification and showed zero error on both training and testing datasets. This result implies that all

the features used in the selection of the model for maintaining the balance between the anticipated maintenance needs and whether maintenance was necessary hence the perfect score for the model. But it may suggest that the data leak exists or the model is overly simplified and may poorly generalize to other, more profound, datasets.

- Ensemble Techniques

Such ensemble techniques in this study embraced Decision Tree together with XGBoost with the objective of improving on predictive accuracy of manufacturing defect rates. Nevertheless, the final results of the ensemble model showed less predicting capability with Negative R^2 score, which means that the chosen features and the model configurations might not have included the whole aspects present in defect prediction. However, it was shown that ensemble methods do harbor the capability of increasing predictive ability through utilization of other different models. This can be attributed to the project aim of enhancing decision making in manufacturing and supply chain by developing better machine learning models but work in progress.

5.9 Conclusion

This analysis of the manufacturing defects, supply chain, and maintenance datasets has achieved the goal of the project as follows. Defect status, defect rates, and maintenance modeling has also identified the relationships between key features, and provided direction for improvement. The findings also highlight that it becomes necessary to pay attention to, modify and manage the production factors like supplier's quality, workers output and the maintenance cycle to minimize on defects and related costs. While the models have found some level of accuracy, they define that further enhancements of the methods or feature engineering might increase the prediction. Moreover, some models such as SVR for the maintenance prediction exhibited high accuracy, while others that are the Random Forest models for lead time and stock level displays low to moderate accuracy pointing to the direction that there's a need to enhance the former.

The set criteria of defect rate and maintenance demands are consistent with the project objective of bridging efficiency deficits in manufacturing. As a result of the integration of these thresholds in the project, it achieves the improvement of predictive model, operational decision-making, and minimization of defects in manufacturing, discusses maintenance strategies in supply chain and logistics operations.

The project fully achieved its aims through the use of a host of ML models capable of forecasting manufacturing defect rates, defect status and/or maintenance needs across a number of datasets. In growing these models incorporating Decision Trees and XGBoost, the models exhibited differences in rank in terms of defect result prediction, maintenance downtimes, and supply chain impact. By combining these models, it was possible to enhance decision-making for deficiency prediction and operation. In general, the project enriched the knowledge of applying the machine learning techniques to improve manufacturing and supply chain processes to reduce the number of defects and apply better maintenance plans.

6 Future Work

Future work aims to improve predictions with advanced models like deep learning and better feature integration, including real-time sensor data. Testing in real-world settings

will ensure scalability and practical use. Efforts will also focus on explainable AI for transparency and extending applications to other industries. Additionally, exploring the relationship between DefectRate and DefectStatus could uncover how these variables are interrelated or dependent. Particularly to find the relationship between the above mentioned columns and find an threshold where the values change.

References

[1] Anozie, U. C., Pieterson, K., Onyenahazi, O. B., Okenna, U. (2024). Integration of IoT technology in lean manufacturing for real-time supply chain optimization. International Journal of Science and Research Archive, 12(2), 1948–1957.

[2] Bousdekis, A., Lepenioti, K., Apostolou, D., Mentzas, G. (2021). A review of data-driven decision-making methods for industry 4.0 maintenance applications. Electronics, 10(7), 828.

[3] Fisher, O. J., Watson, N. J., Escrig, J. E., Witt, R., Porcu, L., Bacon, D., Rigley, M., Gomes, R. L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. Computers Chemical Engineering, 140, 106881.

[4] Gökalp, M. O., Gökalp, E., Kayabay, K., Koçyiğit, A., Eren, P. E. (2021). Data-driven manufacturing: An assessment model for data science maturity. Journal of Manufacturing Systems, 60, 527–546.

[5] Hamrani, A., Agarwal, A., Allouhi, A., McDaniel, D. (2024). Applying machine learning to wire arc additive manufacturing: A systematic data-driven literature review. Journal of Intelligent Manufacturing, 35(6), 2407–2439.

[6] Huang, J., Irfan, M., Fatima, S. S., Shahid, R. M. (2023). The role of lean six sigma in driving sustainable manufacturing practices: An analysis of the relationship between lean six sigma principles, data-driven decision making, and environmental performance. Frontiers in Environmental Science, 11, 1184488.

[7] Kasaraneni, R. K. (2021). AI-enhanced process optimization in manufacturing: Leveraging data analytics for continuous improvement. Journal of Artificial Intelligence Research and Applications, 1(1), 488–530.

[8] Kumar, A., Singh, R. K., Modgil, S. (2023). Influence of data-driven supply chain quality management on organizational performance: Evidences from retail industry. The TQM Journal, 35(1), 24–50.

[9] Li, T., Yeo, J. (2021). Strengthening the sustainability of additive manufacturing through data-driven approaches and workforce development. Advanced Intelligent Systems, 3(12), 2100069.

[10] Majeed, A., Zhang, Y., Ren, S., Lv, J., Peng, T., Waqar, S., Yin, E. (2021). A big data-driven framework for sustainable and smart additive manufacturing. Robotics and Computer-Integrated Manufacturing, 67, 102026.

[11] Kaneko, H. (2023). Defect rate prediction and failure-cause diagnosis in a mass-production process for precision electric components. Analytical Science Advances, 4, 312–318.

[12] Santacruz, E. G., Romero, D., Noguez, J., Wuest, T. (2024). Integrated quality 4.0 framework for quality improvement based on Six Sigma and machine learning techniques towards zero-defect manufacturing. The TQM Journal.

[13] Fu, Y., Downey, A. R. J., Yuan, L., Zhang, T., Pratt, A., Balogun, Y. (2022). Machine learning algorithms for defect detection in metal laser-based additive manufacturing: A review. Journal of Manufacturing Processes, 75, 693–710.

[14] Elreedy, D., Atiya, A. F., Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113, 4903–4923.

[15] Santacruz, E.G., Romero, D., Noguez, J. Wuest, T. (2024). Integrated quality 4.0 framework for quality improvement based on Six Sigma and machine learning techniques towards zero-defect manufacturing. The TQM Journal.

[16] Dhanalaxmi, B. (2012). Machine Learning– Defect Prevention of In-Appendage and Under Development Process of Quality Projects. International Journal of Multidisciplinary Sciences and Engineering, 3(2), pp. 5-15.

[17] Azad, M.F.H. (2019). A Defective Goods Supply Chain Cost Optimization Model Considering Rework of Defectives. M.Sc. Thesis, Bangladesh University of Engineering and Technology.

[18] Sankhye, S. Hu, G. (2020). Machine Learning Methods for Quality Prediction in Production. Logistics, 4(35).

[19] Quadir, M.A., Jha, K., Haneef, S., Sivaraman, A.K. Tee, K.F. (2022). A Review on Data-Driven Quality Prediction in the Production Process with Machine Learning for Industry 4.0. Processes, 10(1966).

[20] Fu, Y., Downey, A.R.J., Yuan, L., Zhang, T., Pratt, A. Balogun, Y. (2022). Machine learning algorithms for defect detection in metal laser-based additive manufacturing: A review. Journal of Manufacturing Processes, 75, pp. 693–710.

[21] Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R.P., Basto, J.P. Alcalá, S.G.S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers Industrial Engineering, 137.

[22] Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N. Nandi, A.K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138.

[23] Kuo, C.J., Ting, K.C., Chen, Y.C., Yang, D.L. Chen, H.M. (2017). Automatic machine status prediction in the era of Industry 4.0: Case study of machines in a spring factory. Journal of Systems Architecture, 81.

[24] Ma, Q. (2023). Product Quality Management in Supply Chains: Applications of Data-Driven Approaches and Incentive Contracts. Ph.D. Dissertation, Aarhus University.

[25] Joe F. Ray, William W. Edge, Gerald P. Raia, Kenneth J. Brown, Jr., "Logistics and Maintenance Concepts for a Future Naval Force," *Master's Thesis*, Naval Postgraduate School, December 2001.

[26] Elreedy, D., Atiya, A.F. Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113, pp. 4903–4923.