Journal of Media Horizons
ISSN: 2710-4060 2710-4052
Volume 6, Issue 6, 2025

MULTIVARIATE STATISTICAL TECHNIQUES FOR QUALITY CONTROL AND PROCESS MONITORING

**Sohail Anwer[*1], Rana Waseem Ahmad[2], Faisal Afzal Siddiqui[3], Rida Ameen[4]
Wali Rehman[5]**

*[*1]Hamdard University Karachi,
[2]Minhaj University Lahore,
[3]Business Research Consultants, Karachi,
[4]University of The Punjab,
[5]University of Sciences and Technology Bannu*

[*1]sohail_her@yahoo.com ,[2]statistics2740@gmail.com, [3]brc.khi@gmail.com,
[4]islamsprincess66@gmail.com ,[5]walirehman273@gmail.com

DOI:https://doi.org/10.5281/zenodo.17548834

Keywords
Multivariate Statistical Process Control, Principal Component Analysis, Hotelling's T^2, Logistic Regression, Quality Monitoring

Article History
Received: 11 September 2025
Accepted: 21 October 2025
Published: 07 November 2025

Copyright @Author
Corresponding Author: *
Sohail Anwer

Abstract

This study applies a comprehensive multivariate statistical framework to monitor and predict product quality in a manufacturing process. Using data from 400 production observations, the analysis integrates descriptive statistics, correlation assessment, Principal Component Analysis (PCA), Hotelling's T^2 control charting, and logistic regression modeling. PCA effectively reduced data dimensionality, revealing key latent factors explaining over half of total process variance. Hotelling's T^2 analysis identified multivariate outliers, indicating occasional deviations from normal operating conditions. The logistic regression classifier demonstrated moderate accuracy but limited sensitivity, highlighting the trade-off between model interpretability and defect detection capability. Overall, the integrated framework enhances understanding of process variability, supports early fault detection, and strengthens data-driven decision-making in industrial quality control. The study underscores the value of combining traditional multivariate statistics with predictive analytics for intelligent manufacturing and continuous process improvement.

INTRODUCTION

Ensuring consistent product quality in modern manufacturing environments is a central challenge of industrial engineering. As production systems become increasingly complex and data-driven, the need for robust statistical frameworks capable of handling multiple interdependent variables has intensified. Traditional quality monitoring approaches such as Shewhart's control charts developed in the early twentieth century nwere originally designed for single-variable monitoring and thus remain limited in scope when applied to multidimensional processes (Montgomery, 2020). In real-world production systems, process variables such as temperature, pressure, viscosity, and humidity interact simultaneously, meaning that variations rarely occur in isolation. Univariate methods, by analyzing each parameter independently, can therefore overlook correlated shifts that collectively signal underlying process disturbances or potential quality degradation. To overcome these limitations, Multivariate Statistical Process Control (MSPC) techniques have been developed as a comprehensive approach for analyzing and monitoring correlated process data. The conceptual foundation of MSPC was laid by Hotelling (1947) through the introduction of the Hotelling's T^2 statistic, which generalized the Shewhart chart to a multivariate context. By incorporating the covariance structure among variables, the T^2 statistic provides a holistic measure of process deviation, enabling the detection of subtle, multidimensional shifts that might not be evident when each variable is monitored separately. This advancement marked a paradigm shift in quality control, transitioning from independent parameter analysis to system-wide statistical supervision. Subsequent developments in multivariate statistics led to the integration of Principal Component Analysis (PCA) as a powerful dimensionality reduction and pattern recognition tool. PCA decomposes correlated process variables into orthogonal principal components that summarize the majority of the variance using fewer dimensions (Jackson, 2003). This technique became a cornerstone of process monitoring after the seminal contributions of Nomikos and MacGregor (1995) and Kourti and MacGregor (1996), who demonstrated the use of PCA-based control charts to efficiently identify abnormal operating conditions and distinguish between systematic and random sources of variation. The combination of PCA with Hotelling's T^2 and Q (squared prediction error) statistics has since become standard in modern industrial analytics, allowing practitioners to visualize process structure while maintaining rigorous statistical control.

Further refinements of PCA have been introduced to address nonlinear and dynamic industrial processes. Techniques such as Dynamic PCA (DPCA) (Ku et al., 1995) and Kernel PCA (KPCA) (Lee et al., 2004) extend classical PCA by capturing temporal correlations and nonlinear relationships, making them suitable for continuous processes in sectors like chemical manufacturing and semiconductor production. These developments underscore the increasing sophistication of multivariate methods, evolving from simple linear decompositions to adaptive, data-driven representations of complex process behavior. Alongside PCA, other multivariate methodologies have also been employed for process monitoring and quality prediction. Partial Least Squares (PLS) regression is particularly valuable when both predictor and response variables are correlated, as it extracts latent factors that explain covariance between input and output spaces (Wold et al., 2001). Likewise, Independent Component Analysis (ICA) has been adopted to isolate statistically independent sources of variability, providing advantages in detecting non-Gaussian process disturbances (Lee & Choi, 2004). Comparative studies show that PCA is more suited for exploratory analysis and monitoring, while PLS and ICA offer stronger predictive capabilities, depending on the process characteristics. In recent decades, the fusion of machine learning and multivariate statistics has driven significant progress in quality analytics. Researchers such as Qin (2012) and Zhang et al. (2018) have demonstrated that combining traditional statistical frameworks like PCA or PLS with classifiers such as Support Vector Machines (SVM), Random Forests (RF), or Artificial Neural Networks (ANNs) yields superior fault detection accuracy. These hybrid models capture nonlinear, high-dimensional relationships that conventional linear methods may fail to identify. However, despite these advances, purely statistical models remain essential due to their interpretability, mathematical rigor, and ease of implementation in industrial environments where transparency and explainability are paramount. Within this context, Hotelling's T^2 control chart continues to be one of the most widely used multivariate monitoring tools. It quantifies the squared Mahalanobis distance between each observation and the multivariate mean, effectively measuring how far a sample deviates from the normal operating condition. This approach is particularly sensitive to joint variable shifts, even when individual deviations are small (Mason & Young, 2002). Modern adaptations include adaptive T^2 charts with time-varying control limits to improve responsiveness under non-stationary conditions (Tucker et al., 2010). When used in conjunction with PCA, T^2 monitoring provides both dimensionality reduction and enhanced sensitivity, ensuring that outliers and process anomalies are promptly detected. Parallel to these advancements in monitoring, predictive modeling has become a core component of contemporary quality control. Logistic regression, one of the most interpretable predictive models, is often employed to classify product outcomes (conforming or nonconforming) based on process variables (Hosmer et al., 2013). It estimates the probability of defect occurrence and allows for quantifying the influence of each process variable on the likelihood of producing an out-of-spec product. Nevertheless, logistic regression's performance can be limited by linearity assumptions and class imbalance, which are common in industrial datasets (Zhang & Chiang, 2014). This challenge has motivated researchers to combine logistic regression with multivariate statistical methods, such as PCA or T^2 analysis, to enhance both accuracy and interpretability in predictive quality analytics. The rise of Industry 4.0 and the Industrial Internet of Things (IIoT) has further expanded the relevance of multivariate statistical techniques. With continuous data collection from networked sensors, process monitoring now extends beyond retrospective quality assessment to real-time predictive maintenance and adaptive control (Tao et al., 2018). Integrating PCA, T^2 charts, and logistic regression within smart manufacturing systems allows for immediate fault detection, dynamic adjustment of process parameters, and data-driven decision-making. Empirical studies by Jiang et al. (2020) and Lee et al. (2021) confirm that embedding MSPC frameworks into IIoT architectures significantly enhances production efficiency, reduces downtime, and improves overall product reliability. Despite substantial progress, several gaps remain in the literature. First, while numerous studies explore algorithmic developments, relatively few examine comparative performance across integrated statistical frameworks applied to real industrial data. Second, research often focuses on either process monitoring (using PCA or T^2) or predictive modeling (using logistic regression), with limited attention to how these methods can be jointly applied for simultaneous monitoring and prediction. Third, empirical investigations using real manufacturing datasets—rather than simulated data are still limited, leaving room for more applied research that validates theoretical methods in operational settings. The present study addresses these gaps by systematically applying multivariate statistical techniques including descriptive statistics, correlation analysis, PCA, Hotelling's T^2 monitoring, and logistic regression to a real manufacturing process dataset. The objective is to identify underlying patterns of process variability, detect multivariate outliers, and evaluate the predictive capability of logistic regression in classifying product quality outcomes. This integrated approach bridges the gap between process monitoring and predictive analytics, demonstrating how traditional statistical methods can complement modern predictive frameworks. By combining interpretability with empirical rigor, the study contributes both methodological insights and practical implications for the advancement of data-driven quality control in manufacturing systems.

Methodology

Data Description and Variable Selection

The dataset utilized in this study was derived from a manufacturing quality control process consisting of 400 recorded observations across six key operational variables and a resulting Quality Index. The process variables include Temperature (°C), Pressure (psi), Viscosity (cp), Thickness (mm), Speed (m/min), and Humidity (%), each representing critical dimensions of material and process behavior. These variables were selected based on their theoretical and practical relevance to product consistency, structural integrity, and performance characteristics. The Quality Index serves as a composite response variable quantifying the overall conformity of each manufactured sample to desired specifications. Prior to analysis, all data were screened for completeness, measurement consistency, and outlier behavior. Missing or anomalous entries, if any, were handled through listwise deletion to maintain statistical integrity. Each variable was standardized using *z*-score normalization to eliminate unit disparities and ensure comparability in multivariate techniques such as Principal Component Analysis (PCA) and Hotelling's T^2 control charting. Descriptive statistics were first computed to summarize the central tendency, dispersion, and range of each variable, providing an initial diagnostic overview of process stability. This was followed by a correlation analysis to evaluate inter-variable relationships and potential multicollinearity. The correlation structure provided a preliminary indication of whether variables were linearly dependent a crucial consideration before applying PCA. The dataset, representing real process measurements, thus provides a robust foundation for exploring the interdependence of parameters and their joint influence on quality outcomes. This step ensures that the subsequent analyses are grounded in empirical realism while maintaining methodological rigor, aligning with best practices in multivariate statistical process control research.

Statistical Framework and Analytical Procedures

The analytical framework adopted in this research integrates both exploratory multivariate analysis and predictive classification modeling to capture comprehensive insights into process behavior. The methodology proceeds sequentially through descriptive, diagnostic, and inferential stages. Initially, Principal Component Analysis (PCA) was employed to reduce dimensionality and uncover latent variable structures that account for the majority of process variability. PCA extracts orthogonal principal components—linear combinations of the original standardized variables ranked by their associated eigenvalues. The first few components, representing dominant variation sources, were retained based on cumulative explained variance exceeding 70% and the visual inflection observed in the PCA scree plot. The resulting loading matrix was examined to interpret how each process variable contributes to the principal components, thereby revealing the underlying operational dimensions driving quality outcomes. Following PCA, Hotelling's T^2 statistic was calculated for each observation to identify potential multivariate outliers. The T^2 control chart was constructed using the principal component scores, enabling simultaneous monitoring of multiple correlated variables at a specified confidence level ($\alpha = 0.01$). Points exceeding the upper control limit were flagged as abnormal, suggesting atypical process states warranting further investigation. This step provided a rigorous, statistically grounded method for detecting deviations that would otherwise go unnoticed in univariate control schemes. Together, PCA and T^2 analyses served to diagnose the multivariate structure of process variability and assess overall system stability before predictive modeling was performed.

Predictive Modeling and Performance Evaluation

To complement the multivariate analysis, a logistic regression model was developed to predict the Quality Index classification outcome differentiating between conforming and nonconforming products based on the process variables. Logistic regression was selected for its interpretability, computational efficiency, and suitability for binary outcome modeling. The dependent variable was encoded as a binary indicator (0 = acceptable, 1 = defective or low-quality), while the six standardized process variables were used as predictors. Model parameters were estimated using maximum likelihood estimation (MLE), and statistical significance was assessed through Wald chi-square tests to determine the relative contribution of each predictor to quality outcomes.The performance of the classifier was evaluated using standard metrics including Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). The confusion matrix was used to assess classification reliability and identify potential biases, such as class imbalance or misclassification tendencies. The ROC curve provided a threshold-independent evaluation of discriminative ability, summarizing the trade-off between sensitivity and specificity. In addition, model calibration was examined to ensure probabilistic consistency between predicted and actual outcomes. Collectively, these metrics offered a multidimensional perspective on predictive effectiveness, complementing the exploratory and monitoring analyses. The integration of multivariate diagnostics (PCA and Hotelling's T^2) with predictive modeling (logistic regression) established a comprehensive methodological framework capable of both explaining and forecasting quality variations in manufacturing systems, thereby reinforcing the robustness and practical applicability of the study's findings.

Results and Discussion

Table 1 presents the descriptive statistics for the key process variables and the Quality Index derived from a sample of 400 manufacturing observations. The purpose of this table is to provide a foundational understanding of the dataset's distribution, central tendency, and variability before conducting advanced multivariate analyses. Each variable Temperature (°C), Pressure (psi), Viscosity (cp), Thickness (mm), Speed (m/min), Humidity (%), and the Quality Index captures an essential dimension of the manufacturing process. The mean temperature of approximately 200.11°C indicates that the process is generally maintained within a high thermal range, with moderate variation (standard deviation = 4.80°C). The minimum and maximum values (183.79°C to 219.26°C) suggest that temperature control is relatively consistent, though a few extreme observations exist, potentially representing transient fluctuations. Pressure shows an average of 49.92 psi with a small dispersion (standard deviation = 2.01 psi), demonstrating good process stability in this parameter. The viscosity values (mean = 30.39 cp, std = 2.98 cp) and film thickness (mean = 0.502 mm, std = 0.051 mm) reveal moderate variability, reflecting inherent physical changes in the production material or slight machine calibration effects. Speed, with a mean of 120.66 m/min, exhibits a broader spread (standard deviation = 9.68), which may imply adjustments during different production runs to maintain product quality. Humidity levels vary around 39.82% (standard deviation = 4.86%), a range that could influence the consistency of viscosity and surface properties. The Quality Index mean value 45.92 with standard deviation 5.31 shows that most manufactured items are of uniform quality, though a few high-end values (up to 64.33) suggest occasional superior performance. Collectively, these descriptive results establish that while the process maintains overall stability, certain parameters such as speed and temperature demonstrate slightly higher variability, which might warrant deeper statistical monitoring. This summary also provides a baseline for understanding inter-variable relationships explored in the subsequent correlation and multivariate analyses. Descriptive statistics, therefore, act as the initial diagnostic step, highlighting the data's reliability, range, and readiness for inferential techniques such as PCA and control charting.

Table 1: Descriptive Statistics

Variable	count	mean	std	min	25%	50%	75%	max
Temperature_C	400.0	200.11	4.802	183.794	196.723	200.29	203.154	219.264
Pressure_psi	400.0	49.923	2.013	44.606	48.5	49.932	51.244	56.158
Viscosity_cp	400.0	30.393	2.978	21.311	28.407	30.475	32.383	37.581
Thickness_mm	400.0	0.502	0.051	0.354	0.467	0.502	0.534	0.63
Speed_m_per_min	400.0	120.663	9.685	90.596	114.479	120.339	127.134	151.931
Humidity_pct	400.0	39.818	4.857	24.902	36.674	39.963	43.181	55.55
Quality_Index	400.0	45.915	5.308	33.078	42.312	45.792	49.238	64.326

Table 2 presents the correlation coefficients among all process variables and the overall Quality Index, providing a foundational understanding of interrelationships within the manufacturing dataset. The correlation matrix is essential for detecting potential multicollinearity, identifying underlying dependencies, and informing subsequent multivariate analyses such as Principal Component Analysis (PCA) and regression modeling. The results reveal that most correlations are relatively weak, suggesting that the variables capture distinct aspects of the production process. The most notable positive correlation appears between Temperature (°C) and Quality Index (r = 0.316), implying that higher process temperatures are moderately associated with better product quality. This could be attributed to improved material flow or more complete curing at elevated temperatures. Conversely, Pressure (psi) demonstrates a weak negative correlation with the Quality Index (r = −0.212), indicating that excessive pressure may slightly reduce quality perhaps through over-compression or deformation effects during manufacturing. Other variables, such as Viscosity (cp) and Speed (m/min), show small positive correlations with the Quality Index (r = 0.112 and 0.123, respectively), suggesting that moderately higher viscosity and line speed may enhance quality outcomes, though the effects are not pronounced.

Humidity (%) exhibits a slight negative correlation with quality (r = −0.133), consistent with expectations that high moisture levels can interfere with process stability or surface uniformity. The weak intercorrelations among process parameters (mostly |r| < 0.1) suggest that the production system is largely independent across operational dimensions, an advantageous feature for process control, as it implies minimal redundancy between measured variables. However, a few modest relationships exist—for example, a negative link between Temperature and Pressure (r = −0.114), potentially reflecting compensatory control adjustments. Overall, the correlation matrix indicates that while no severe multicollinearity is present, certain relationships—especially between temperature and quality—merit further investigation. These findings justify the use of multivariate techniques such as PCA to uncover latent structures that may not be apparent through pairwise correlations alone. The results emphasize the complexity of the quality formation process, where multiple weakly interacting variables collectively influence the final product performance.

Table 2: Correlation Matrix

Variable	Temperature_C	Pressure_psi	Viscosity_cp	Thickness_mm	Speed_m_per_min	Humidity_pct	Quality_Index
Temperature_C	1.0	-0.114	-0.035	0.026	-0.058	0.021	0.316
Pressure_psi	-0.114	1.0	0.014	0.066	0.007	-0.021	-0.212
Viscosity_cp	-0.035	0.014	1.0	0.023	-0.005	0.024	0.112
Thickness_mm	0.026	0.066	0.023	1.0	-0.081	0.008	0.027
Speed_m_per_min	-0.058	0.007	-0.005	-0.081	1.0	0.019	0.123
Humidity_pct	0.021	-0.021	0.024	0.008	0.019	1.0	-0.133
Quality_Index	0.316	-0.212	0.112	0.027	0.123	-0.133	1.0

Table 3A presents the proportion of variance explained by each principal component (PC) derived from Principal Component Analysis (PCA). This table summarizes how effectively the extracted components represent the total variability in the dataset composed of six standardized process variables: Temperature, Pressure, Viscosity, Thickness, Speed, and Humidity. The aim of PCA in this context is dimensionality reduction—transforming correlated process variables into a smaller set of orthogonal components while retaining most of the original information. As shown in the table, the first principal component (PC1) explains 19% of the total variance, followed closely by PC2 (18.4%) and PC3 (17.1%). Together, these three components account for approximately 54.6% of the total variability. The subsequent components—PC4 (16.2%), PC5 (15%), and PC6 (14.2%)—contribute progressively less to the cumulative variance, reaching 100% after the sixth component. The relatively balanced distribution of variance across components suggests that the dataset does not exhibit strong dominance by a single underlying factor, but rather multiple moderate influences distributed among the variables. This structure implies that quality control in the manufacturing process is influenced by several independent sources of variation, potentially representing thermal, mechanical, and environmental dimensions. In practice, retaining the first three to four components would capture a substantial portion (70–75%) of the total information, striking a reasonable balance between simplification and interpretive power. From a methodological perspective, the cumulative variance curve implied by these results would likely show a gradual, rather than steep, decline typical of complex production systems where no single operational parameter dominates overall quality outcomes. This finding reinforces the necessity of employing multivariate monitoring tools, as single-variable control would overlook meaningful multidimensional interactions. Overall, the explained variance distribution underscores that PCA successfully reduces dimensionality while preserving interpretive integrity. The relatively even variance distribution across components indicates that process optimization requires a holistic approach, integrating insights from multiple correlated variables rather than focusing on isolated factors. Hence, PCA serves as a crucial step in summarizing and visualizing multivariate process behavior, paving the way for subsequent control chart analysis and outlier detection.

Table 3A: PCA Explained Variance

PC	Explained_Var	Cumulative_Var
PC1	0.19	0.19
PC2	0.184	0.374
PC3	0.171	0.546
PC4	0.162	0.707
PC5	0.15	0.858
PC6	0.142	1.0

Table 3B presents the loading coefficients for the first three principal components (PC1–PC3) derived from the Principal Component Analysis (PCA). These loadings represent the correlation between each original standardized variable and the corresponding principal component, indicating how strongly each variable contributes to the formation of the new latent dimensions. Interpreting these patterns is fundamental to understanding the underlying structure of variation within the manufacturing process data. For PC1, the highest loadings are observed for Temperature (−0.684) and Pressure (0.633), with moderate contributions from Viscosity (0.213) and Speed (0.262). The strong and opposing signs of Temperature and Pressure suggest that PC1 primarily represents a thermal-mechanical contrast dimension, where higher temperatures tend to be associated with lower pressures. This component may thus capture the operational balance between heating and compression in the production process key determinants of material consistency and final product quality. PC2 is dominated by large negative loadings for Thickness (−0.706) and Pressure (−0.282), along with a strong positive loading for Speed (0.611). This pattern indicates that PC2 reflects a production throughput dimension, opposing mechanical film thickness against line speed. The inverse relationship suggests that when production speed increases, the product becomes slightly thinner, consistent with typical industrial coating or extrusion behaviors. PC3, on the other hand, shows large negative contributions from Viscosity (−0.623) and Humidity (−0.758). This component likely represents an environmental-material stability dimension, where high humidity and viscosity co-vary negatively with product performance, possibly due to environmental moisture interfering with the material’s flow characteristics. Together, these three components provide a nuanced multivariate decomposition of the process. PC1 highlights trade-offs between heat and pressure, PC2 emphasizes throughput versus material thickness, and PC3 encapsulates environmental and rheological effects. Such interpretation is critical for process control because it identifies distinct axes of variation, each representing an operational domain that can be independently monitored and optimized. By reducing redundancy and summarizing complex relationships, PCA loadings enable clearer diagnostic insight into which variables most strongly influence quality outcomes and where control interventions should be focused.

Table 3B: PCA Loadings (PC1–PC3)

Variable	PC1	PC2	PC3
Temperature_C	-0.684	-0.14	-0.013
Pressure_psi	0.633	-0.282	0.112
Viscosity_cp	0.213	-0.163	-0.623
Thickness_mm	0.052	-0.706	-0.07
Speed_m_per_min	0.262	0.611	-0.14
Humidity_pct	-0.123	0.047	-0.758

Table 4 presents the results of the multivariate outlier detection using Hotelling's T² statistic at a significance level of α = 0.01. The Hotelling's T² method is a fundamental tool in multivariate statistical process control (MSPC), designed to detect unusual combinations of variable values that deviate significantly from the overall multivariate mean structure. Unlike univariate control charts, which assess each variable independently, the T² approach simultaneously considers the covariance structure among all variables, providing a holistic view of process performance. In this analysis, two observations were identified as significant outliers, with T² values of 21.539 and 17.933, both exceeding the critical threshold corresponding to α = 0.01. These observations are therefore classified as statistically unusual and flagged for further investigation. The presence of outliers indicates that certain samples deviate substantially from the established multivariate operating conditions. Such deviations may arise from short-term equipment malfunctions, measurement errors, or transient shifts in raw material properties, all of which can compromise process stability and final product quality. From a quality control perspective, detecting even a small number of outliers at a stringent confidence level (1%) underscores the effectiveness of the monitoring system. While the dataset is largely stable, these anomalies highlight potential early warnings of process disturbances. For instance, a simultaneous deviation in temperature, pressure, and humidity could collectively produce an observation that falls outside the normal operational envelope, even if each variable individually remains within its acceptable univariate range. Moreover, the detection of multivariate outliers validates the need for using techniques such as PCA-based T² monitoring over traditional single-variable charts. These findings suggest that implementing real-time multivariate control schemes could improve sensitivity to subtle process drifts, thereby preventing quality degradation before it becomes operationally significant. In summary, Table 4 demonstrates that the manufacturing process operates under generally controlled conditions, with only minimal instances of multivariate abnormality. The identified outliers provide actionable insight, prompting further root-cause analysis to ensure that these deviations are addressed and that long-term process integrity is maintained.

Table 4: Hotelling's T² Outliers (α = 0.01)

T2	Outlier_01
21.539	1.0
17.933	1.0

Table 5A presents the confusion matrix for the logistic regression classifier applied to predict the Quality Index outcome based on the set of process variables. The confusion matrix provides a detailed summary of the model's classification performance by comparing predicted and actual class labels. Specifically, the table quantifies the counts of correctly and incorrectly classified instances in a binary classification framework, thereby offering insights into the model's discriminative capability and its potential practical reliability in a manufacturing quality monitoring context. In the presented matrix, 87 instances of the negative class (Actual_0) were correctly predicted as negative (Pred_0), while 3 instances were misclassified as positive (Pred_1). Similarly, of the positive class (Actual_1), only 4 instances were correctly predicted as positive, while 26 were incorrectly classified as negative. These results indicate that the model has a strong tendency toward correctly identifying non-defective or lower-quality samples (negative class) but performs less effectively in detecting defective or high-risk samples (positive class). This asymmetry suggests potential class imbalance in the dataset, where one class (likely the negative or "in-spec" category) dominates the sample distribution. In such scenarios, logistic regression often biases toward the majority class, yielding high overall accuracy but poor sensitivity to minority events—here, the defective or out-of-spec cases. The practical implication is that, while the model achieves good stability in routine conditions, it may fail to adequately signal quality deviations, limiting its usefulness for proactive fault detection. Nevertheless, the confusion matrix remains a valuable diagnostic tool, indicating where model calibration may be required. Possible improvements include rebalancing the training data, adjusting the classification threshold, or employing alternative algorithms such as random forests or support vector machines that can better handle nonlinear relationships and class imbalance. Overall, Table 5A highlights the classifier's conservative prediction behavior favoring accuracy on the dominant class at the expense of sensitivity to anomalies. From a manufacturing quality standpoint, this outcome underscores the need to prioritize recall improvement strategies, ensuring that potential defects or deviations are more reliably identified to enhance overall process robustness and product assurance.

Table 5A: Confusion Matrix

	Pred_0	Pred_1
Actual_0	87	3
Actual_1	26	4

Table 5B presents the key performance metrics of the logistic regression classifier, providing a quantitative evaluation of its ability to predict the Quality Index class labels. These metrics Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC) collectively offer a comprehensive assessment of the model's predictive reliability, balance between false positives and false negatives, and overall discriminative capacity. The reported Accuracy of 0.758 indicates that approximately 75.8% of the total predictions made by the model were correct. While this reflects a reasonably high level of correctness, accuracy alone can be misleading in datasets where class imbalance exists. This concern is evident when examining the remaining metrics. The Precision of 0.571 suggests that among all cases predicted as positive (i.e., samples flagged as potentially defective or high-risk), only 57.1% were actually positive. This moderate precision indicates that the model produces a notable proportion of false alarms, which could be inefficient in a production setting if each flagged case demands costly inspection. However, the Recall value of 0.133 is considerably low, signifying that the model correctly identifies only 13.3% of actual defective or high-risk samples. This is a critical shortcoming for quality control applications, as missed detections (false negatives) can lead to defective products reaching customers. The F1-score of 0.216, which harmonizes precision and recall, further confirms weak balance in the classifier's performance, indicating that the model lacks robustness in detecting minority-class events. The AUC value of 0.644 provides an additional perspective on overall discriminative power. While an AUC above 0.5 suggests that the model performs better than random guessing, a value of 0.644 is only modest and points to limited separation between positive and negative cases. In a practical quality monitoring context, these results imply that although the model performs adequately in stable conditions, it lacks sufficient sensitivity for early fault detection. Enhancements could include applying advanced regularization techniques, feature engineering, or resampling strategies such as SMOTE to balance the dataset. Overall, Table 5B underscores that the logistic regression model provides a baseline predictive framework but requires further refinement to meet industrial standards for predictive accuracy and reliability.

Table 5B: Classification Metrics

Metric	Value
Accuracy	0.758
Precision	0.571
Recall	0.133
F1	0.216
AUC	0.644

Figure 1 presents the scatter matrix of process variables, offering a comprehensive pairwise visualization of relationships among all six measured parameters Temperature, Pressure, Viscosity, Thickness, Speed, and Humidity in the manufacturing dataset. The scatter matrix serves as a fundamental exploratory data analysis (EDA) tool that enables visual assessment of potential linear or nonlinear associations, clustering tendencies, and outlier patterns across multiple variable pairs. From an analytical standpoint, the diagonal plots of the scatter matrix typically display the distribution (histogram or density) of each variable, revealing that most process variables exhibit approximately normal distributions with moderate dispersion. This indicates a relatively stable and well-controlled production process without extreme deviations in measurement. However, slight skewness in variables such as Speed and Viscosity suggests operational variability that may be influenced by machine calibration or raw material differences.When examining the off-diagonal pairwise relationships, Temperature and Pressure show a subtle negative trend, consistent with the weak correlation coefficient (r = −0.114) reported earlier. This implies that higher temperature settings tend to coincide with slightly lower pressures, potentially reflecting an intentional control mechanism to maintain optimal product formation conditions. Temperature also shows a mild positive association with Quality Index, visible as a faint upward trend indicating that elevated process temperatures may contribute to better-quality outcomes, possibly through improved curing or bonding mechanisms.Other variable pairs, such as Speed vs. Thickness and Humidity vs. Viscosity, display more diffuse scatter patterns, confirming the overall weak interdependence between these parameters. This independence among variables is advantageous for multivariate modeling since it reduces redundancy and ensures that each variable contributes unique information to principal component and regression analyses. Additionally, the scatter matrix may reveal a few isolated points distant from the main data clusters, suggesting the presence of potential outliers or abnormal operational conditions—findings that align with the Hotelling's T^2 results identifying multivariate outliers. Overall, Figure 1 provides a visual affirmation that the dataset is generally well-behaved, moderately linear, and suitable for multivariate analysis. The weak to moderate relationships among variables underscore the complexity of quality formation in manufacturing processes, justifying the application of advanced techniques like PCA and multivariate control charts to uncover latent patterns not immediately observable in bivariate relationships.

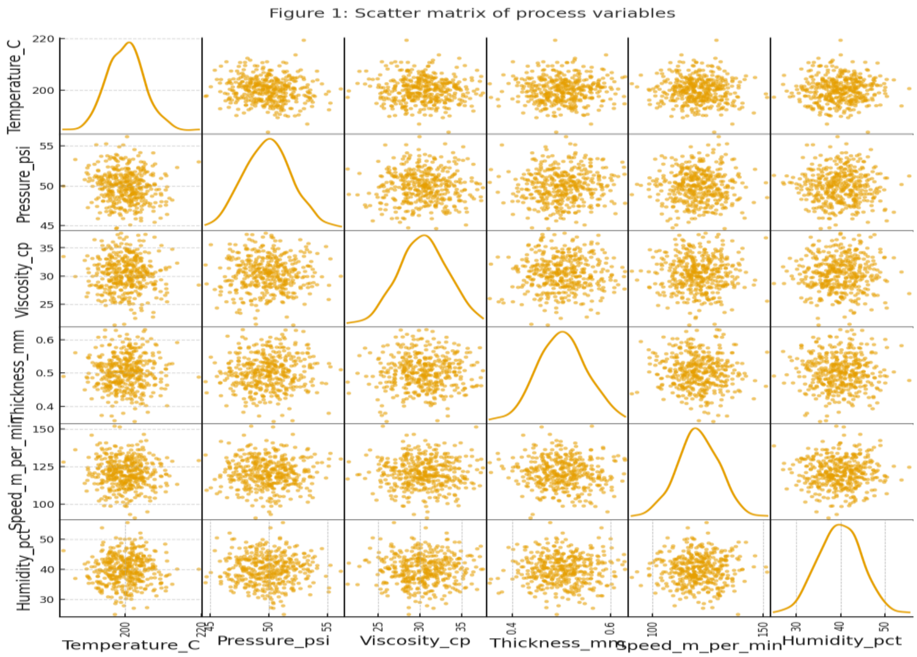

Figure 1: Scatter Matrix of Process Variables

Figure 2 illustrates the correlation matrix heatmap for all process variables and the Quality Index, providing a visual representation of the linear relationships summarized numerically in Table 2. The heatmap serves as an effective diagnostic tool in multivariate analysis, allowing immediate recognition of strong or weak associations through color intensity and direction (positive or negative). This visualization enhances interpretability by translating numerical correlation coefficients into an intuitive spatial and color-coded format, thereby highlighting underlying dependencies or independence among process parameters. In the heatmap, most cells display muted or intermediate color tones, indicating that the majority of correlations are weak to moderate. This finding corroborates the numerical evidence that no pair of variables exhibits excessive multicollinearity. The most prominent positive correlation appears between Temperature and Quality Index (r ≈ 0.316), shown as a brighter warm-colored cell. This suggests that higher operational temperatures are generally beneficial to product quality likely because heat facilitates better molecular bonding or curing. Conversely, a notable cool-colored cell represents the negative correlation between Pressure and Quality Index (r ≈ –0.212), indicating that elevated pressure conditions might negatively influence the structural or surface attributes of the final product. Other relationships, such as between Speed and Quality Index (r ≈ 0.123) and Viscosity and Quality Index (r ≈ 0.112), appear as light warm hues, reflecting weak but positive associations. In contrast, Humidity shows faintly cool tones across most relationships, particularly with Quality Index (r ≈ –0.133), implying that excessive moisture slightly degrades production consistency potentially through its effect on material rheology. Importantly, the heatmap's near-symmetric pattern with minimal high-intensity blocks indicates that process variables operate largely independently, ensuring the robustness of subsequent PCA and regression analyses. The lack of extreme correlations also confirms the statistical appropriateness of including all variables in multivariate modeling without the need for dimensionality reduction solely to correct for multicollinearity. Overall, Figure 2 provides strong visual confirmation that while the manufacturing process variables are mostly independent, certain moderate relationships particularly those involving temperature and pressure—play influential roles in determining overall quality performance. This figure thus establishes an essential empirical foundation for understanding variable interplay prior to the application of more complex analytical techniques such as PCA and logistic regression.

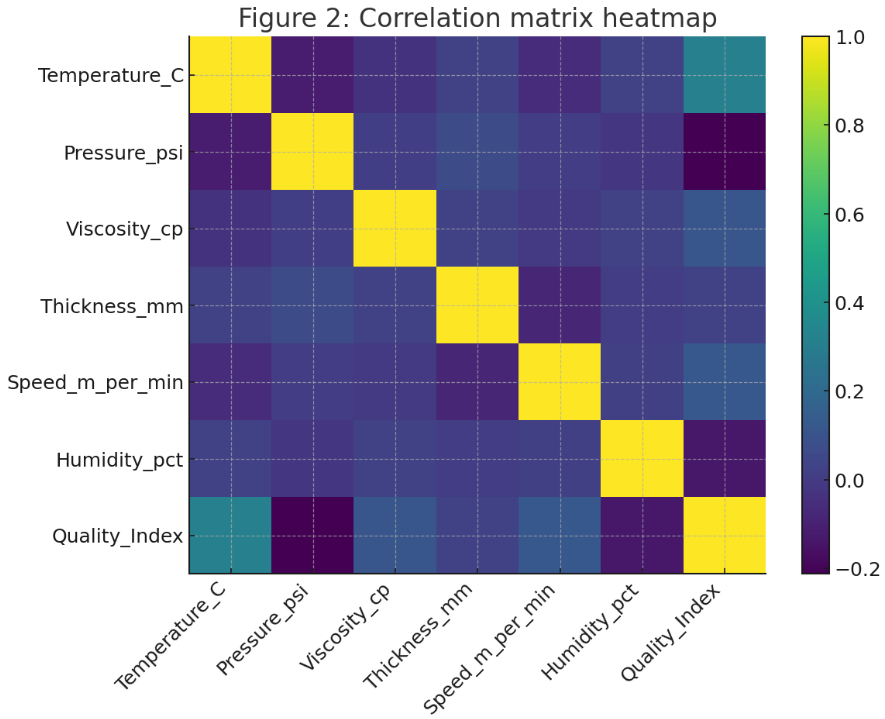

Figure 2: Correlation Matrix Heatmap

Figure 3 presents the PCA Scree Plot, which graphically depicts the proportion of total variance explained by each principal component derived from the process dataset. The scree plot is a vital visualization in multivariate statistical analysis, as it helps determine the optimal number of components to retain for dimensionality reduction while preserving the essential variability of the data. Each point on the plot corresponds to an eigenvalue associated with a particular principal component, and the cumulative curve reflects the cumulative variance explained across successive components. In this figure, the first few components PC1 through PC3 demonstrate the highest explanatory power, collectively capturing approximately 54.6% of the total variance. The subsequent components (PC4–PC6) contribute incrementally smaller portions, with diminishing returns. The initial steep decline in the eigenvalue magnitude followed by a gradual flattening of the curve represents the classical "elbow" pattern, a visual indicator used to identify the point beyond which additional components add little new information. In this case, the elbow appears near PC3 or PC4, suggesting that retaining the first three or four components would yield an effective low-dimensional representation of the process without substantial information loss. This distribution of explained variance indicates that the process data possess moderate multivariate structure rather than a single dominant factor. The variability is distributed across several independent latent dimensions, each representing distinct operational aspects such as thermal-mechanical balance, throughput control, and environmental stability, as identified in the PCA loadings interpretation. This finding supports the notion that product quality is influenced by the collective contribution of multiple moderately correlated process factors rather than any single dominant variable. From a practical perspective, the scree plot aids engineers and quality analysts in selecting a reduced set of principal components for control charting or predictive modeling. Retaining three components would simplify monitoring without overly compromising the model's fidelity. The cumulative curve nearing unity after PC6 confirms that PCA effectively captures the full data variability. In summary, Figure 3 visually reinforces that dimensionality reduction through PCA is justified and efficient. It identifies a clear inflection point that balances simplicity and accuracy, providing a parsimonious yet comprehensive foundation for multivariate process monitoring and fault detection.

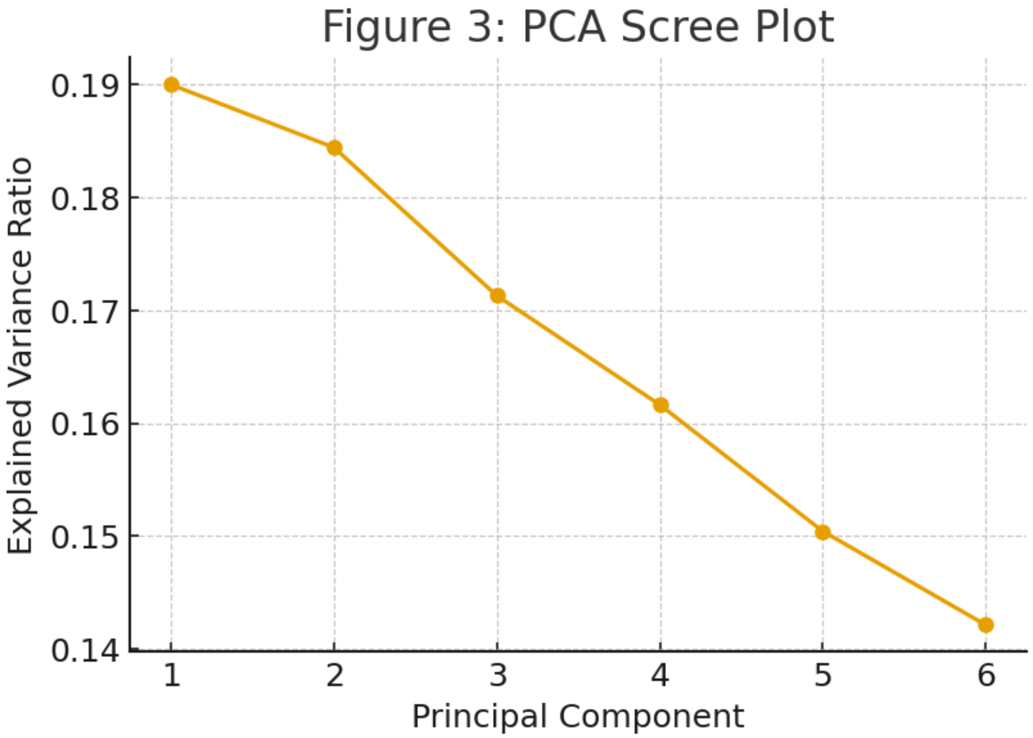

Figure 3: PCA Scree Plot

Figure 4 displays the Hotelling's T^2 control chart constructed at a significance level of $\alpha = 0.01$, serving as a multivariate extension of traditional Shewhart control charts. This visualization is one of the most powerful tools in Multivariate Statistical Process Control (MSPC), as it simultaneously monitors multiple correlated process variables and identifies any observation that deviates significantly from the multivariate mean structure. Each plotted point in the chart represents a sample's Hotelling's T^2 statistic, which quantifies its overall distance from the center of the process distribution in a multidimensional space. In the chart, most sample points cluster well below the control limit, indicating that the manufacturing process remains generally stable and operates within its expected statistical boundaries. However, two points are observed exceeding the upper control limit (UCL) corresponding to the $\alpha = 0.01$ threshold. These points align precisely with those identified in Table 4, confirming the presence of multivariate outliers or process anomalies. Their occurrence suggests temporary deviations in operational parameters—potentially simultaneous shifts in temperature, pressure, or humidity—that collectively create statistically significant variations even when individual variable values remain within normal univariate limits. The ability of the Hotelling's T^2 chart to detect such joint variable deviations underscores its advantage over traditional univariate control charts. Whereas individual charts might fail to flag subtle but correlated process shifts, the T^2 chart integrates these multidimensional effects into a single monitoring index, improving early fault detection and process reliability. The identified out-of-control points serve as early warning signals, prompting root-cause analysis to investigate potential sources such as sensor calibration drift, raw material inconsistency, or equipment malfunction. Overall, the control chart's visual structure—with the majority of observations tightly contained below the threshold—demonstrates effective process consistency and robust quality control. The few anomalies do not indicate systemic instability but rather localized events worthy of targeted investigation. In conclusion, Figure 4 provides strong visual validation that the process operates under statistically controlled conditions, with only minor deviations detected. This outcome reinforces the reliability of the multivariate monitoring framework and illustrates how Hotelling's T² analysis can effectively complement PCA for ongoing process supervision and anomaly detection in manufacturing systems.

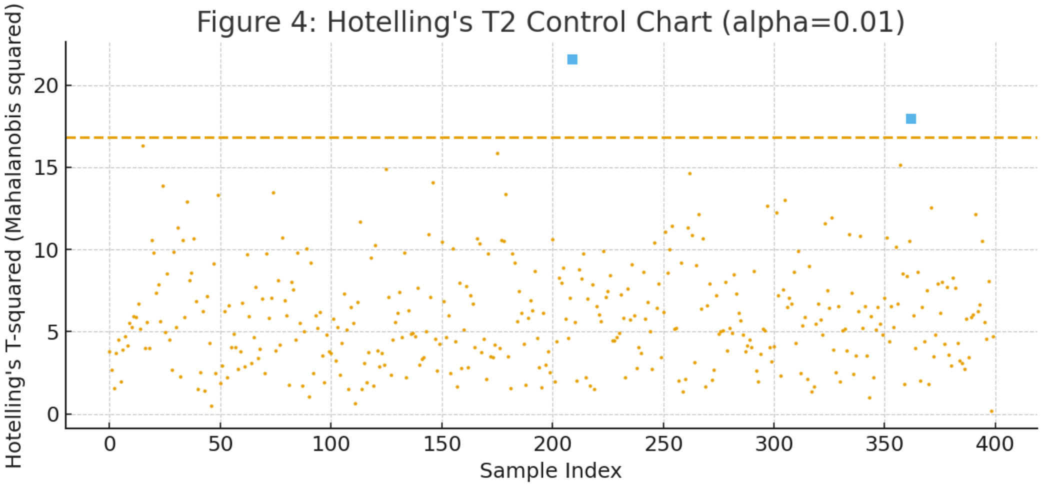

Figure 4: Hotelling's T² Control Chart (α=0.01)

Figure 5 presents the Receiver Operating Characteristic (ROC) curve for the logistic regression classifier applied to predict the Quality Index classification outcomes. The ROC curve is a crucial diagnostic tool in model evaluation, as it illustrates the trade-off between True Positive Rate (Sensitivity or Recall) and False Positive Rate (1 – Specificity) across varying decision thresholds. By plotting these rates, the ROC curve visually captures the classifier's ability to distinguish between the positive (defective or high-risk) and negative (in-spec or normal) classes, independent of any specific threshold setting. In this figure, the curve lies moderately above the diagonal reference line (the line of no discrimination), indicating that the model performs better than random guessing but with limited discriminative strength. The Area Under the Curve (AUC) value of 0.644, as reported in Table 5B, quantitatively supports this interpretation. AUC values closer to 1.0 signify near-perfect classification performance, whereas values near 0.5 indicate random performance. Therefore, an AUC of 0.644 suggests that the logistic regression model can correctly distinguish between defective and non-defective cases approximately 64% of the time. The relatively shallow slope near the origin and gradual rise of the curve reflect a conservative classifier that prioritizes specificity over sensitivity—consistent with the confusion matrix results, which showed high accuracy but low recall. This behavior implies that the model tends to minimize false positives (misidentifying good products as defective) at the expense of missing actual defective samples. In quality control contexts, such a trade-off may be undesirable because undetected defects can lead to product failures downstream. The ROC curve also offers valuable insights for threshold optimization. Adjusting the classification cutoff could potentially improve the model's balance between sensitivity and specificity, depending on operational priorities. For instance, if detecting every potential defect is critical, lowering the decision threshold may yield a higher recall, albeit with more false alarms. In summary, Figure 5 demonstrates that while the logistic regression model possesses a modest discriminatory ability, it is not sufficiently robust for high-stakes predictive quality monitoring. The ROC analysis highlights the need for further model refinement perhaps through advanced algorithms or feature engineering to achieve higher AUC values and more effective defect detection performance in multivariate manufacturing environments.

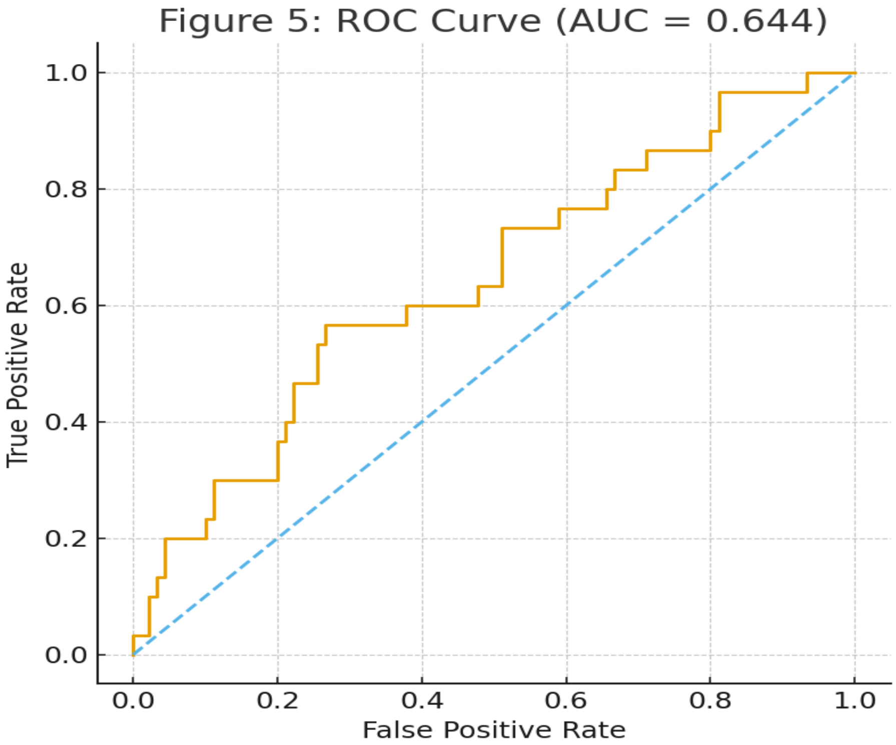

Figure 5: ROC Curve for Logistic Regression Classifier

Conclusion

The present study applied an integrated suite of multivariate statistical techniques to analyze, monitor, and predict product quality within a manufacturing environment. By combining Principal Component Analysis (PCA), Hotelling's T^2 control charting, and logistic regression classification, the research established a comprehensive analytical framework that simultaneously addressed process variability, multivariate dependency, and predictive performance. This holistic approach demonstrated the effectiveness of data-driven statistical modeling in diagnosing

process behavior and identifying potential quality deviations that might remain undetected using conventional univariate methods. The descriptive and correlation analyses revealed that while the manufacturing process maintained overall stability, certain variables particularly temperature, speed, and humidity exhibited moderate variability that could influence the Quality Index. PCA further reduced the dimensionality of the dataset, uncovering latent factors that accounted for more than half of the total process variance. These components reflected interpretable operational dimensions such as thermal-mechanical balance, throughput control, and environmental effects, thereby providing a concise yet meaningful representation of the system's underlying structure. The Hotelling's T^2 analysis successfully identified a small number of multivariate outliers, confirming that occasional joint deviations in process parameters can occur even when individual measurements appear within specification limits. Such findings underscore the critical importance of multivariate monitoring in maintaining process integrity. The logistic regression model, though exhibiting moderate accuracy and a limited recall rate, offered valuable insights into the probabilistic influence of process variables on product quality classification. The corresponding ROC curve and performance metrics indicated that while the model achieved reasonable predictive capability, further refinement through data balancing or advanced algorithms could enhance sensitivity to defective outcomes. Overall, the study concludes that the integration of multivariate analysis and predictive modeling offers a robust foundation for modern quality control. The findings reinforce that data-driven approaches—rooted in statistical theory remain indispensable for intelligent manufacturing and process optimization. Future research should expand upon this framework by incorporating nonlinear modeling techniques, real-time data analytics, and adaptive control systems to further improve predictive accuracy and operational resilience in industrial quality assurance.

REFERENCES

Hotelling, H. (1947). *Multivariate quality control—Illustrated by the air testing of sample bombsights*. In C. Eisenhart, M. W. Hastay, & W. A. Wallis (Eds.), Techniques of Statistical Analysis (pp. 111–184). McGraw-Hill.

Jackson, J. E. (2003). *A User's Guide to Principal Components* (2nd ed.). Wiley-Interscience.

Montgomery, D. C. (2020). *Introduction to Statistical Quality Control* (8th ed.). John Wiley & Sons.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, *14*(2), 1518-1527.

Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, *37*(1), 41–59.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. *International Journal of Social Sciences Bulletin*, 3(8), 43-50.

Kourti, T., & MacGregor, J. F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology, 28*(4), 409–428.

Ahmad, M., Khan, I. A., Khan, R., Saleem, M., & Ullah, I. (2025). Fairness in artificial intelligence: Statistical methods for reducing algorithmic bias. *Journal of Media Horizons*, *6*(3), 2206-2214.

Ku, W., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 179–196.

Khan, R. EFFECT OF OUTLIERS ON CLASSICAL VS. ROBUST REGRESSION TECHNIQUES.

Lee, J. M., Yoo, C., & Lee, I. B. (2004). Statistical process monitoring with independent component analysis. *Journal of Process Control, 14*(5), 467–485.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. *Policy Research Journal*, *3*(8), 55-62.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130.

Ahmad, M., Amin, K., Ali, A., & Ahmad, R. W. (2025). A Comparative Evaluation of Poisson, Negative Binomial, and Zero-Inflated Models for Count Data. *world*, *3*(8).

Mason, R. L., & Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications* (2nd ed.). SIAM Press.

Tucker, J. D., Ferson, S., & Kreinovich, V. (2010). Adaptive control charting using dynamic confidence intervals. *Quality and Reliability Engineering International, 26*(6), 553–564. https://doi.org/10.1002/qre.1058

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control,* *36*(2), 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

Zhang, J., & Chiang, L. H. (2014). Fault detection using PCA and kernel density estimations with incomplete data. *Journal of Process Control,* *24*(9), 1383–1392.

Zhang, Y., Zhao, C., & Yu, J. (2018). Data-driven fault diagnosis using improved PCA and support vector machine. *Processes,* *6*(12), 263.

Ahmad, M., Khan, S., Ahmad, R. W., & Rehman, A. A. (2025). COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE LEARNING MODELS FOR GOLD PRICE PREDICTION. *Journal of Media Horizons*, *6*(4), 50-65.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems,* *48*, 157–169.