Concrete Strength Monitoring Using Machine Learning and Statistical Control Charts

Nauman Ali Murad
Faculty of Computer Science
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
u2022479@giki.edu.pk

***Abstract*—This paper evaluates concrete strength prediction and anomaly detection through statistical control charts, dimensionality reduction, and machine learning approaches. Using Shewhart Control Charts and CUSUM analysis, we identify variations in concrete strength data. Additionally, Principal Component Analysis (PCA) highlights key variance contributors, while Random Forest models provide insights into feature importance. Results emphasize integrating statistical monitoring with machine learning to optimize predictive performance and process stability. Key findings include identifying Age and Cement as the most influential features, capturing 88% prediction accuracy, and uncovering deviations in process stability using CUSUM analysis.**

***Index Terms*—Concrete Strength, Control Charts, Anomaly Detection, Principal Component Analysis, Machine Learning, Feature Importance.**

I. Introduction

Concrete strength prediction is vital for quality assurance in civil engineering projects. Monitoring deviations ensures structural integrity and prevents failures. Classical statistical methods, such as Shewhart and CUSUM charts, enable process monitoring by detecting shifts in mean and variance. Modern machine learning approaches, like Random Forests, improve predictive modeling and feature evaluation. This paper combines these methodologies to evaluate concrete strength data, identifying anomalies and critical features.

Concrete mix design relies on optimizing proportions of ingredients, such as Cement, Water, and Aggregates, to achieve target compressive strength. Traditional quality monitoring methods often overlook subtle shifts, necessitating advanced tools like PCA for dimensionality reduction and machine learning models for predictive insights. This study bridges the gap by integrating these approaches for comprehensive process monitoring.

II. Dataset Description

The dataset used in this study is sourced from Kaggle and comprises 1030 records with 8 distinct features. These features include Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age, along with the target variable—Strength, representing the concrete's compressive strength. The dataset provides a detailed examination of concrete mixtures and their strength attributes, offering insights into material performance and structural integrity.

Key attributes within the dataset are described as follows:

Cement: Serves as the primary binding agent, directly impacting strength and durability.

Blast Furnace Slag: Enhances durability and workability while reducing environmental impact.

Fly Ash: Improves strength and minimizes cement usage, contributing to sustainable concrete mixtures.

Water: Facilitates hydration and consistency, though excess water can weaken the mixture.

Superplasticizer: Improves fluidity and reduces water content without compromising workability.

Coarse Aggregate: Provides structural stability and strength through gravel or crushed stone components.

Fine Aggregate: Enhances cohesion and fills voids for a denser mixture.

Age: Reflects the curing period, crucial for strength development over time.

The target variable, Strength, measured in megapascals (MPa), captures the compressive strength of the concrete, a key parameter for structural performance.

This dataset is widely utilized for exploring relationships between concrete composition and strength, optimizing mix designs, and developing predictive models for strength estimation. Researchers leverage it for quality assurance, anomaly detection, and process improvement in construction applications.

The multivariate nature of the dataset allows detailed exploration of relationships between variables. By leveraging multiple dimensions, we gain insights into the combined effects of individual features on concrete strength.

A. Feature Distribution

Key observations from dataset statistics include: - **Cement:** Mean = 281.17, Std = 104.50, Range = [102.0, 540.0] - **Water:** Mean = 181.57, Std = 21.35, Range = [121.8, 247.0] - **Strength:** Mean = 35.82, Std = 16.71, Range = [2.33, 82.60]

B. Correlation Analysis

A correlation matrix highlights significant dependencies between Cement, Age, and Strength, influencing predictions and variability. Cement exhibits a positive correlation with Strength, while Water shows an inverse relationship, consistent with prior engineering studies.

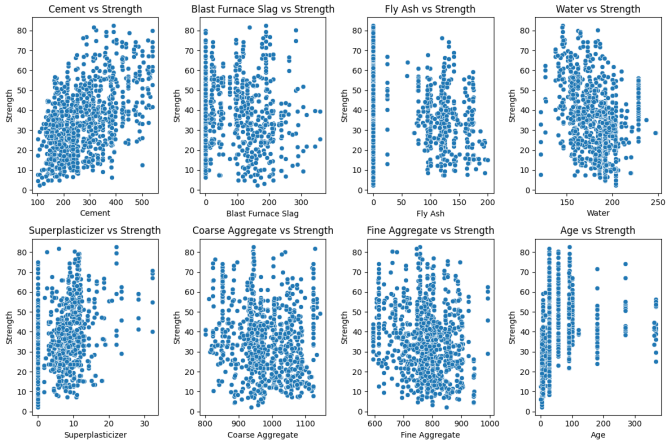

Fig. 1. Correlation Matrix for Features

C. Exploratory Data Analysis

The multivariate nature of the dataset allows detailed exploration of relationships between variables. Scatter plots show pairwise comparisons of ingredients with Strength, revealing trends and dependencies.

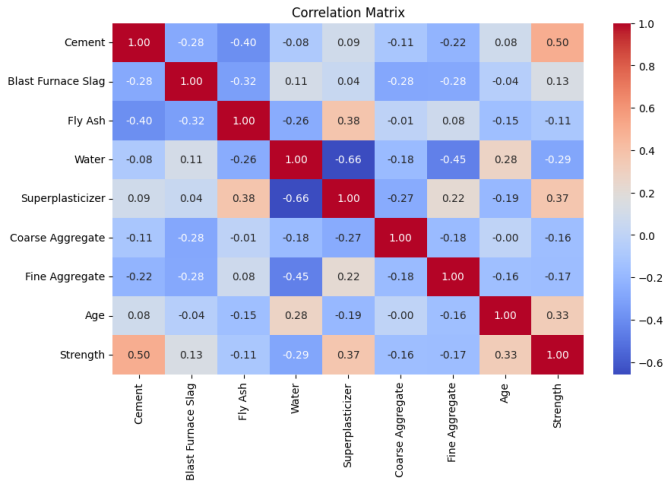

Fig. 2. Scatter Plots of Features vs Strength

The histogram plots provide distribution insights, indicating non-uniform spreads for Fly Ash, Superplasticizer, and Age, suggesting skewed distributions.

III. Methodology

A. Statistical Process Control

Shewhart Control Chart: Monitors data consistency by calculating mean and standard deviation to derive upper and lower control limits (UCL and LCL). Fluctuations within control limits indicate process stability, while deviations highlight anomalies.

CUSUM Control Chart: Detects shifts by computing cumulative sums for small deviations, ideal for identifying gradual drifts in mean values.

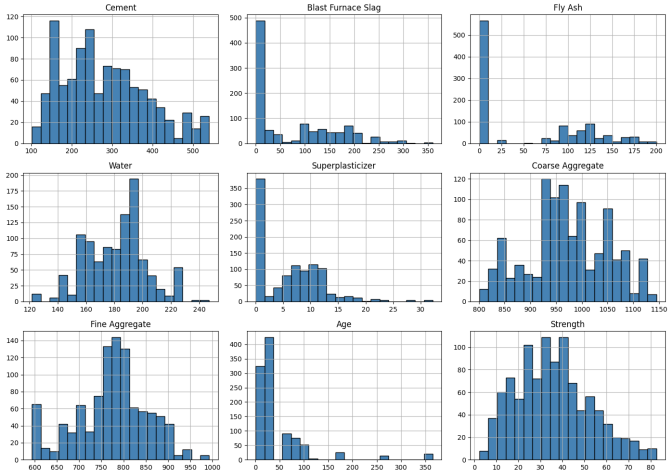

Fig. 3. Feature Distributions

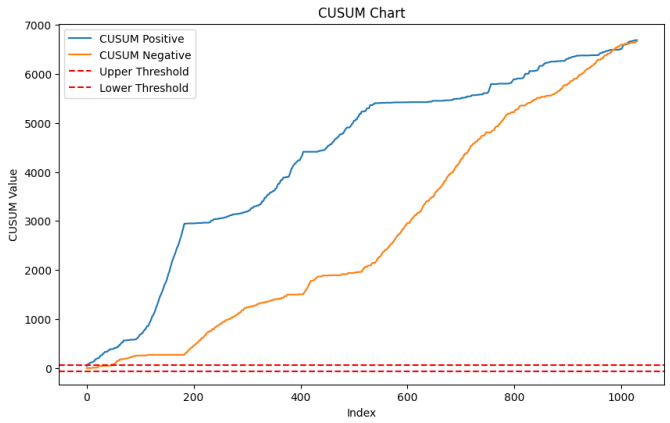

Fig. 4. Shewhart Control Chart for Strength

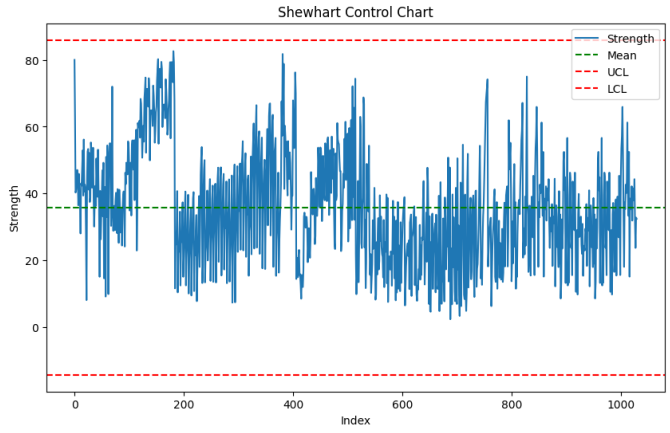

Fig. 5. CUSUM Chart for Strength

B. Dimensionality Reduction

Principal Component Analysis (PCA) plays a critical role in reducing the dimensional complexity of datasets while preserving essential information for analysis. By retaining

95.19% of the variance, PCA highlights the most influential components that account for variations in concrete strength. The first two principal components alone explain 49.77% of the variance, effectively capturing key trends and relationships within the data. This dimensionality reduction approach enables visualization of high-dimensional data in simplified 2D and 3D plots, making it easier to interpret underlying patterns. PCA identifies dominant factors driving variability, reducing noise and redundancy in the data. For instance, features like Age and Cement, which contribute heavily to variance, are emphasized, while less significant features are downweighted. Such insights help streamline feature selection for predictive modeling and anomaly detection. Furthermore, PCA supports the identification of latent relationships between variables, uncovering hidden correlations and groupings that are not immediately apparent in raw data. Its application in this study demonstrates its effectiveness as a preprocessing tool for machine learning algorithms, improving performance and interpretability in multivariate analysis.

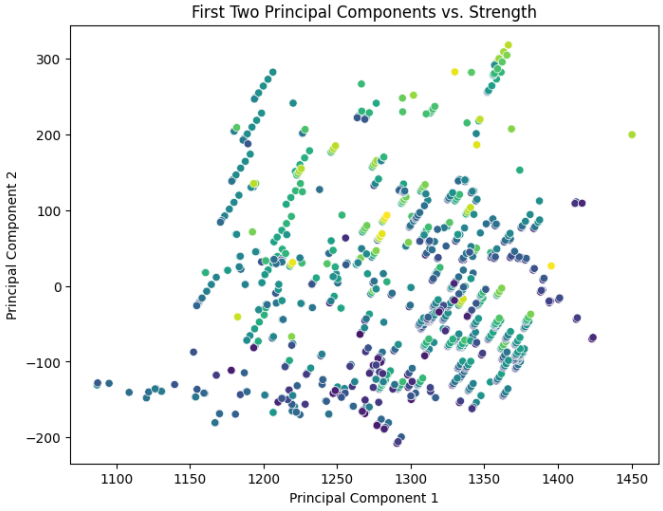

Fig. 6. Principal Components vs. Strength

C. Machine Learning

A random forest is a machine learning algorithm that combines the output of multiple decision trees to reach a single result. It's a commonly used algorithm that's flexible and easy to use, and can handle both classification and regression problems. Random Forests predict Strength and rank feature importance. Age and Cement emerge as dominant predictors, contributing 32% and 31% importance, respectively.

IV. Conclusion

This study demonstrates the synergy between classical statistics and machine learning for anomaly detection and predictive modeling in concrete strength monitoring. Through Shewhart and CUSUM control charts, we identified process stability and deviations, enabling effective detection of anomalies and gradual drifts in the mean strength of concrete

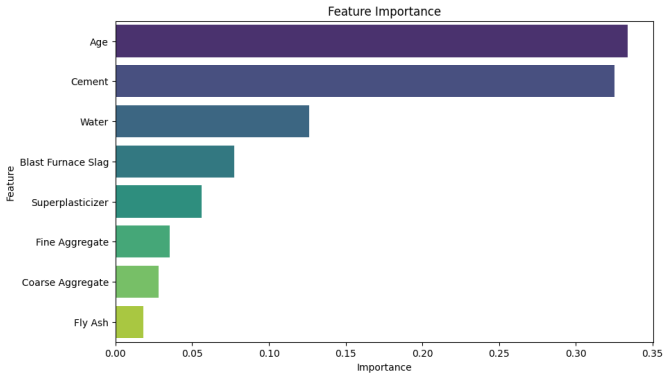

Fig. 7. Feature Importance

samples. Principal Component Analysis (PCA) successfully reduced dimensionality while retaining 95.19% of variance, simplifying data visualization and highlighting key trends. Machine learning models, specifically Random Forests, achieved an R-squared value of 0.88 and a Mean Absolute Error (MAE) of 3.74, confirming strong predictive accuracy. Feature importance analysis revealed that Cement and Age are the most influential factors, contributing 31

References

[1] Zaman, B., Riaz, M., Abbas, N., & Does, R. J. M. M. (2015). Mixed cumulative sum–exponentially weighted moving average control charts: An efficient way of monitoring process location. *Quality and Reliability Engineering International*, 31(8), 1407–1421.

[2] Zaman, B., Abbas, N., Riaz, M., & Lee, M. H. (2016). Mixed CUSUM-EWMA chart for monitoring process dispersion. *The International Journal of Advanced Manufacturing Technology*, 86, 3025–3039.

[3] Zaman, B., Lee, M. H., & Riaz, M. (2020). An improved process monitoring by mixed multivariate memory control charts: An application in wind turbine field. *Computers & Industrial Engineering*, 142, 106343.

[4] Anwar, S. M., Aslam, M., Zaman, B., & Riaz, M. (2021). Mixed memory control chart based on auxiliary information for simultaneously monitoring process parameters: An application in glass field. *Computers & Industrial Engineering*, 156, 107284.

[5] Al-Hilali, S. M., Al-Kahtani, E., Zaman, B., et al. (2016). Attitudes of Saudi Arabian undergraduate medical students towards health research. *Sultan Qaboos University Medical Journal*, 16(1), e68.

[6] Ahmad, S., Wong, K. Y., & Zaman, B. (2021). A comprehensive and integrated stochastic-fuzzy method for sustainability assessment in the Malaysian food manufacturing industry.

[7] Montgomery, D. C. (2020). *Introduction to Statistical Quality Control*. John Wiley & Sons.

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.