Electrocoagulation Turbidity Analysis Report

March 17, 2025

Contents

1 Introduction

Electrocoagulation is a process used to improve water clarity by applying an electrical potential to induce the aggregation and removal of suspended particles. In this study, we analyze a dataset that records:

- **Voltage (V: Volts)** – the electrical potential applied.
- **Time (s: seconds)** – the elapsed time during treatment.
- **Turbidity (NTU: Nephelometric Turbidity Units)** – a measure of water cloudiness.

Our primary objectives are to:

(1) Determine the optimal operating condition at 5 minutes (300 s) by evaluating the turbidity achieved.

(2) Identify the absolute minimum turbidity and the fastest time to achieve it across test runs.

(3) Develop a machine learning model (Random Forest Regressor) to predict turbidity based on Voltage and Time.

Additionally, the dataset is segmented into runs (a new run is defined whenever `Second` equals 2) to better capture repeated experimental conditions.

2 Data Description and Preprocessing

2.1 Dataset Overview

The dataset (`dataset(combined).csv`) contains three columns:

(a) **Voltage (V: Volts)** (numeric): The applied voltage.

(b) **Time (s: seconds)** (numeric): Elapsed time in seconds.

(c) **Turbidity (NTU: Nephelometric Turbidity Units)** (numeric): Measured turbidity.

A snippet of the data is shown below:

Voltage	Time (s)	Turbidity (NTU)
10.0	0	52.3
10.0	60	40.1
10.0	120	29.4
15.0	0	51.8
15.0	60	20.5

2.2 Descriptive Statistics

The dataset contains 240 observations. The summary statistics are:

- **Voltage (V: Volts):** Mean = 19.48 V, Std = 5.91 V, Range = [10.0, 25.0] V.
- **Time (s: seconds):** Mean = 1888.99 s, Std = 1947.20 s, Range = [2, 7502] s.
- **Turbidity (NTU: Nephelometric Turbidity Units):** Mean = 27.34 NTU, Std = 19.71 NTU, Range = [0.31, 107.25] NTU.

These values help us understand the central tendencies and variability of the dataset.

2.3 Data Cleaning and Run Segmentation

Data cleaning involved converting all values to numeric types and dropping rows with missing data. For run segmentation, a new run is defined each time `Second` equals 2. After segmentation, the total number of runs was determined to be **18**. The dataset, with run identifiers, is saved as `dataset_droppednull_with_runs.csv`.

3 Methodology

3.1 Exploratory Data Analysis (EDA)

Our EDA consisted of:

- **Histograms:** Visualizing the distributions of Voltage, Time, and Turbidity (see Figure 1).
- **Scatter Plot:** Plotting Time (s: seconds) versus Turbidity (NTU: Nephelometric Turbidity Units), with points colored by Voltage (V: Volts), to observe trends (see Figure 2).

3.2 Run-Level Analysis

Two key analyses were performed on each run:

1. **Turbidity at 5 Minutes (300 s):**
   - For runs extending to at least 300 s, the data point closest to 300 s was selected.
   - For example, in run 18 (Voltage = 25.0 V: Volts), the closest measurement was at 302 s with a turbidity of 27.45 NTU (Nephelometric Turbidity Units).
2. **Fastest Time to Reach the Absolute Lowest Turbidity:**
   - For each run, the minimum turbidity value and the earliest time it was achieved were recorded.
   - The overall lowest turbidity was 0.31 NTU, achieved in run 7 at 2402 s with a Voltage of 15.0 V.

Figures 3 and 4 visualize these findings.

3.3 Machine Learning Analysis

To predict turbidity, we evaluated two modeling approaches:

1. **Baseline Linear Regression:**
   - This serves as a benchmark model that assumes linear relationships between Voltage, Time, and Turbidity.
   - **Performance Metrics:**
     - R^2 (Coefficient of Determination): 0.167 (indicates that about 16.7% of the variability in turbidity is explained by the model).
     - RMSE (Root Mean Square Error): 16.304 (average magnitude of prediction errors in NTU).
     - MAE (Mean Absolute Error): 12.281 (average absolute error in NTU).

2. **Tuned Random Forest Regressor:**
   - Enhanced by incorporating additional non-linear features:
     - `Second_squared` = $(\text{Time (s)})^2$
     - `Voltage_squared` = $(\text{Voltage (V)})^2$
     - `Voltage_x_Second` = Voltage (V) $\times$ Time (s)
   - Hyperparameters were optimized via GridSearchCV. The best parameters found were:
     - `max_depth = 5`: Maximum depth of each decision tree.
     - `min_samples_split = 2`: Minimum number of samples required to split an internal node.
     - `n_estimators = 200`: Number of decision trees in the ensemble.
   - **Performance Metrics on Test Data:**
     - $R^2 = 0.630$ (indicating that 63% of the variance in turbidity is explained by the model).
     - RMSE = 10.867 NTU.
     - MAE = 7.706 NTU.
   - **5-Fold Cross-Validation (CV) $\mathbf{R^2}$ Scores:** [0.4265, 0.7177, 0.7688, 0.3088, -0.3313] with a mean CV R^2 of 0.378.

Additional visualizations include:

- A predicted versus actual turbidity plot (see Figure 5) showing that the Random Forest model closely tracks the true values.
- A feature importance chart (see Figure 6) indicating that time-related features are highly influential.

3.3.1 Interpretation of Machine Learning Outcomes

Baseline Linear Regression:

- **$R^2 = 0.167$:** Only 16.7% of the variance in turbidity is explained by the linear model, suggesting a weak linear relationship.
- **RMSE = 16.304 NTU:** On average, the predictions deviate from the actual values by 16.3 NTU.
- **MAE = 12.281 NTU:** The average absolute difference between predictions and actual turbidity values is 12.3 NTU.

Tuned Random Forest Regression:

- **Best Parameters:**
  - `max_depth = 5`, `min_samples_split = 2`, `n_estimators = 200`.
- **$R^2 = 0.630$:** The model explains 63% of the variability in turbidity, a significant improvement over the linear model.
- **RMSE = 10.867 NTU:** The average prediction error is reduced to 10.87 NTU.
- **MAE = 7.706 NTU:** The average absolute error is reduced to 7.71 NTU.
- **5-Fold CV R^2 Scores:** The CV scores vary, with a mean of 0.378, indicating some variability in performance across different subsets of the data. A negative score in one fold suggests that in that subset, the model performed worse than a naive baseline.

4 Results and Discussion

4.1 EDA Results

Descriptive Statistics:

- Voltage (V: Volts): Mean = 19.48 V, Std = 5.91 V.
- Time (s: seconds): Mean = 1888.99 s, Std = 1947.20 s.
- Turbidity (NTU: Nephelometric Turbidity Units): Mean = 27.34 NTU, Std = 19.71 NTU.

The histograms (Figure 1) and scatter plot (Figure 2) reveal that while turbidity generally decreases over time, both Time and Voltage are significant factors.

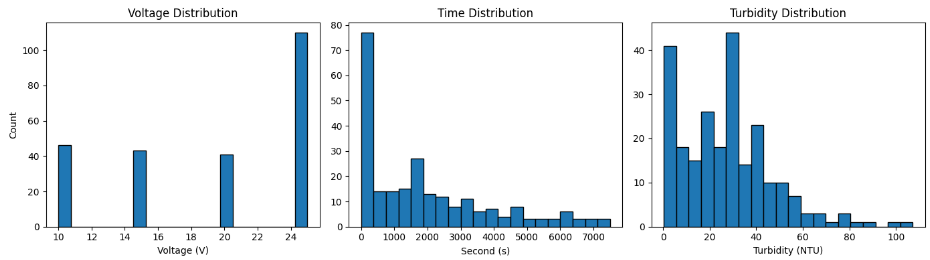

Figure 1: Histograms of Voltage (V: Volts), Time (s: seconds), and Turbidity (NTU: Nephelometric Turbidity Units).

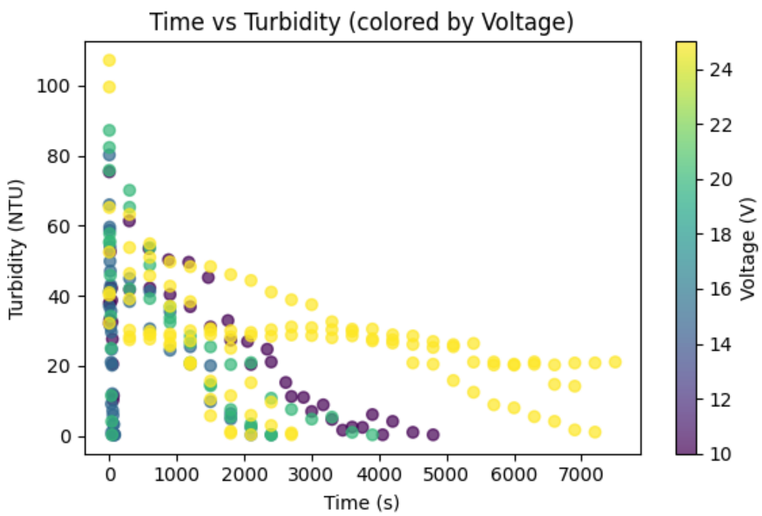

Figure 2: Scatter plot of Time (s: seconds) versus Turbidity (NTU: Nephelometric Turbidity Units), colored by Voltage (V: Volts).

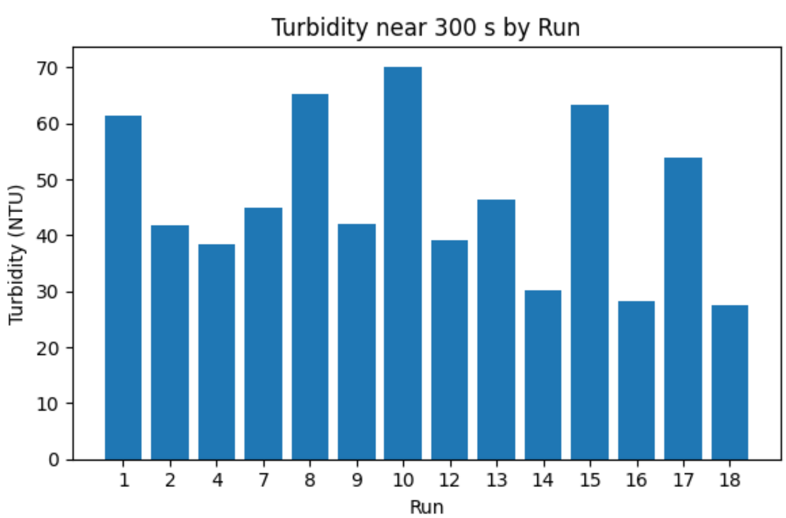

Figure 3: Turbidity at 5 minutes (closest to 300 s) for each run.

4.2 Run-Level Findings

Turbidity at 5 Minutes (300 s):
For runs extending to at least 300 s, the measurement closest to 300 s was extracted. For instance, run 18 (Voltage = 25.0 V: Volts) had a measurement at 302 s with a turbidity of 27.45 NTU (Nephelometric Turbidity Units). See Figure 3 for a visualization.

Fastest Time to Reach Minimum Turbidity:
The overall minimum turbidity of 0.31 NTU was achieved in run 7 at 2402 s with a Voltage of 15.0 V (Volts). Figure 4 summarizes the fastest times across runs.

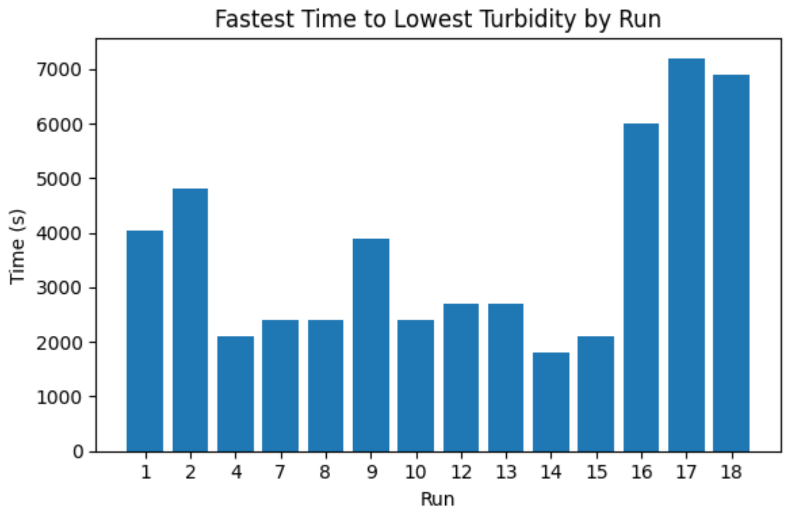

Figure 4: Fastest time to reach minimum turbidity for each run.

4.3 Machine Learning Results

As described in Section 3.3, two models were evaluated.

Baseline Linear Regression:

- $R^2 = 0.167$, RMSE = 16.304 NTU, MAE = 12.281 NTU.

Tuned Random Forest Regression:

- Best Parameters: `max_depth = 5`, `min_samples_split = 2`, `n_estimators = 200`.
- $R^2 = 0.630$, RMSE = 10.867 NTU, MAE = 7.706 NTU.
- 5-Fold CV R^2 Scores: [0.4265, 0.7177, 0.7688, 0.3088, -0.3313] (Mean CV $R^2 = 0.378$).

The detailed interpretation of these results is provided in Section 3.3.1.
Additional figures include:

- **Predicted vs. Actual Plot:** (Figure 5) shows the Random Forest model's predictions closely tracking the actual turbidity values.
- **Feature Importance Chart:** (Figure 6) illustrates that time-related features, including quadratic and interaction terms, are highly influential.

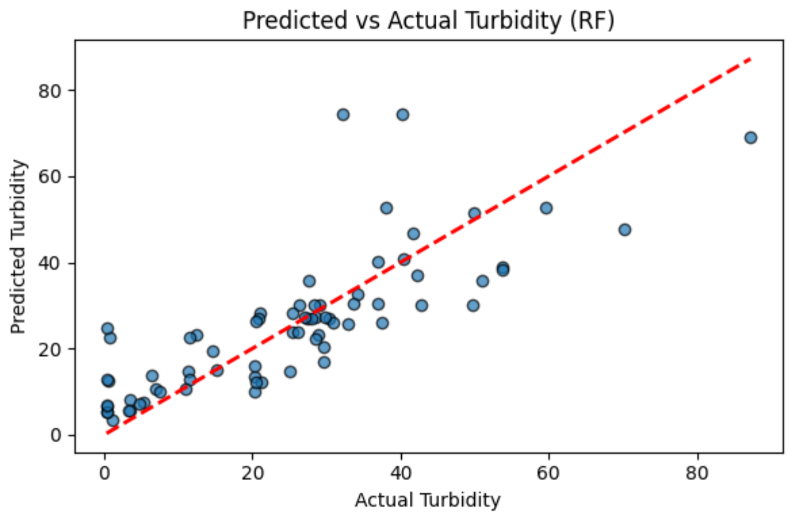

Figure 5: Predicted versus actual turbidity using the tuned Random Forest model.

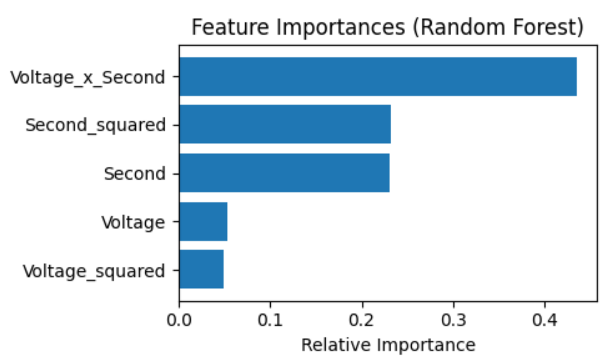

Figure 6: Feature importances from the tuned Random Forest model, highlighting the influence of time-related features.

5 Conclusions

1. **Optimal 5-Minute Turbidity:** The lowest turbidity near 300 s was 27.45 NTU, observed in run 18 at Voltage = 25.0 V (Volts). This serves as an indicator of process performance at the 5-minute mark.

2. **Fastest Achievement of Minimum Turbidity:** The overall minimum turbidity of 0.31 NTU was achieved in run 7 at 2402 s with a Voltage of 15.0 V (Volts), indicating optimal conditions for rapid turbidity reduction.

3. **Machine Learning Insights:** The tuned Random Forest model, incorporating additional non-linear features and optimized hyperparameters, achieved an R^2 of 0.630. This performance suggests that approximately 63% of the variability in turbidity is explained by the model, significantly outperforming the baseline Linear Regression model.

6 Recommendations

Based on the analysis:

1. **For 5-Minute Turbidity Reduction:** Operate at the voltage corresponding to the lowest turbidity at 300 s (e.g., 25.0 V in run 18).

2. **For Quick Clearance:** Adopt conditions similar to run 7 (15.0 V at 2402 s) to achieve the fastest reduction in turbidity.

3. **Further Investigations:** Future studies should include additional parameters (e.g., pH, electrode spacing) and larger datasets to further refine the predictive model.

7 Limitations and Future Work

- The analysis is based on three measured variables; additional process parameters might offer deeper insights.

- The run segmentation strategy (new run when `Second` equals 2) is based on the current experimental design and may require adjustment with further data.

- Variability in instrumentation and measurement techniques could impact the robustness of the conclusions.

8 Appendix: Code and Analysis Summary

The analysis was implemented in Python using:

- **Data Processing:** `pandas` for data cleaning and run segmentation.

- **Visualization:** `matplotlib` (and optionally `seaborn`) for generating histograms, scatter plots, and bar charts.
- **Machine Learning:**
  - A baseline Linear Regression model.
  - A tuned Random Forest Regressor, optimized via GridSearchCV and incorporating additional non-linear features.
  - Performance was evaluated using metrics such as R^2 (Coefficient of Determination), RMSE (Root Mean Square Error), and MAE (Mean Absolute Error).