Predictive Health Scoring for HT Cables

Sagar Kumar

August 13, 2025

Contents

1 Data Pipeline Procedure **3**

1.1 End-to-End Workflow 3

1.2 Step 1: SCADA Data Aggregation 4

1.3 Step 2: Building the Master Table 4

2 ML Model & Health Score **5**

2.1 Fault Prediction Process 5

2.2 The 9 Factors of Cable Health 6

2.3 Health Score Calculation Logic 6

2.4 Understanding the Risk Factors 7

2.5 Final Score Calculation Steps 8

3 Results & Validation **8**

3.1 Model Performance (2024 Validation) 8

1 Data Pipeline Procedure

1.1 End-to-End Workflow

The entire process is an automated pipeline that transforms raw, disconnected data into a single, actionable Health Score for each cable.

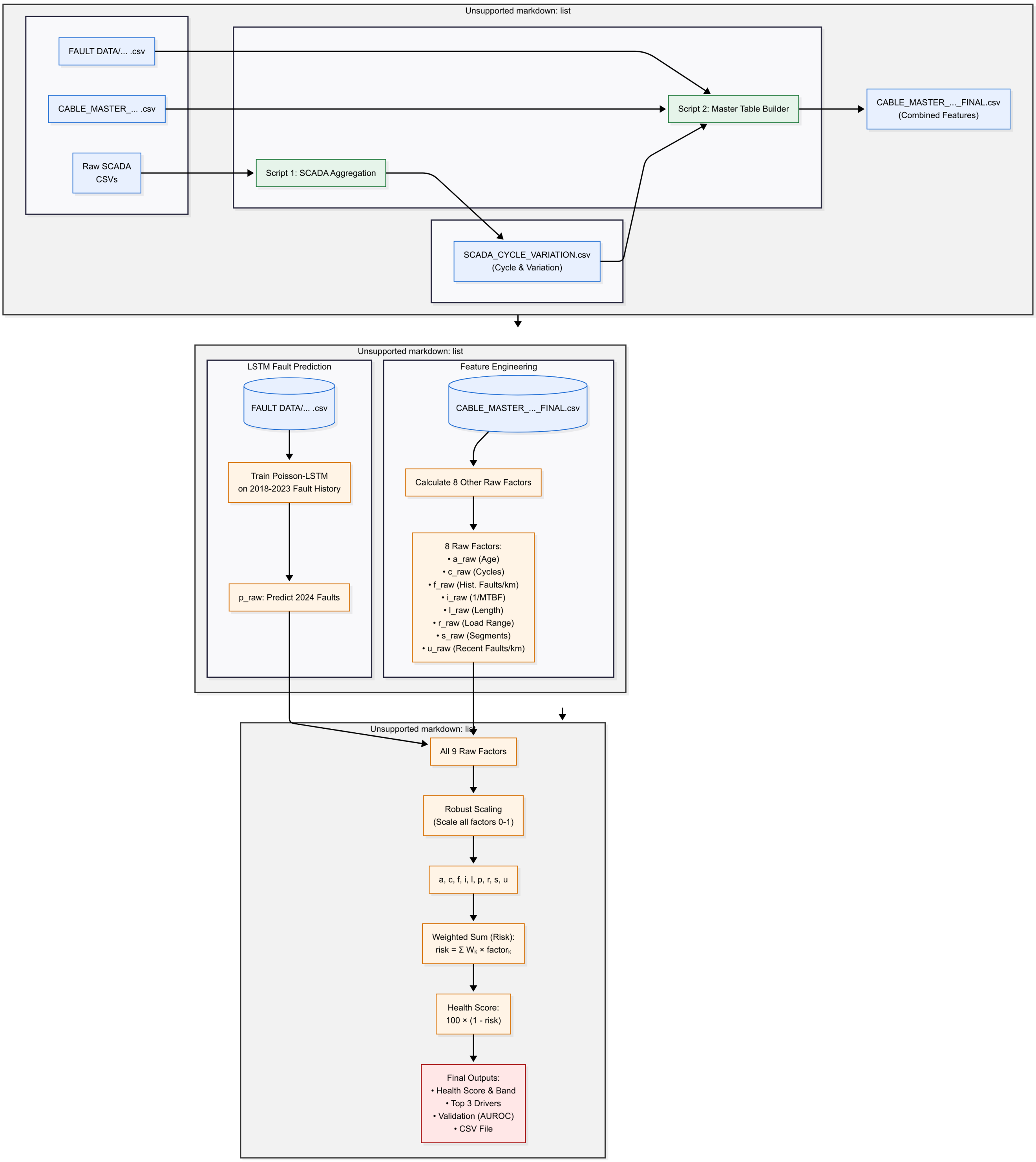
Unsupported markdown: list
FAULT DATA/... .csv
CABLE_MASTER_... .csv
Raw SCADA CSVs
Script 1: SCADA Aggregation
Script 2: Master Table Builder
CABLE_MASTER_..._FINAL.csv (Combined Features)
SCADA_CYCLE_VARIATION.csv (Cycle & Variation)
Unsupported markdown: list
LSTM Fault Prediction
FAULT DATA/... .csv
Train Poisson-LSTM on 2018-2023 Fault History
p_raw: Predict 2024 Faults
Feature Engineering
CABLE_MASTER_..._FINAL.csv
Calculate 8 Other Raw Factors
8 Raw Factors:
• a_raw (Age)
• c_raw (Cycles)
• f_raw (Hist. Faults/km)
• i_raw (1/MTBF)
• l_raw (Length)
• r_raw (Load Range)
• s_raw (Segments)
• u_raw (Recent Faults/km)
Unsupported markdown: list
All 9 Raw Factors
Robust Scaling (Scale all factors 0-1)
a, c, f, i, l, p, r, s, u
Weighted Sum (Risk): risk = Σ W_k × factor_k
Health Score: 100 × (1 - risk)
Final Outputs:
• Health Score & Band
• Top 3 Drivers
• Validation (AUROC)
• CSV File

Figure 1: The project workflow: from raw data ingestion and processing, through ML-based fault prediction, to the final calculation of the 9-factor Health Score.

1.2 Step 1: SCADA Data Aggregation

Input
Raw, time-series SCADA data from multiple CSV files.

Process
A Python script uses parallel processing to:

- Calculate daily peak-to-peak current variation.
- Count high-load ”cycles” per month.

Output
`SCADA_CYCLE_VARIATION.csv`

Key SCADA Aggregations
The script processes the raw current readings to extract two key metrics for each cable per month:

- **Average Daily Variation:** It finds the difference between the maximum and minimum current for each day, and then averages these daily variations over the month. This measures the typical daily stress on the cable.
- **High-Load Cycles:** It counts how many times the current exceeded a high-load threshold during the month. This quantifies how often the cable was under significant strain.

1.3 Step 2: Building the Master Table

Goal: Create One Master File
The second script joins multiple data sources to create the final feature set.

Inputs Joined:

- Cable asset data
- Aggregated SCADA data
- Historical fault records

Key Transformations:

- Aggregating fault counts & MTBF.
- Extracting cable path from comments.
- Merging all tables on Switch ID.

Output
`CABLE_MASTER_COMBINED_FULL_FINAL.csv`

2 ML Model & Health Score

2.1 Fault Prediction Process

This diagram illustrates the process of using historical data to train the ML model, which then forecasts the probability of failure for each individual cable for the year 2024.

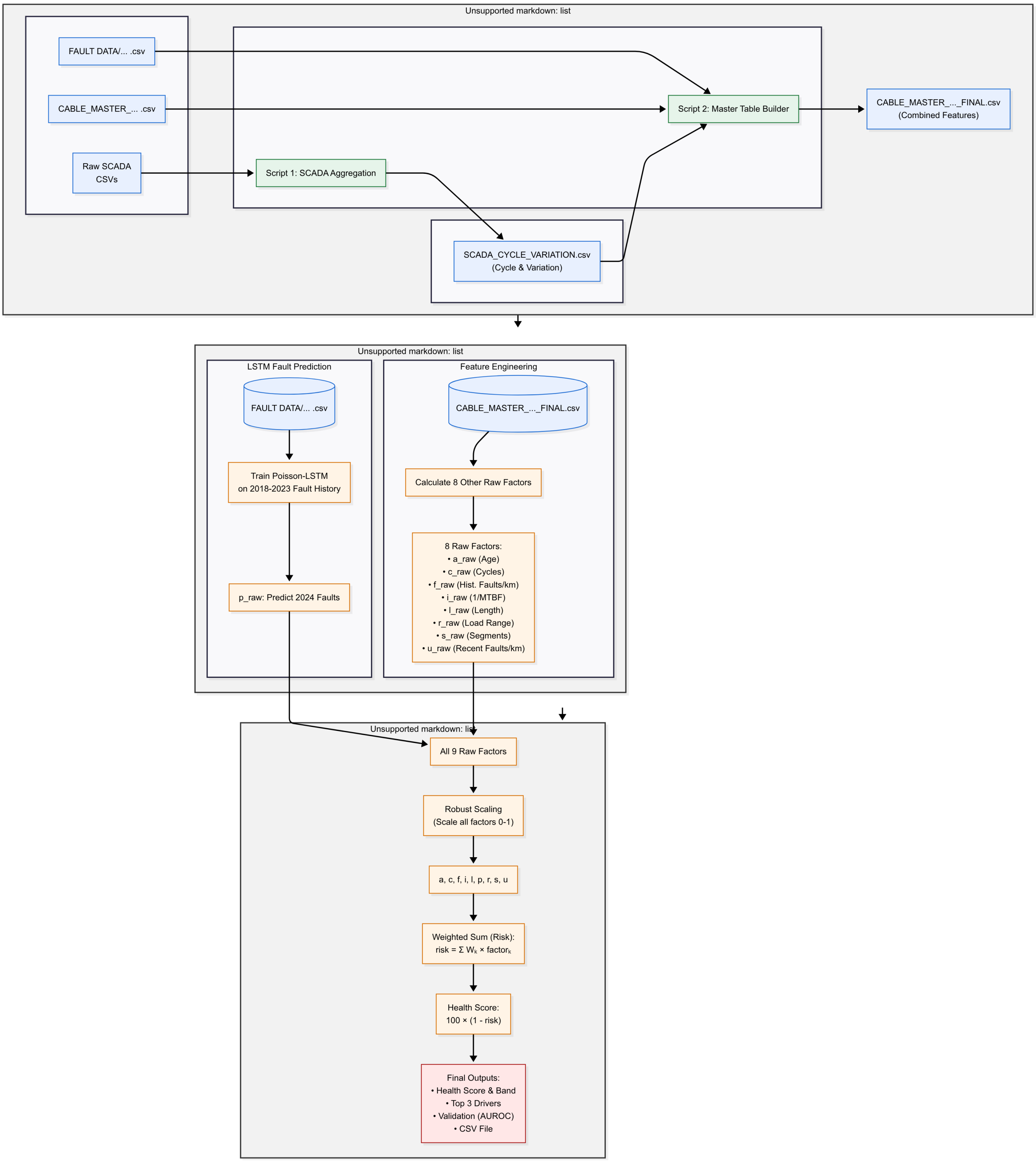
Unsupported markdown: list
FAULT DATA/... .csv
CABLE_MASTER_... .csv
Raw SCADA CSVs
Script 1: SCADA Aggregation
Script 2: Master Table Builder
CABLE_MASTER_..._FINAL.csv (Combined Features)
SCADA_CYCLE_VARIATION.csv (Cycle & Variation)
Unsupported markdown: list
LSTM Fault Prediction
FAULT DATA/... .csv
Train Poisson-LSTM on 2018-2023 Fault History
p_raw: Predict 2024 Faults
Feature Engineering
CABLE_MASTER_..._FINAL.csv
Calculate 8 Other Raw Factors
8 Raw Factors:
• a_raw (Age)
• c_raw (Cycles)
• f_raw (Hist. Faults/km)
• i_raw (1/MTBF)
• l_raw (Length)
• r_raw (Load Range)
• s_raw (Segments)
• u_raw (Recent Faults/km)
Unsupported markdown: list
All 9 Raw Factors
Robust Scaling (Scale all factors 0-1)
a, c, f, i, l, p, r, s, u
Weighted Sum (Risk): risk = Σ W_k × factor_k
Health Score: 100 × (1 - risk)
Final Outputs:
• Health Score & Band
• Top 3 Drivers
• Validation (AUROC)
• CSV File

Figure 2: Process flow for predicting 2024 cable faults using the trained ML model.

2.2 The 9 Factors of Cable Health

The health score is a weighted combination of 9 distinct risk factors. **Asset & Operational Factors:**

- **a**: Age
- **l**: Length (log)
- **s**: Segments / Joints
- **c**: Cyclic Loading (SCADA)
- **r**: Load Range (SCADA)

Fault History & Prediction:

- **f**: Historical Faults / km
- **i**: Interruption Frequency (1/MTBF)
- **u**: Recent Faults (last 8-mo)
- p: **Predicted Faults (ML Model)**

Key Innovation

The model doesn't just rely on the past; it uses an LSTM network to **forecast future faults**, which becomes a key factor ('p').

2.3 Health Score Calculation Logic

This diagram shows how the 9 scaled factors are combined using specific weights to produce a single risk value, which is then converted into the final Health Score from 0 to 100.

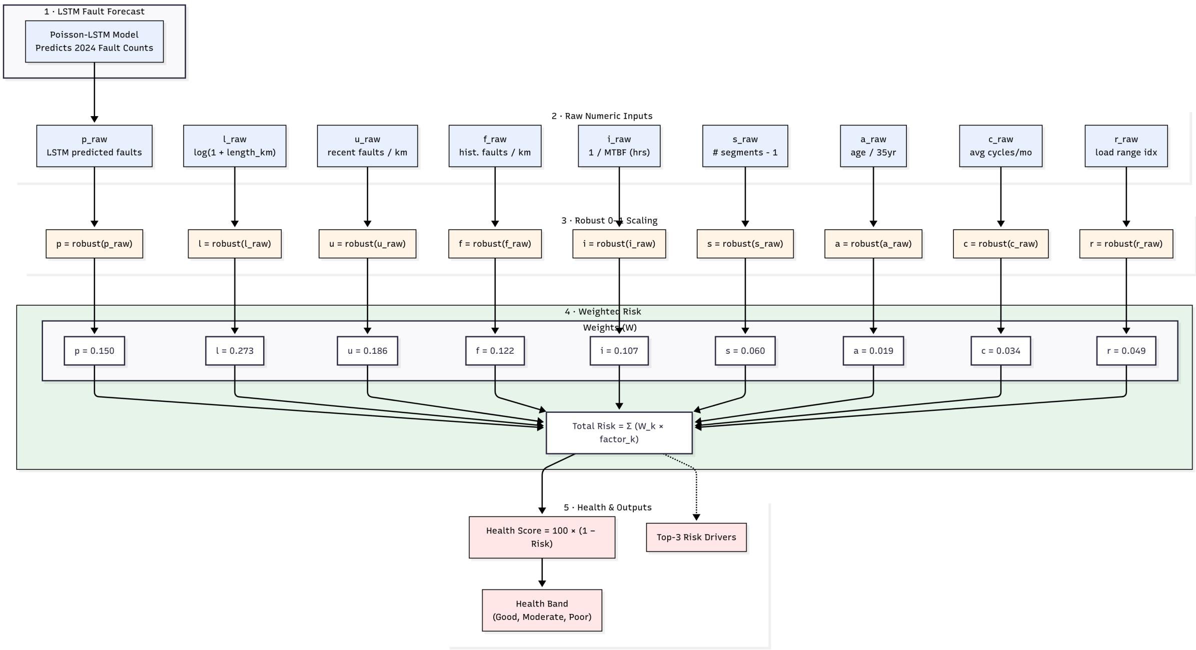

Figure 3: The logic for combining the 9 factors into a final Health Score.

2.4 Understanding the Risk Factors

- **Age (a):** Date installed/35.
- **Length (l):** log(1+length-km).
- **Segments (s):** from HT cable analyze the segment based on the JT NO 1B - JT NO 2B it is the one segment.
- **Cyclic Loading (c):** Frequent, large changes in load cause thermal stress.
  *Example: cycle count is if one cable have 10 median load than how many number of times is going greater than 10 *1.6 or lower than 10/1.6 and aggregate all days within all months. A cable with 20 high-load events (cycle count = 20) in a month is riskier than a cable with a steady load (cycle count = 0).*
- **Load Range (r):** The average daily variation between min/max current indicates thermal stress.
  Example: daily variation = daily daily max - daily min. aggregate and take mean for all each month
- **Faults/km (f):** Normalizes fault history by cable length.
  Example: A 2km cable with 4 faults (2 faults/km) is riskier than a 5km cable with 5 faults (1 fault/km).
- **Interruption Freq (i):** Inverse of Mean Time Between Failures (MTBF).
  Example: A cable failing every 2 years (0.5 freq) is riskier than one failing every 10 years (0.1 freq).
- **Recent Faults (u):** Gives more weight to failures that happened recently.
  Example: A fault 3 months ago contributes more risk than one 3 years ago.
- **Predicted Faults (p):** The ML model's forecast for faults in the next 12 months.

2.5 Final Score Calculation Steps

Step 1: Robust Scaling

- **Sets Boundaries:** It defines the "minimum" as the 5th percentile and the "maximum" as the 95th percentile of the data
- **Clips Outliers:** Any data point below the 5th percentile is treated as the minimum (scaled to 0). Any point above the 95th percentile is treated as the maximum (scaled to 1).
- **Scales Everything Else:** All the data between these percentiles is scaled proportionally to fit within the 0-1 range.

Step 2: Weighted Risk Calculation
A weighted sum of the scaled factors produces a final risk value. $risk = \sum_{k\in\{a,c,f,...\}} W_k \times \text{factor}_k$

Step 3: Health Score
The final score is a simple transformation of the risk, from 0 (highest risk) to 100 (perfect health). Health Score = round(100 × (1 − risk))

Output
`CABLE_HEALTH_SCORE_2024.csv`

3 Results & Validation

3.1 Model Performance (2024 Validation)

From Health Score to a Prediction
To validate the score, we must define a cutoff. We categorize the score into bands and flag cables that are not "Good":

- **Poor (0-32)** or **Moderate (33-62)** → Predicted to Fail (Positive Class)
- **Good (63-100)** → Predicted to be Healthy (Negative Class)

Confusion Matrix
This matrix compares our predictions against the actual faults in 2024. From the values in the matrix, we calculate the model's sensitivity. **Sensitivity (Recall): ≈ 78%**

Key Performance Metrics
The confusion matrix compares our predictions (cables flagged as 'Poor' or 'Moderate') against the actual outcomes in 2024.

- **Accuracy: 75.4%**
  Overall, how often is the model correct?
  Formula: $\frac{TP+TN}{Total} = \frac{54+139}{256}$
- **Precision: 54.5%**
  When it predicts a failure, how often is it right?
  Formula: $\frac{TP}{TP+FP} = \frac{54}{54+45}$

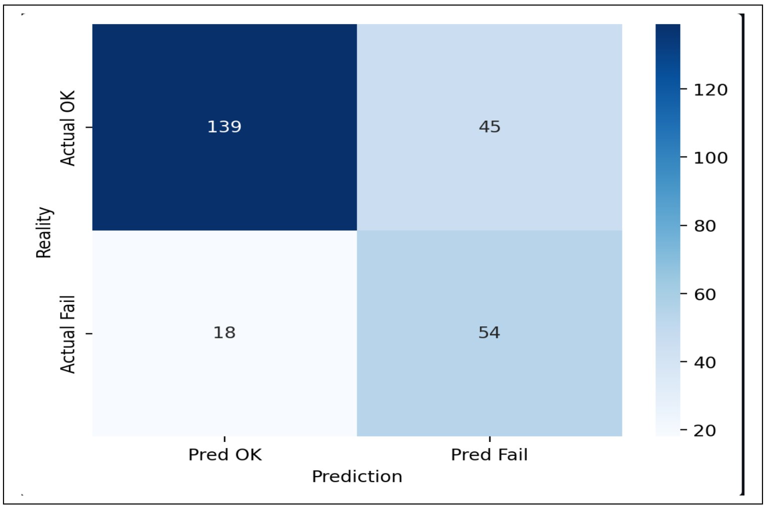

Figure 4: Confusion Matrix for 2024 Validation.

- **Sensitivity (Recall): 75.0%**
  Of all actual failures, what fraction did we identify?
  Formula: $\frac{TP}{TP+FN} = \frac{54}{54+18}$

Intuition
A sensitivity of 78% means: **By focusing on cables with "Poor" or "Moderate" health, we would have successfully preempted 78% of all the failures that occurred in 2024.**