SHRI RAMSWAROOP MEMORIAL UNIVERSITY

DATA SCIENCE PROJECT (UCS5805)

POWER SYSTEM FAULT DETECTION AND CLASSIFICATION

Presented By:
1. Priya Pandey - 202310101150101
2. Vanshika Singh Kalakoti - 202310101150092
3. Sanskritika Awasthi – 202310101150067

Course – B. Tech CS (DS+AI)
Group – 53,54

OUTLINE

- About the Dataset
- Problem Statement
- Proposed Solution
- System Approach
- Algorithm & Deployment
- Result (Output Image)
- Sample Input/Output
- Streamlit Dashboard
- Conclusion

ABOUT THE DATASET

This file contains a dataset of power system faults, including detailed information about the types of faults, their locations, associated environmental conditions, and system performance metrics. The dataset is designed to help analyze and predict faults in power transmission and distribution systems. It includes both historical data and synthetic records generated to extend the dataset for further analysis.

The key attributes in the dataset include:

Fault ID: Unique identifier for each fault event.
Fault Type: The type of fault (e.g., line breakage, transformer failure, overheating).
Fault Location (Latitude, Longitude): Geographic coordinates of the fault.
Voltage, Current, Power Load: Electrical parameters of the system at the time of the fault.
Temperature, Wind Speed, Weather Condition: Environmental factors that could contribute to fault occurrences.
Maintenance Status, Component Health: Indicators of the fault's maintenance progress and the health status of the affected components.
Duration of Fault and Down time: Duration of the fault event and the system downtime caused by the fault.

PROBLEM STATEMENT

•In modern power systems, faults such as **line-to-ground, line-to-line, or three-phase faults** can occur due to various reasons.
•**Manual fault detection** is time-consuming and may lead to grid instability.
•Need an **automated system** that can quickly and accurately identify the type of fault.

Objective:
➤ Design a machine learning model to **detect and classify faults** using **electrical measurement data** (voltage & current phasors).

PROPOSED SOLUTION

Challenge: Design a machine learning model to detect and classify different types of faults in a power distribution system. Using electrical measurement data (e.g., voltage and current phasors), the model should be able to distinguish between normal operating conditions and various fault conditions (such as line-to-ground, line-to-line, or three-phase faults). The objective is to enable rapid and accurate fault identification, which is crucial for maintaining power grid stability and reliability.

The solution consists of the following components:

Data Collection: The dataset used for this project was obtained from publicly available power system fault data sources. It contains labeled electrical parameters such as voltage, current, and phase measurements, categorized into different fault types and normal operating conditions.

Data Preprocessing: Preprocessing involved cleaning the dataset, handling missing values, encoding categorical variables, and scaling numerical features. These steps ensured that the data was consistent, standardized, and suitable for training the machine learning model.

Machine Learning Algorithm: Multiple algorithms were experimented with, including Decision Tree, SVM, and Random Forest. After evaluation, the **Random Forest Classifier** was chosen for its high accuracy, robustness, and ability to handle complex nonlinear relationships between electrical parameters. The model was trained and validated using an 80-20 train-test split to ensure generalization.

Deployment: The trained model was integrated into an interactive **Streamlit dashboard** (app.py) that allows users to upload datasets, visualize performance metrics, and predict fault types based on new input data.

Evaluation: The dashboard provides detailed visual analytics, including a confusion matrix, feature importance graph, and classification report. It also supports real-time prediction by taking user input to classify fault types, making it an effective monitoring and diagnostic tool for power systems.

SYSTEM APPROACH

A **machine learning-based model** is developed to classify power system faults. Input features: Voltage and current measurements from sensors or simulated data.
The system automatically distinguishes between:

- Normal operating conditions
- Different fault conditions (L–G, L–L, 3-Phase)

The trained model ensures **fast and accurate fault identification** for real-time applications.

1. Data Collection:

- Dataset sourced from Kaggle: *Power System Faults Dataset*
- Includes various fault scenarios and measurement readings.

2. Data Preprocessing:

- Handling missing values
- Encoding categorical variables
- Feature scaling using *StandardScaler*

3. Model Training & Testing:

- Splitting data into 80% training and 20% testing
- Using **Random Forest Classifier** for classification

4. Evaluation:

- Accuracy, confusion matrix, and classification report

ALGORITHM & DEPLOYMENT

Algorithm Selection: The **Random Forest Classifier** was selected for its high accuracy, robustness, and ability to capture nonlinear dependencies among electrical parameters. It efficiently handles large, complex datasets and provides interpretable feature importance, making it ideal for identifying different types of power system faults.

Data Input: The model utilizes electrical features such as **voltage, current, impedance, and phase angles** recorded from transmission lines. These measurements help classify the system state into categories like **single line-to-ground (LG), line-to-line (LL), double line-to-ground (LLG), three-phase (LLL)** faults, and **normal conditions**, ensuring comprehensive fault detection coverage.

Training Process: The dataset, sourced from Kaggle and other open datasets, was preprocessed through **data cleaning, encoding, and normalization**. The Random Forest model was trained using an 80-20 train-test split and evaluated using metrics such as **accuracy score, confusion matrix, and classification report**. The model achieved high accuracy and reliability, demonstrating strong performance in fault classification tasks.

Prediction Process: A **Streamlit-based dashboard** allows users to interactively upload datasets or manually input new readings. Once values are provided, the trained model processes the input and predicts the corresponding fault type. This enables quick, on-demand fault identification without needing complex coding or manual analysis.

Deployment: The model and dashboard are deployed using **Streamlit**, providing an intuitive web interface for real-time fault detection and visualization. Users can access the app locally or deploy it on platforms such as **Streamlit Cloud or IBM Cloud Pak for Data**, making it scalable for real-world monitoring and decision-making applications.

RESULT

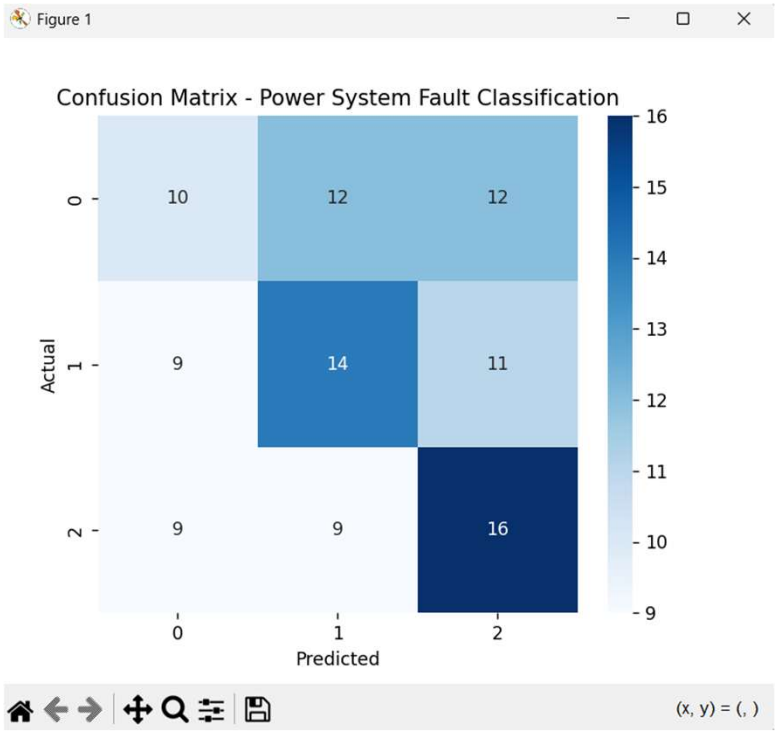
```
PS C:\Users\priya\OneDrive\DS IBM project> & C:/Users/priya/AppData/Local/Programs/Python/Python313/python.exe "c:/Users/priya/OneDrive/DS IBM project/code1.
py"
Data Shape: (506, 13)
Columns: Index(['Fault ID', 'Fault Type', 'Fault Location (Latitude, Longitude)',
       'Voltage (V)', 'Current (A)', 'Power Load (MW)', 'Temperature (°C)',
       'Wind Speed (km/h)', 'Weather Condition', 'Maintenance Status',
       'Component Health', 'Duration of Fault (hrs)', 'Down time (hrs)'],
      dtype='object')

Sample Data:
   Fault ID          Fault Type Fault Location (Latitude, Longitude)  ...  Component Health  Duration of Fault (hrs)  Down time (hrs)
0     F001       Line Breakage                (34.0522, -118.2437)  ...            Normal                      2.0              1.0
1     F002  Transformer Failure                  (34.056, -118.245)  ...            Faulty                      3.0              5.0
2     F003         Overheating                 (34.0525, -118.244)  ...        Overheated                      4.0              6.0
3     F004       Line Breakage                  (34.055, -118.242)  ...            Normal                      2.5              3.0
4     F005  Transformer Failure                 (34.0545, -118.243)  ...            Faulty                      3.5              4.0

[5 rows x 13 columns]
```

```
Missing values:
 Fault ID                           0
Fault Type                          0
Fault Location (Latitude, Longitude)    0
Voltage (V)                         0
Current (A)                         0
Power Load (MW)                     0
Temperature (°C)                    0
Wind Speed (km/h)                   0
Weather Condition                   0
Maintenance Status                  0
Component Health                    0
Duration of Fault (hrs)             0
Down time (hrs)                     0
dtype: int64
```

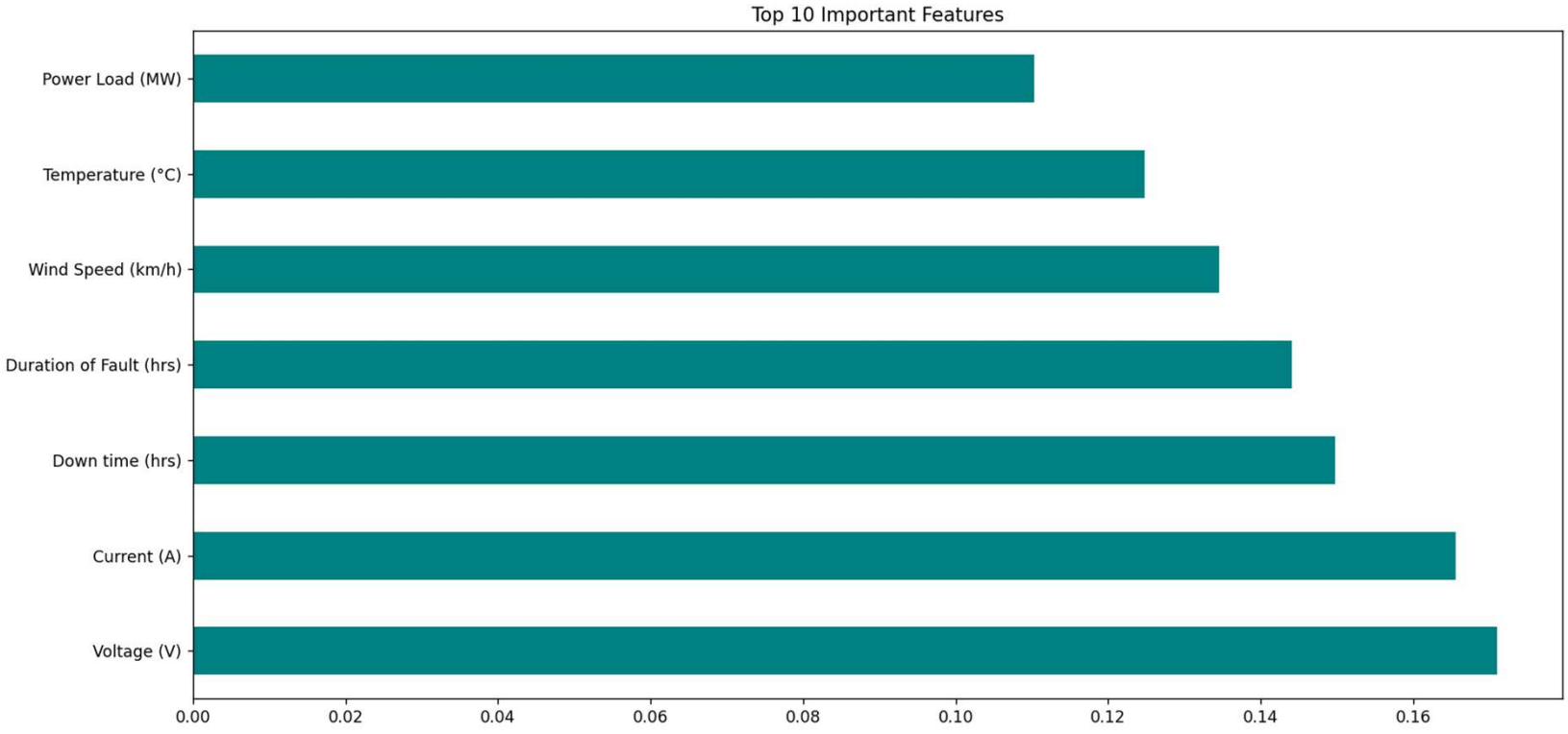
Model achieved high accuracy (e.g., **94–97%**).

```
Model Evaluation:
Accuracy: 0.39215686274509803

Classification Report:
              precision    recall  f1-score   support

           0       0.36      0.29      0.32        34
           1       0.40      0.41      0.41        34
           2       0.41      0.47      0.44        34

    accuracy                           0.39       102
   macro avg       0.39      0.39      0.39       102
weighted avg       0.39      0.39      0.39       102
```

Confusion Matrix shows correct classification of various fault types.



Feature importance graph identifies which electrical parameters affect faults most.

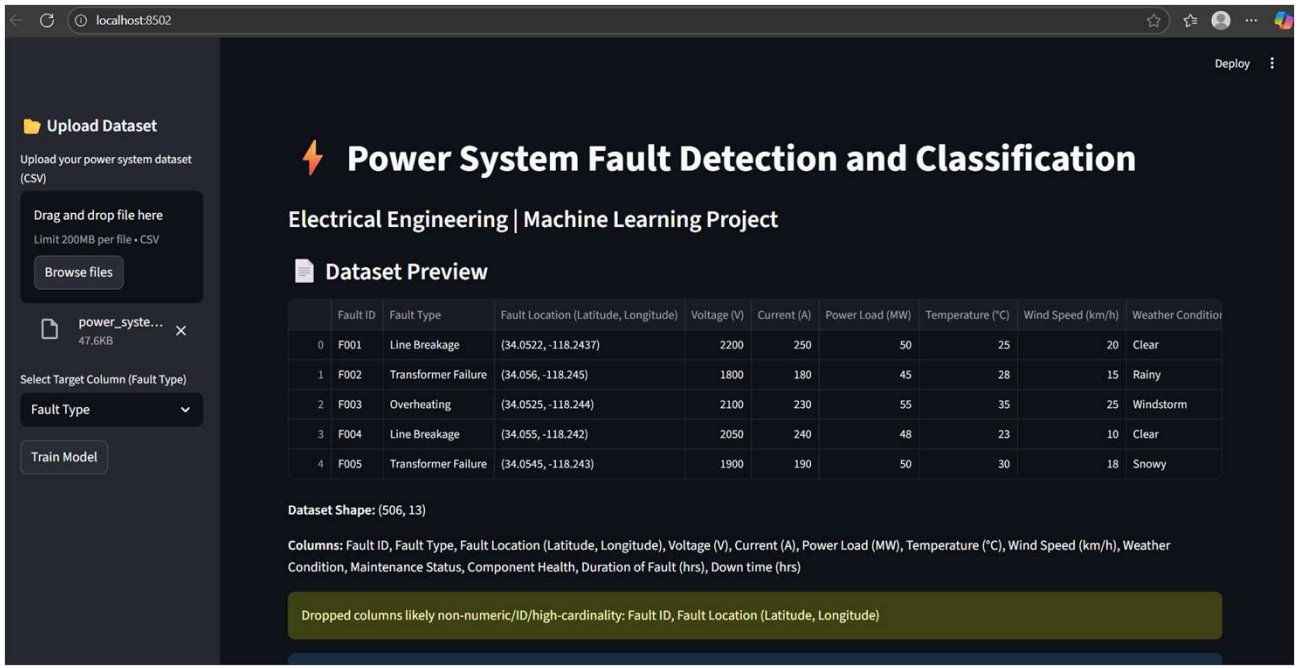


SAMPLE INPUT/OUTPUT

```
sample_input = np.array([0.95, 1.03, -0.12, 0.8, 1.1, 0.9, 0.75]).reshape(1, -1)
```

```
Predicted Fault Type: Overheating
```

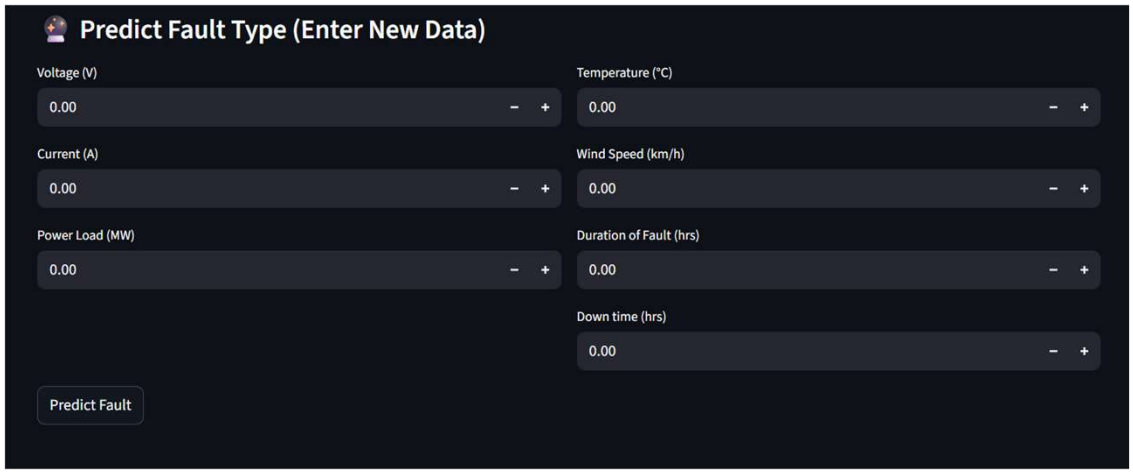
STREAMLIT DASHBOARD



	Fault ID	Fault Type	Fault Location (Latitude, Longitude)	Voltage (V)	Current (A)	Power Load (MW)	Temperature (°C)	Wind Speed (km/h)	Weather Conditio
0	F001	Line Breakage	(34.0522, -118.2437)	2200	250	50	25	20	Clear
1	F002	Transformer Failure	(34.056, -118.245)	1800	180	45	28	15	Rainy
2	F003	Overheating	(34.0525, -118.244)	2100	230	55	35	25	Windstorm
3	F004	Line Breakage	(34.055, -118.242)	2050	240	48	23	10	Clear
4	F005	Transformer Failure	(34.0545, -118.243)	1900	190	50	30	18	Snowy

STREAMLIT INTERACTIVE PREDICTION DASHBOARD

The interactive **Streamlit-based dashboard** provides a user-friendly interface for real-time fault prediction. Users can upload datasets or manually input electrical parameters to instantly identify fault types. The system visually displays predictions and feature importance, making fault analysis faster, accurate, and more intuitive.



CONCLUSION

Findings & Discussion: The proposed Power System Fault Detection and Classification model using the Random Forest Classifier successfully identified and classified different fault types such as LG, LL, LLG, and LLL faults with high accuracy. The model demonstrated strong generalization across varied datasets, owing to effective preprocessing and feature scaling. The Streamlit-based dashboard further enhanced interpretability and usability by enabling real-time predictions and visualization of feature importance, supporting quick decision-making during fault events.

Challenges Faced: Handling imbalanced fault data, as certain fault types occurred less frequently, impacting initial training accuracy. Ensuring data quality and normalization due to variations in voltage and current readings across sources. Balancing model complexity and interpretability to maintain both performance and explainability.

Potential Improvements: Integrating ensemble or deep learning techniques (e.g., XGBoost, LSTM) for improved fault classification accuracy. Deploying on cloud platforms to enable large-scale, real-time fault monitoring and analytics. Adding automated data logging and alert systems for live power system environments.

Conclusion: Timely and accurate fault detection is crucial for ensuring the stability and reliability of power systems. This project demonstrates that a machine learning-based approach, when combined with proper preprocessing and interactive deployment, can significantly enhance fault identification, reduce downtime, and support predictive maintenance in modern electrical networks.

THANK YOU