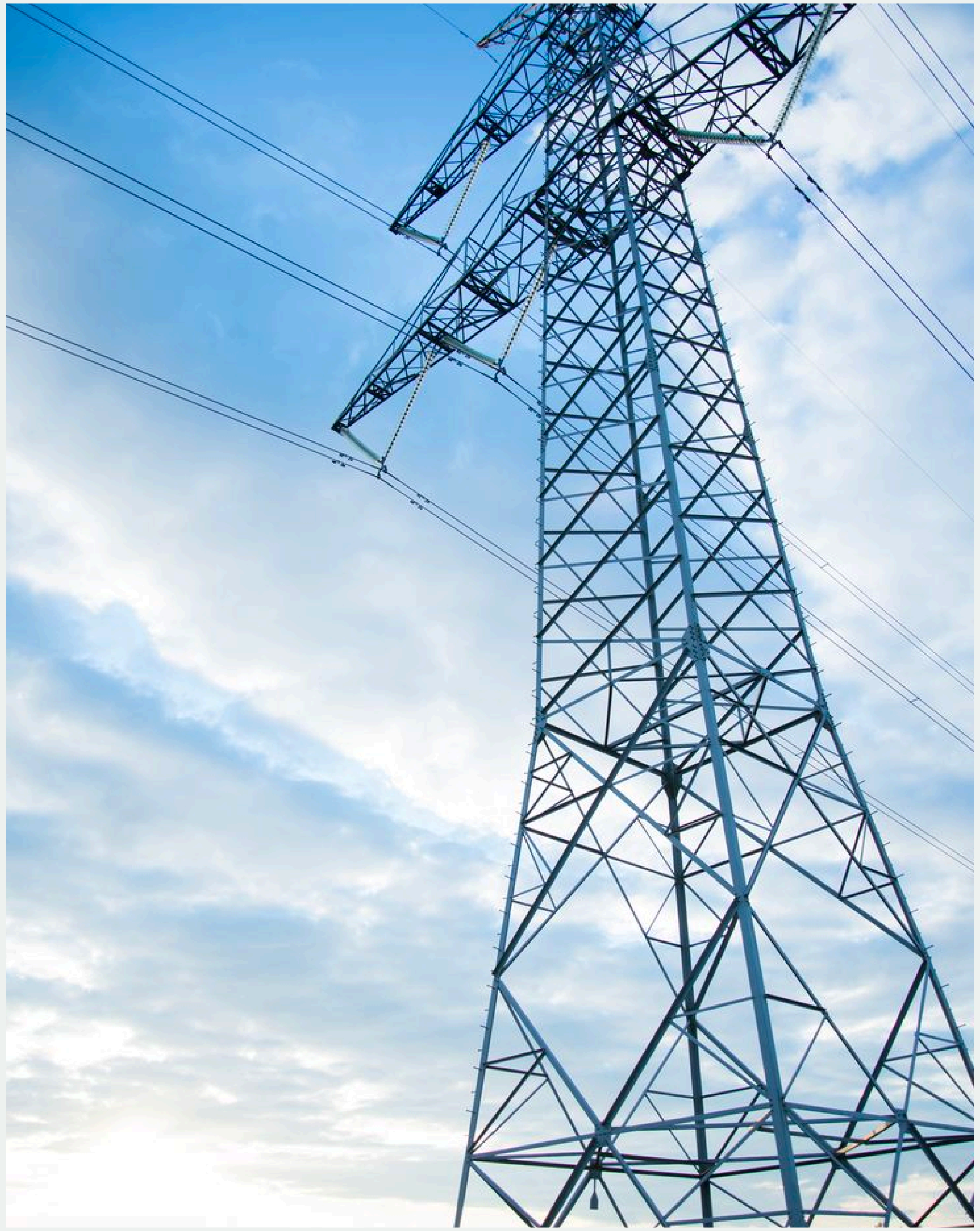
Machine Learning 1

Predicting Electrical Faults in Transmission Lines

Prediction and Classification of Faults in Electrical Systems



Servanez, GPT
LT4

Objectives:

- Discuss an **overview of electrical system** and **electrical faults**
- Discuss **significance of fault classification**
- Create a machine learning model that **accurately predicts** the type of **faults** in an electrical system
- Utilize **f1-score** as metric and utilize **class-wise analysis**

The Electrical Grid

Generation

Generates electricity from power plants

Transmission

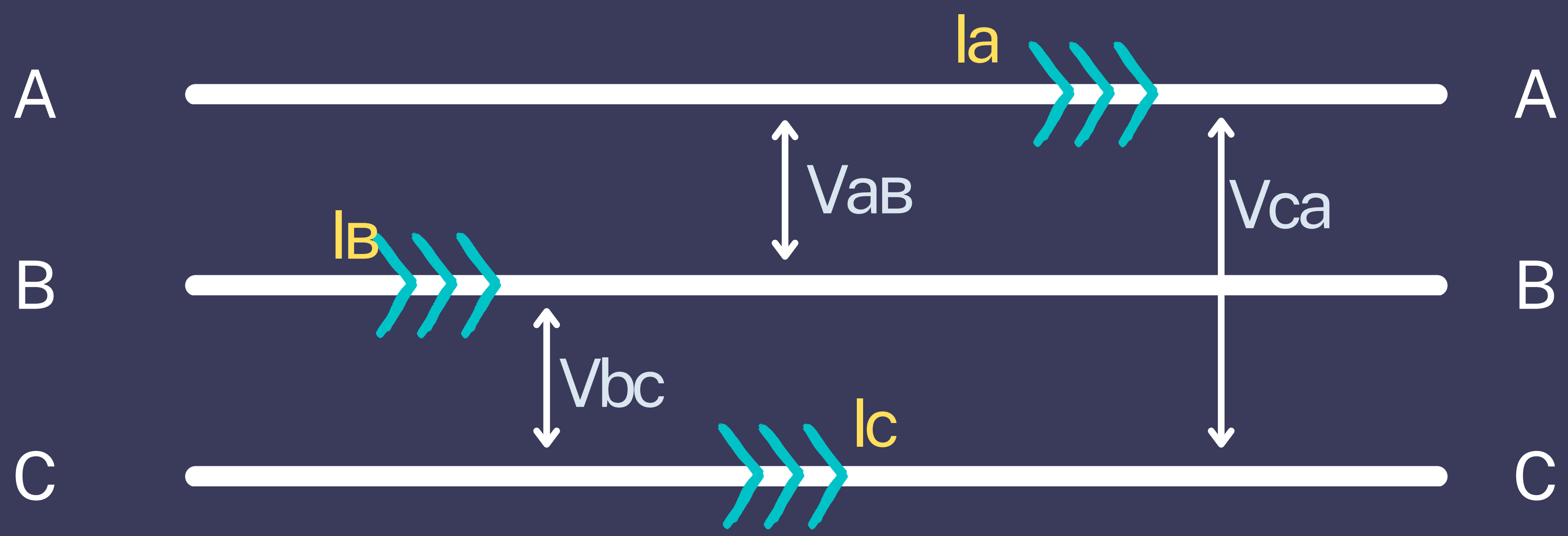
Transmits electricity across long distances

Distribution

Distributes electricity to users

The Electrical Grid

Transmission Line



Line Representation

Electrical Faults

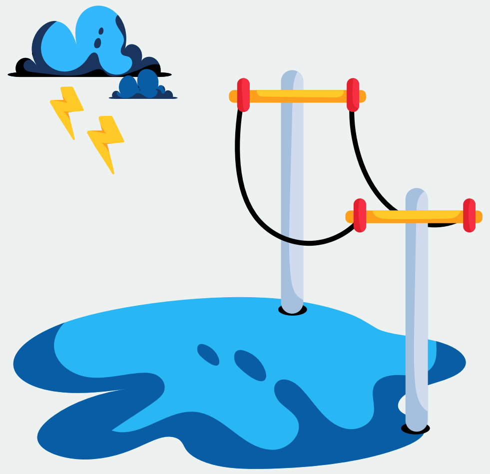
- Electrical Faults
  - **Open Circuit Fault**
    - Open Conductor Faults
  - **Short Circuit Fault**
    - **Asymmetrical Faults**
      - Line-to-Line
      - Line-to-Ground
      - Line-to-Line-to-Ground
    - **Symmetrical Faults**
      - Line-to-Line-to-Line
      - Line-to-Line-Line-Ground

Line-to-ground fault

Impact of Faults

Safety Hazards

Equipment Damage



Service Interruption

Fault Classification Model

Enhanced Safety Measures

Improved System Reliability

Informed Protection System Design

Customer Satisfaction

Preventive Actions

Optimized Maintenance

Grid Modernization

Fault Classification Model

Leveraging **machine learning** to **enhance the precision** of **fault classification**

<u>Goal:</u>

95%

F1 Score

Model Used:

- KNN Classifier,
- Logistic Regression ,
- SVM,
- Decision Trees,
- Random Forest,
- Gradient Boosting,
- XGBoost
- LightGBM
- CatBoost

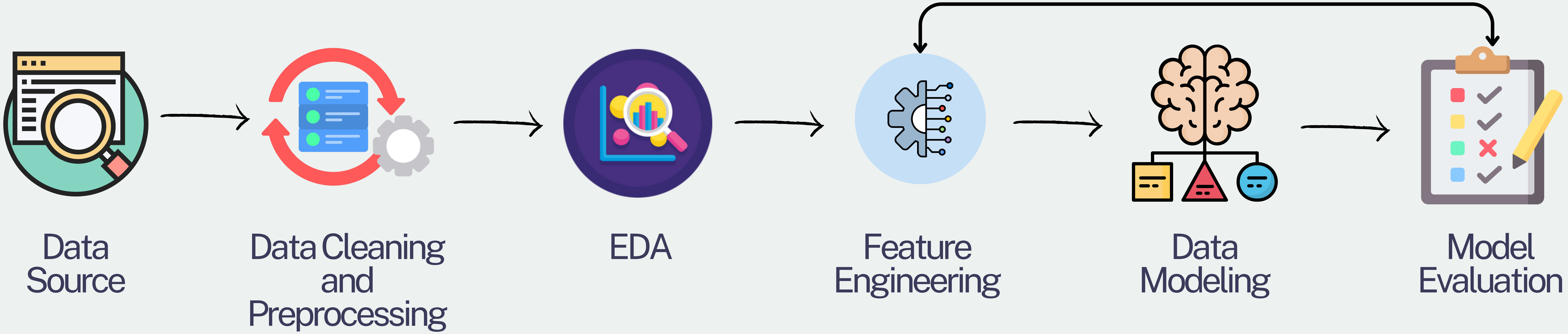
Metric Used:

Precision, Recall, F1 Score

Fault Classes

Fault	A	B	C	G	Class
No fault	0	0	0	0	0000
Fault in Line A	1	0	0	0	1000
Fault in Line B	0	1	0	0	0100
Fault in Line C	0	0	1	0	0010
LG fault (Phase A and Ground)	1	0	0	1	1001
LG fault (Phase B and Ground)	0	1	0	1	0101
LG fault (Phase C and Ground)	0	0	1	1	0011
LL fault (Phase A and Phase B)	1	1	0	0	1100
LL fault (Phase A and Phase C)	1	0	1	0	1010
LL fault (Phase B and Phase C)	0	1	1	0	0110
LLG Fault (Phases A, B and Ground)	1	1	0	1	1101
LLG Fault (Phases A, C and Ground)	1	0	1	1	1011
LLG Fault (Phases B, C and Ground)	0	1	1	1	0111
LLL Fault(All three phases)	1	1	1	0	1110
LLLG fault (Three phase symmetrical fault)	1	1	1	1	1111

Methodology



Dataset

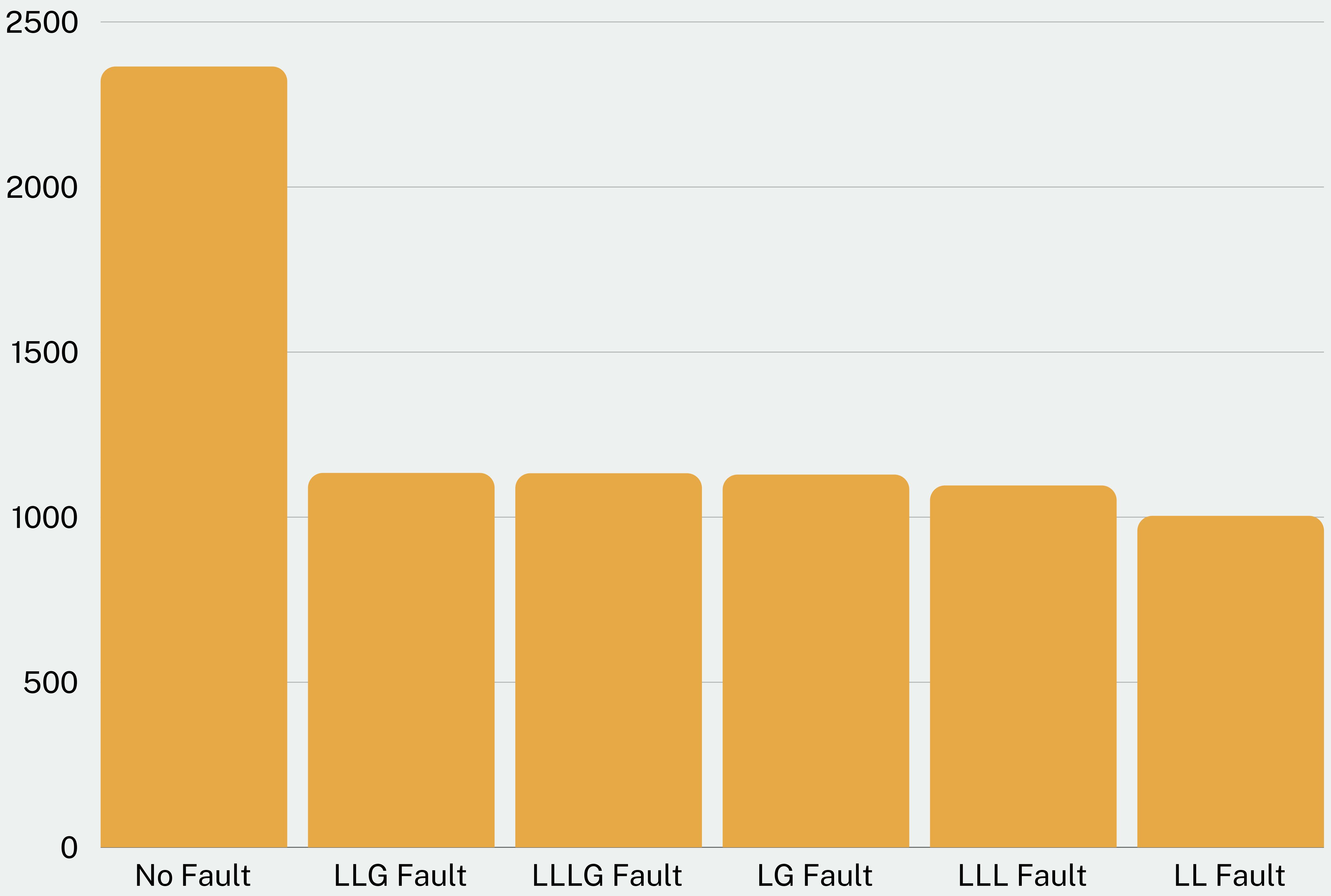
10 features and 7,681 observations

	G	C	B	A	Ia	Ib	Ic	Va	Vb	Vc
0	1	0	0	1	-151.291812	-9.677452	85.800162	0.400750	-0.132935	-0.267815
1	1	0	0	1	-336.186183	-76.283262	18.328897	0.312732	-0.123633	-0.189099
2	1	0	0	1	-502.891583	-174.648023	-80.924663	0.265728	-0.114301	-0.151428
3	1	0	0	1	-593.941905	-217.703359	-124.891924	0.235511	-0.104940	-0.130570
4	1	0	0	1	-643.663617	-224.159427	-132.282815	0.209537	-0.095554	-0.113983

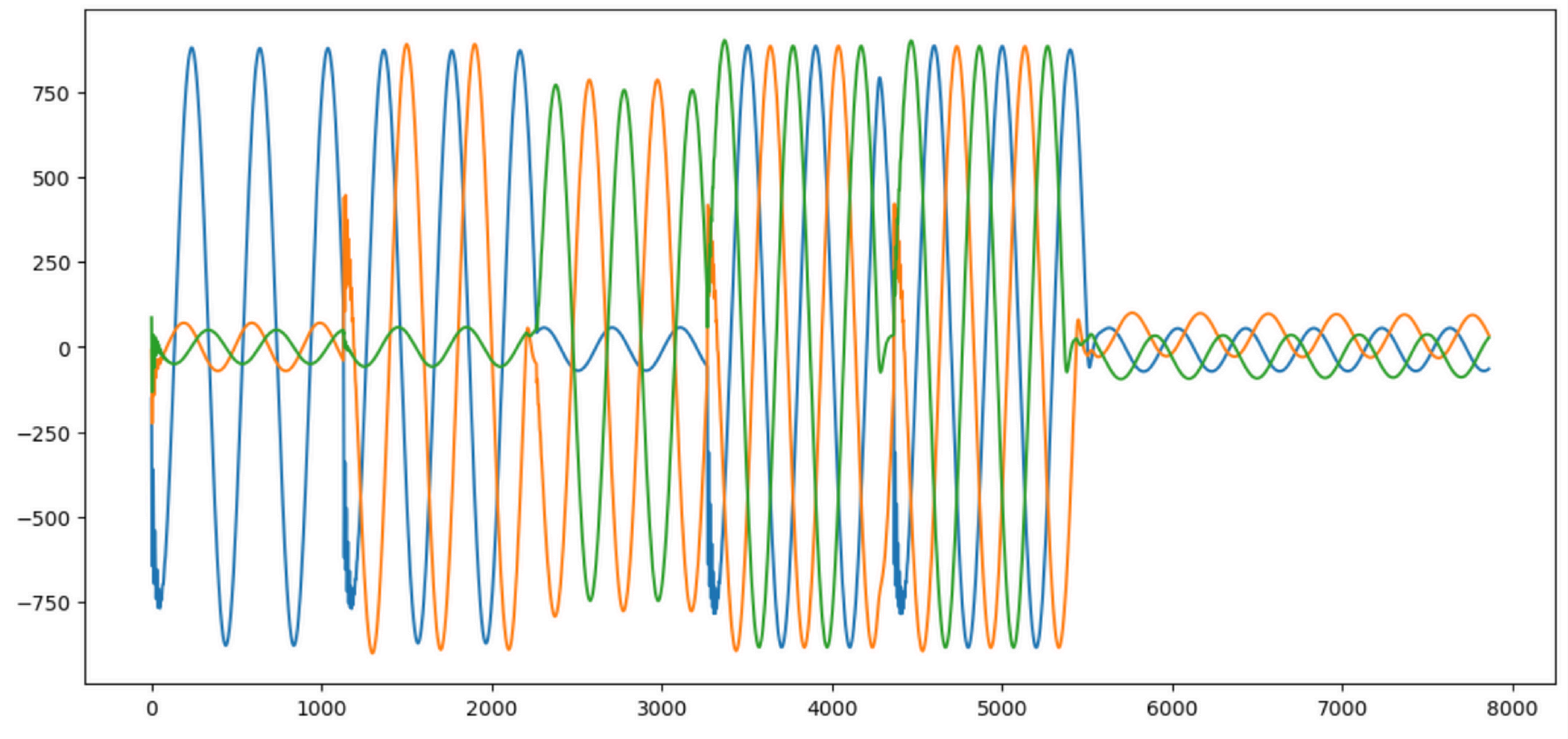
fault_class_mapping = {
'**No fault**': 0,
'**LLG Fault, involving Phases A, B, and Ground**': 1,
'**LLLG Fault, a three-phase symmetrical fault**': 2,
'**LG Fault, between Phase A and Ground**': 3,
'**LLL Fault, involving all three phases**': 4,
'**LL Fault, between Phases B and C**': 5
}

	A	B	C	G	Ia	Ib	Ic	Va	Vb	Vc	Fault Type	Fault Class	Fault Numeric Class
0	1	0	0	1	-151.291812	-9.677452	85.800162	0.400750	-0.132935	-0.267815	1001	LG Fault, between Phase A and Ground	3
1	1	0	0	1	-336.186183	-76.283262	18.328897	0.312732	-0.123633	-0.189099	1001	LG Fault, between Phase A and Ground	3
2	1	0	0	1	-502.891583	-174.648023	-80.924663	0.265728	-0.114301	-0.151428	1001	LG Fault, between Phase A and Ground	3
3	1	0	0	1	-593.941905	-217.703359	-124.891924	0.235511	-0.104940	-0.130570	1001	LG Fault, between Phase A and Ground	3
4	1	0	0	1	-643.663617	-224.159427	-132.282815	0.209537	-0.095554	-0.113983	1001	LG Fault, between Phase A and Ground	3

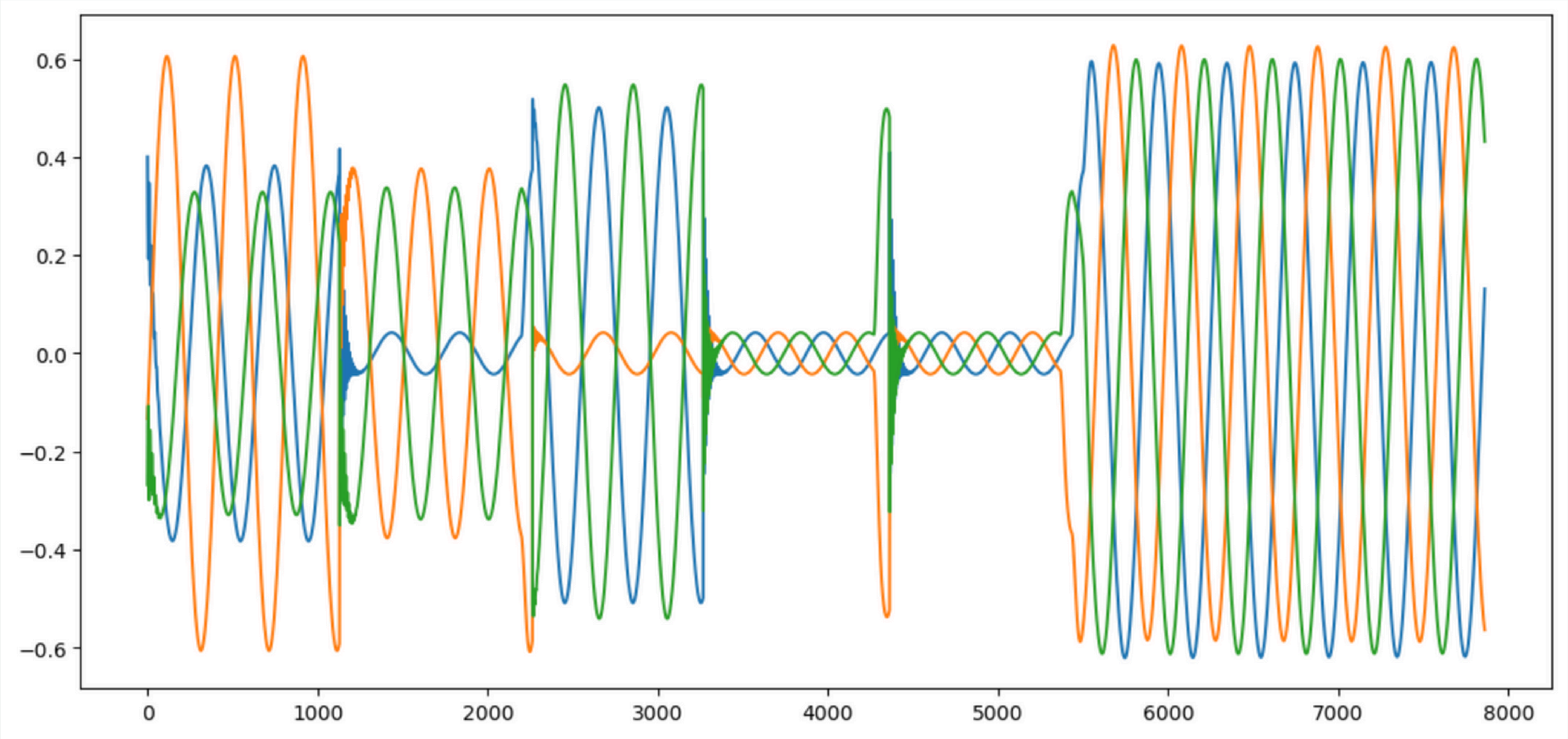
Fault Classes



Current Distribution

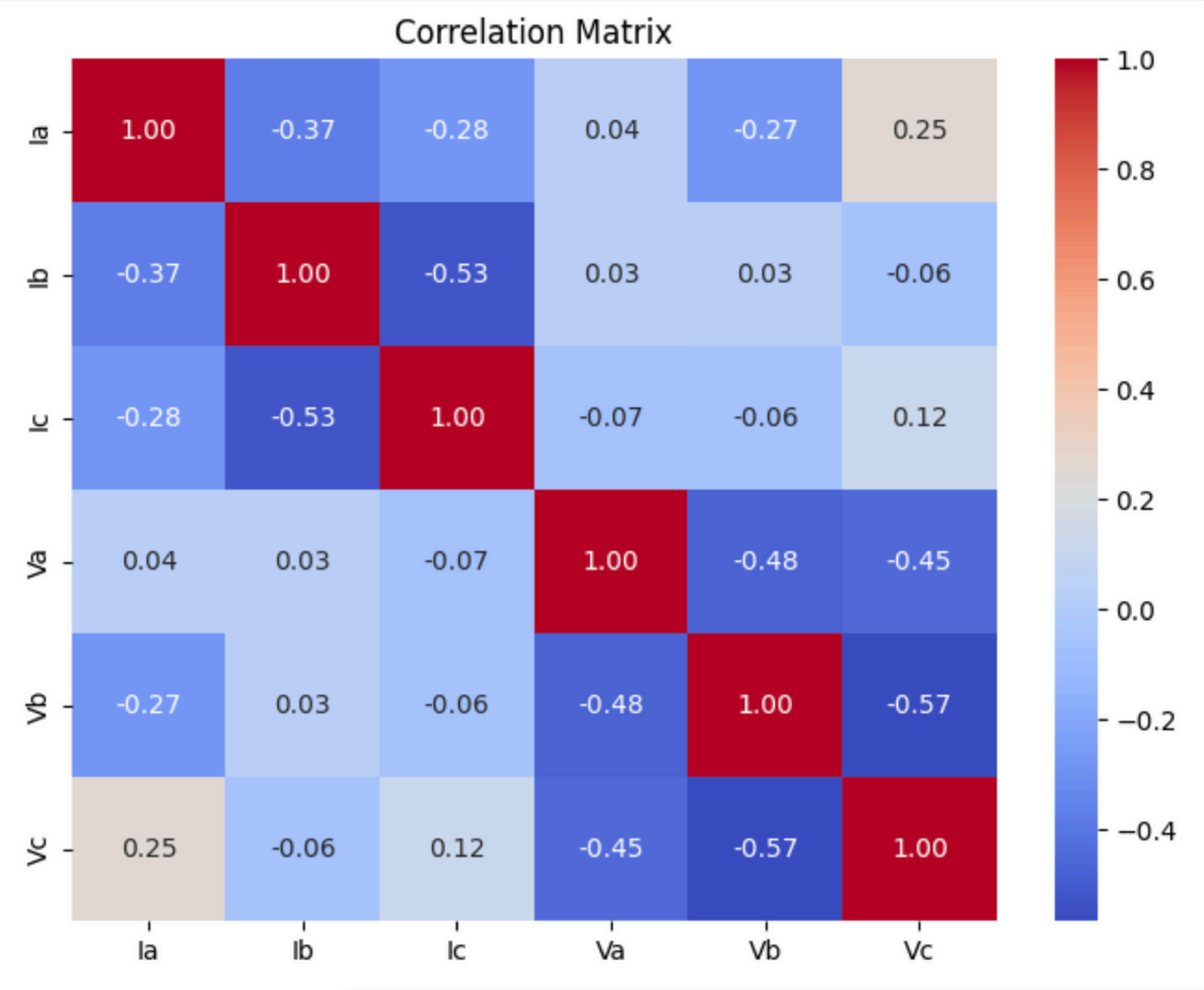


Voltage Distribution

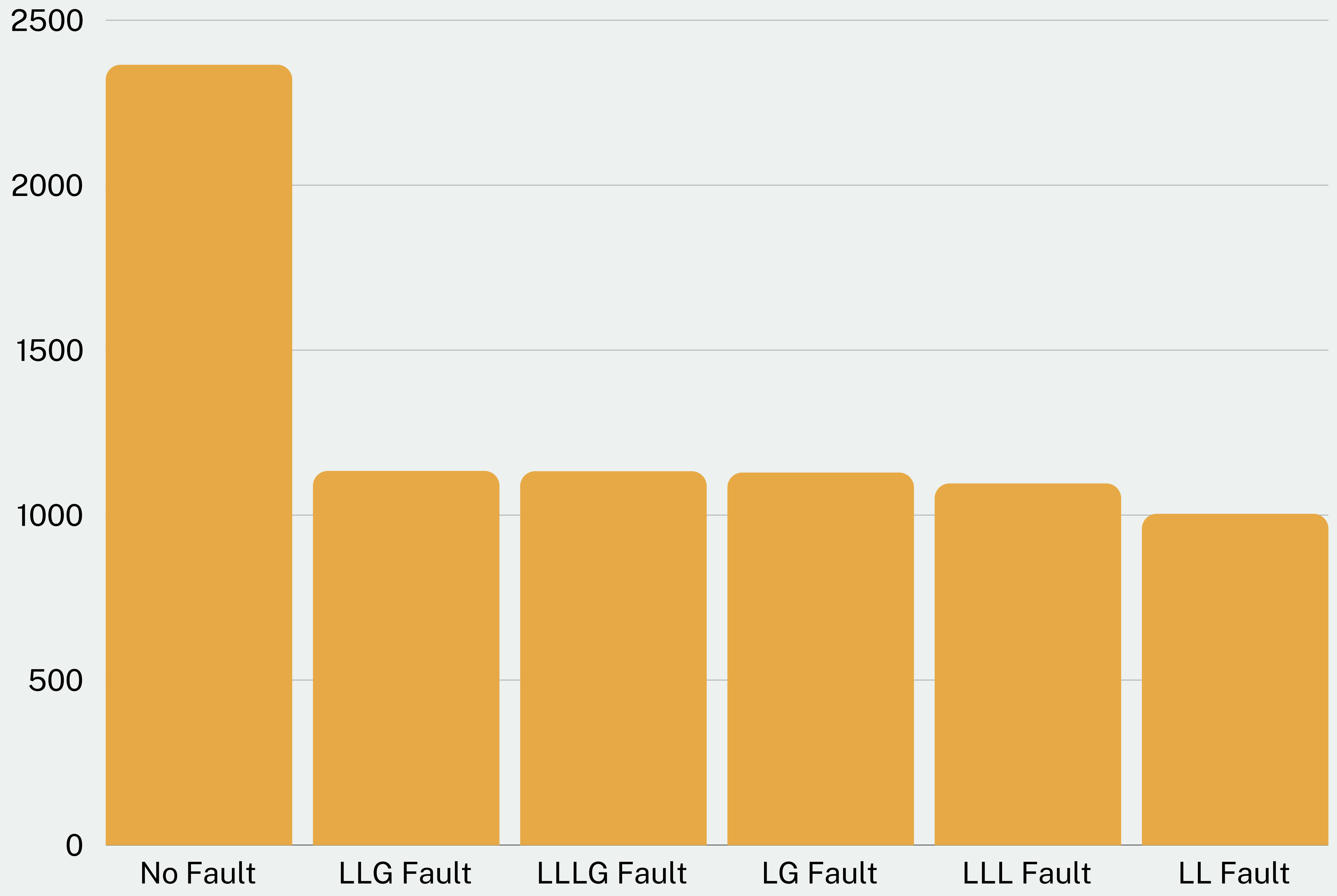


Correlation Matrix

- Ia & Ib
- Ib & Ic
- Va, Vb, & Vc



Model Training



18.85%
PCC

23.56%
1.25*PCC

Model Training

80:20
TrainVal : Holdout

MinMaxScaler
Scaling

SMOTE
Sampling*

5 folds
cross-validation

f1-scorer
scorer

**Only the training data is resampled*

Model	F1 Score (TrainVal)	Runtime
KNN Classifier	0.802	0.43 seconds
LogisticRegression	0.152	0.83 seconds
DecisionTree	**0.840**	0.41 seconds
RandomForest	**0.838**	7.15 seconds
GradientBoosting	0.815	60.14 seconds
XGBClassifier	0.830	356.72 seconds
SVM	0.770	6.07 seconds
LightGBM	0.800	2.77 seconds
CatBoost	0.791	278.71 seconds

Confusion Matrix for Decision Tree:

True Values		0	1	2	3	4	5
No fault	0	500		1			
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2		3	178		27	
LG Fault, between Phase A and Ground	3			1	235		
LLL Fault, involving all three phases	4		1	154		61	
LL Fault, between Phases B and C	5			1			202

Predicted Values

Confusion Matrix for RandomForest:

True Values		0	1	2	3	4	5
No fault	0	500		1			
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2			128		80	
LG Fault, between Phase A and Ground	3				235	1	
LLL Fault, involving all three phases	4			86		130	
LL Fault, between Phases B and C	5						203

Predicted Values

Decision Tree:

		0	1	2	3	4	5
No fault	0	500		1			
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2		3	178		27	
LG Fault, between Phase A and Ground	3			1	235		
LLL Fault, involving all three phases	4		1	154		61	
LL Fault, between Phases B and C	5			1			202

Lowest precision and recall in

- 'Three-Phase' and
- 'Three-Phase with Ground' categories.

RandomForest:

		0	1	2	3	4	5
No fault	0	500		1			
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2			128		80	
LG Fault, between Phase A and Ground	3				235	1	
LLL Fault, involving all three phases	4			86		130	
LL Fault, between Phases B and C	5						203

Perfect precision and recall in

- 'LLG Fault, involving Phases A, B, and Ground'

Lowest precision and recall in

- 'Three-Phase' and
- 'Three-Phase with Ground' categories.

From **6 features** to **40 features** using **Polynomial features** and **domain knowledge with hypertuning:**

['Ia', 'Ib', 'Ic', 'Va', 'Vb', 'Vc',] → ['Ia', 'Ib', 'Ic', 'Va', 'Vb', 'Vc', 'Ia^2', 'Ia Ib', 'Ia Ic', 'Ia Va', 'Ia Vb', 'Ia Vc', 'Ib^2', 'Ib Ic', 'Ib Va', 'Ib Vb', 'IbVc', 'Ic^2', 'Ic Va', 'Ic Vb', 'Ic Vc', 'Va^2', 'Va Vb', 'Va Vc', 'Vb^2', 'Vb Vc', 'Vc^2']

New Features	Formula	Description
Zero Sequence Currents, Zero Sequence Voltages	Io = (Ia + Ib + Ic) / 3 Vo = (Va + Vb + Vc) / 3	The unbalanced current flows in the circuit during the ground fault.
Voltage and Current Ratios	VI_a = Va / Ia VI_b = Vb / Ib VI_c = Vc / Ic	Ratio of voltage and current in a line, can be referred to as ‘impedance’.

Model	F1 Score (Holdout)	Best Parameters	Prediction Time
DecisionTree	0.998	{'classifier__criterion': 'entropy', 'classifier__max_depth': None, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 10}	0.00 seconds
RandomForest	0.999	{'classifier__bootstrap': False, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100}	0.04 seconds
XGBClassifier	0.999	{'n_estimators': 100, 'max_delta_step': None, 'max_depth': None, 'max_leaves': None, 'min_child_weight': 1, 'random_state': 143, 'verbosity': None}	0.01 seconds
LightGBM	0.999	{'boosting_type': 'gbdt', 'colsample_bytree': 1.0, 'importance_type': 'split', 'learning_rate': 0.1, 'min_child_samples': 20, 'min_child_weight': 0.001, 'n_estimators': 100, 'num_leaves': 31, 'random_state': 42, 'subsample': 1.0, 'subsample_for_bin': 200000}	0.03 seconds
CatBoost	1.000	{'classifier__depth': 1, 'classifier__l2_leaf_reg': 1, 'classifier__learning_rate': 0.1, ‘verbose’: 0}	0.03 seconds

Confusion Matrix for RandomForest:

True Values

No fault	0	501					
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2			207		1	
LG Fault, between Phase A and Ground	3				236		
LLL Fault, involving all three phases	4					216	
LL Fault, between Phases B and C	5						203
		0	1	2	3	4	5

Predicted Values

Confusion Matrix for XGBoost:

True Values								
	No fault	0	501					
	LLG Fault, involving Phases A, B, and Ground'	1		209				
	LLLG Fault, a three-phase symmetrical fault'	2			207	1		
	LG Fault, between Phase A and Ground	3				236		
	LLL Fault, involving all three phases	4					216	
	LL Fault, between Phases B and C	5						203
			0	1	2	3	4	5

Predicted Values

Confusion Matrix for LightGBM:

True Values	No fault	**0**	**501**				
	LLG Fault, involving Phases A, B, and Ground'	**1**		**209**			
	LLLG Fault, a three-phase symmetrical fault'	**2**			**208**		
	LG Fault, between Phase A and Ground	**3**				**235**	**1**	
	LLL Fault, involving all three phases	**4**					**216**	
	LL Fault, between Phases B and C	**5**						**203**
			0	**1**	**2**	**3**	**4**	**5**

Predicted Values

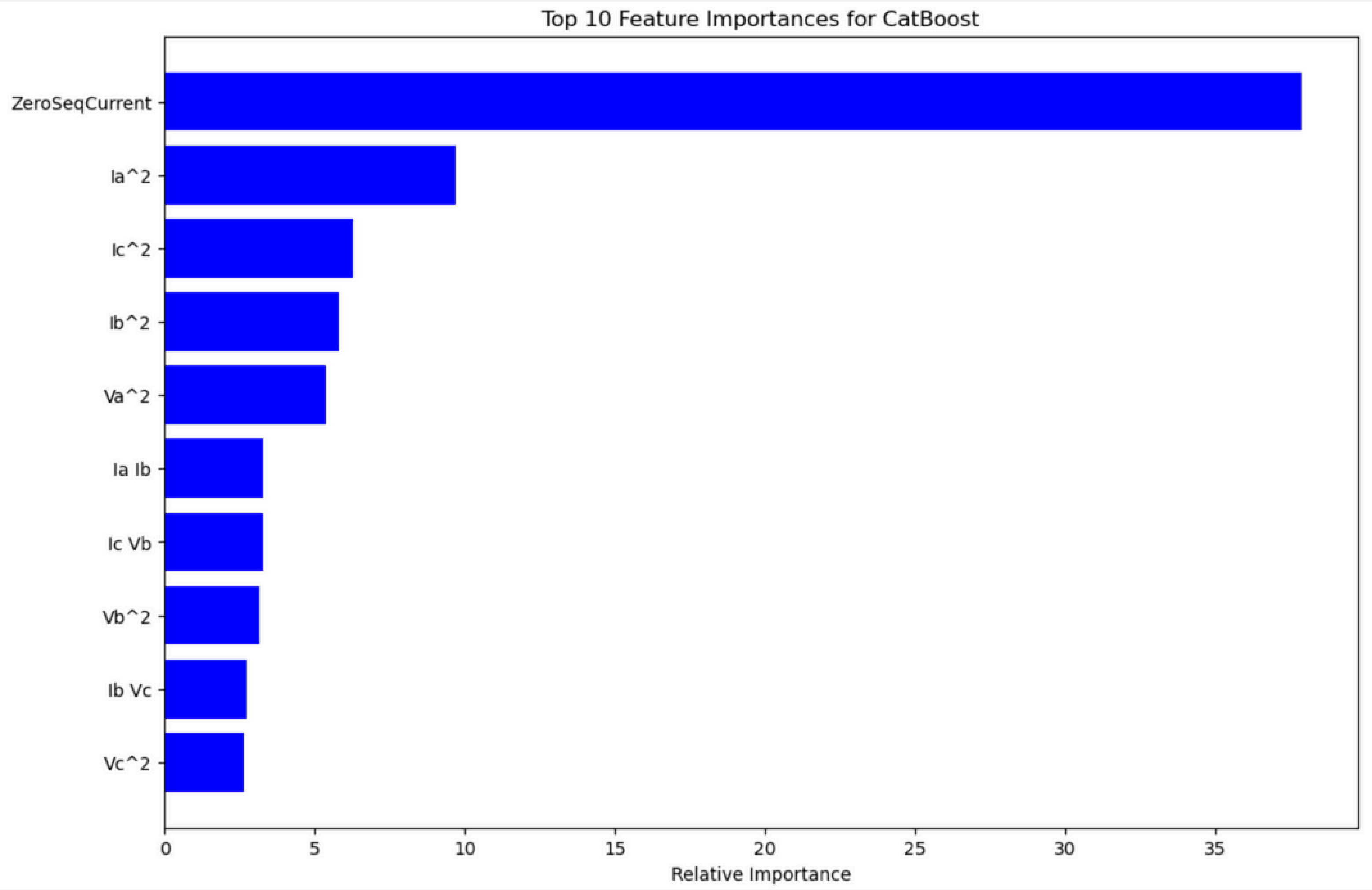
Confusion Matrix for CatBoost:

True Values

No fault	0	501					
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2			208			
LG Fault, between Phase A and Ground	3				236		
LLL Fault, involving all three phases	4					216	
LL Fault, between Phases B and C	5						203
		0	1	2	3	4	5

Predicted Values

Simpler Model using Top 10 Features of CatBoost:



Confusion Matrix:

		0	1	2	3	4	5
No fault	0	501					
LLG Fault, involving Phases A, B, and Ground'	1		209				
LLLG Fault, a three-phase symmetrical fault'	2			208			
LG Fault, between Phase A and Ground	3				236		
LLL Fault, involving all three phases	4					216	
LL Fault, between Phases B and C	5						203

Conclusion:

- CatBoost proven to be the best model for this problem, achieving a 100% f1-score.
- The top 10 features hold most of the predictive power for an accurate classification in the dataset with 100% F1 score.
- Using fewer features simplifies the model, reduces the risk of overfitting, and can improve computational efficiency.
- Feature Engineering is the most important step as it was highly effective in getting the best derived feature.

Recommendations:

- Expanding the dataset to include a wider variety of fault classes
- Introduce more electrical characteristics such as phase angles, phase differences, and frequency and economic factors such as repair costs, and maintenance costs.
- Deploying the model in a simulated real-time environment to evaluate its performance under operational conditions
- Cost-implications on prioritizing f0.5 over f1