Electrical Fault detection With SVM and Logistic regression

Arjun kandel [B200387PE] Vishakha [B210888EE] Unnati [B210776EE]
National Institute of Technology, Calicut

Abstract - Transmission line electrical faults present serious obstacles to the dependable and effective functioning of power systems, frequently resulting in equipment damage, service interruptions, and safety risks. It is imperative to promptly and precisely identify these errors in order to avert extensive disruptions and minimize possible hazards [1]. Machine learning, or ML, techniques like logistic regression (LR) and support vector machines (SVM) provide practical options for defect diagnosis in order to address this problem. Phase angle, voltage, and current readings from sensors along transmission lines are used by these algorithms to spot patterns and project events to come.

Logistic regression, which excels in binary classification, determines if a fault exists or not. SVM, on the other hand, can handle both linear and non-linear classifications more skillfully by accommodating complex fault patterns. Real-time monitoring is made possible by the integration of these ML models into fault detection systems, which facilitates quick fault localization and prompt intervention.

Index terms- Support vector Machine (SVM), Logistic regression, power systems faults, Modeling

1 Introduction

The transmission line is the electricity system's most crucial element. Electric power must be sent from the source region to the distribution network via a transmission line in order to fulfill the contemporary era's enormous need for electricity and its allegiance. The many intricate, dynamic, and interacting components that make up the electrical power system are constantly vulnerable to interference or electrical malfunctions. Normal operating conditions for a power system are regulated and balanced. When the system becomes unbalanced because of insulation failures at any point or because live wires come into touch with one another, there is a short-circuit or fault in the line. Numerous factors, including as natural disturbances (such as lightning, strong winds, earthquakes), deteriorating insulation, falling trees, bird shorting, and more, may result in power system faults. There are two main categories of faults:

1. Open-circuit fault
2. Short-circuit faults

2 Support vector machine (SVM)

Supervised learning algorithms, or SVMs for short, are useful for problems involving regression or classification. Support Vector Machines also known as SVM works on principle in which it identifies a hyperplane that efficiently divides different classes in the training dataset. Finding the hyperplane that maximizes the margin which is the distance between the hyperplane and the nearest data points from each class is how this is accomplished. New data may be classified according to which side of the hyperplane it stays on after it has been created earlier by using training data set. SVMs are especially useful for datasets with a large number of characteristics or when class boundaries are well defined. [3].

In the SVM, the hyperplane is represented as follows.:

where w is the weight vector perpendicular to the hyperplane, b is the bias, and xi is the input data.

The magnitude of the margin is calculated using the following equation:

$$margin = \frac{2}{||w||} + C\sum_{i=1}^{n}\zeta_i \qquad (2)$$

Where ζ_i represents a slack variable incorporated to accommodate potential misclassifications when linear classification of data isn't feasible. Meanwhile, 'C' stands for a user-specified parameter regulating the extent of permissible misclassification. A higher 'C' value corresponds to a stricter tolerance for misclassification.

In SVMs, the objective is to identify a hyperplane that maximizes the margin size. This pursuit involves determining the minimum value for 'w', a process that constitutes an optimization problem. The optimization involves a set of constraints (Equation 1) and an objective function (Equation 2) [2].

For optimization, the Lagrange multiplier approach was used. An approach for determining a function's maximum or minimum under restrictions is the Lagrange multiplier method. α or λ is a new variable that is used to incorporate the restrictions into the objective function. By multiplying the restrictions by this new variable and adding them to the initial goal function, an equation is created using this procedure. [9]. Following is the expression for the final function of the SVM that was produced by using the Lagrange multiplier method:

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b \qquad (3)$$

where N is the number of samples in the training data, y_i is the output data, a_i is the Lagrange multiplier, and K() is the kernel.

The kernel serves the purpose of projecting data into higher dimensions particularly in non-linear classifications where linear classification isn't viable. In this research, we opted for a radial basis function (RBF) kernel [10]. The formulation of the RBF kernel is articulated as follows:

$$K(x_i, x) = \exp\left(\frac{-||x_i - x||^2}{2\gamma^2}\right) \qquad (4)$$

where the decision boundary's degree of flexibility is controlled by the user-defined parameter γ.

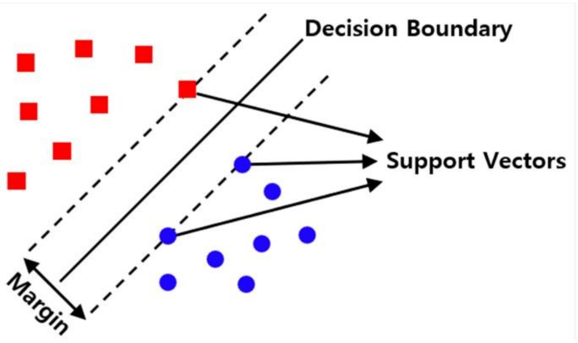


Fig. 1: Support vector machine architecture

3 Logistic regression

In order to estimate the probability that an input belongs to a certain class, supervised machine learning techniques like logistic regression are often used for classification issues.

We fit a "S" shaped logistic function, which predicts one of two maximum values (0 or 1), as an alternative to creating a regression line. [4]. The logistic function, also referred to as the sigmoid function, is a mathematical formula that changes predicted values into probabilities by normalizing the actual values between 1 and 0. This function is pivotal in logistic regression, ensuring that predictions fall within this boundary, creating a characteristic S shaped curve. In logistic regression, the notion of a threshold value is employed to delineate the probability of either 0 or 1. Values surpassing this threshold are inclined toward 1, while those below it trend to favor 0. [4].

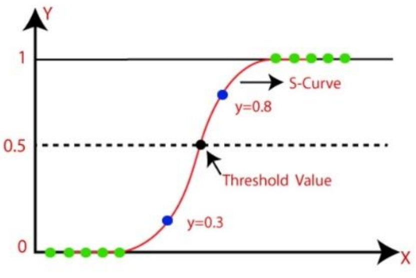


Fig. 2: Sigmoid function (S curve)

$$Log\frac{y}{1-y} = b_0 + b_1x_1 + b_2x_2 + \cdots b_nx_n$$

The above equation is the equation for logistic regression.

4 Modeling of Problem using python

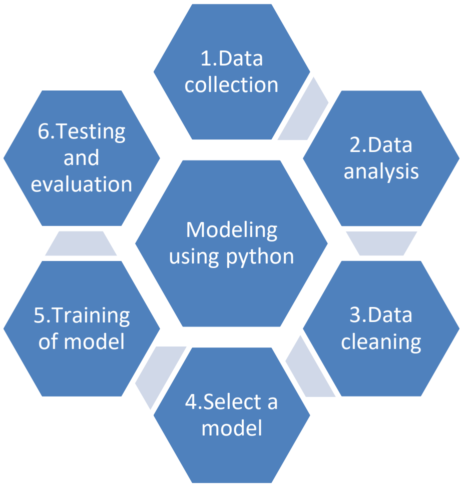


Fig. 3: work flow

4.1 About dataset used

The Power system MATLAB model was developed for simulating fault analysis within a power system. However, the dataset comprising over 7,861 labeled data points, encompassing measurements of Line Voltages and Line Currents, was acquired from Kaggle, a trusted external source [6,7].

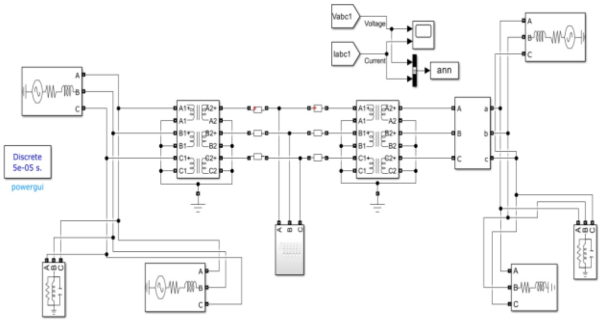


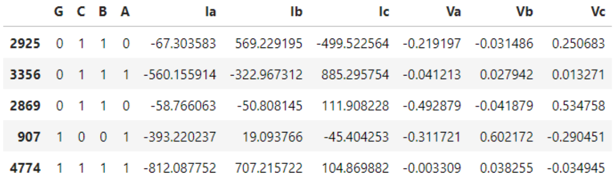
Fig. 4: Power system circuit

Inputs - [Ia, Ib, Ic, Va, Vb, Vc]

Ia = Current of line A
Ib = Current of line B
Ic = Current of line C
Va = Voltage of line A
Vb = Voltage of line B
Vc = Voltage of line C

Outputs - [G C B A]

Each sequence in output of binary digits corresponds to specific fault types occurring within the system involving the four components. When all four digits are 0 (i.e., [0 0 0 0]), it indicates a normal operating condition with no faults detected. Conversely, the sequence [1 1 1 1] represents the most severe scenario - a three-phase symmetrical fault (LLL fault) affecting all three phases and the ground. Intermediate fault scenarios are denoted by different combinations. For instance, [1 0 1 1] indicates an LLG fault occurring between Phases A, B, and the ground. Another example is [0 1 1 1], signifying an LLL fault, indicating a fault occurring simultaneously between all three phases without involving the ground.



	G	C	B	A	Ia	Ib	Ic	Va	Vb	Vc
2925	0	1	1	0	-67.303583	569.229195	-499.522564	-0.219197	-0.031486	0.250683
3356	0	1	1	1	-560.155914	-322.967312	885.295754	-0.041213	0.027942	0.013271
2869	0	1	1	0	-58.766063	-50.808145	111.908228	-0.492879	-0.041879	0.534758
907	1	0	0	1	-393.220237	19.093766	-45.404253	-0.311721	0.602172	-0.290451
4774	1	1	1	1	-812.087752	707.215722	104.869882	-0.003309	0.038255	-0.034945

Fig. 5: Sample of Dataset

4.2 Exploratory Data Analysis of dataset

- By the categorical faults analysis, it was found according to their fault type

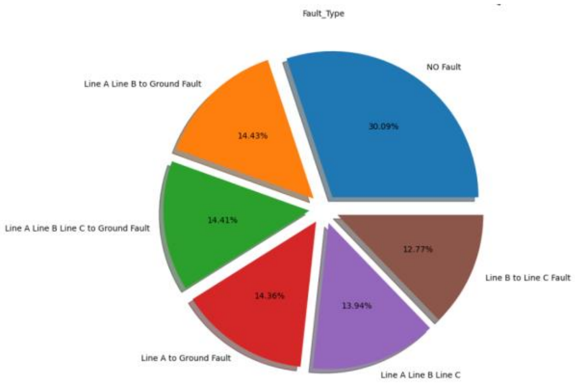


Fig. 5: Category wise faults

- In voltage or current graph, where there is large fluctuation in the graph, there faults have occurred.

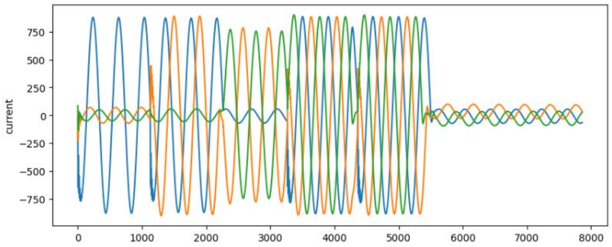


Fig. 6: Current fluctuation graph

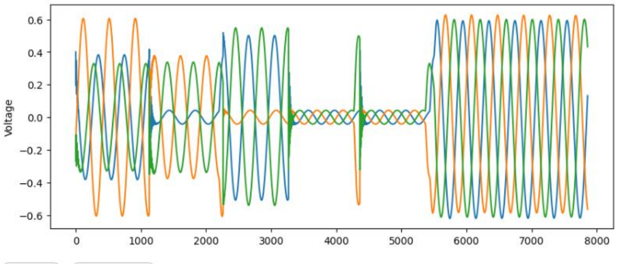


Fig. 7: Voltage fluctuation graph

- The predicted symmetrical and sinusoidal behavior with a 120-degree phase shift in both current and voltage is seen in the current and voltage graphs under the "not Faulty" condition. Curves don't exhibit any distortion. On the other hand, noticeable aberrations may be seen on the voltage and current curves when the system is experiencing a "Faulty" condition.

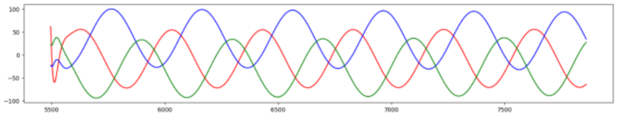


Fig. 8a: current variation in no fault condition

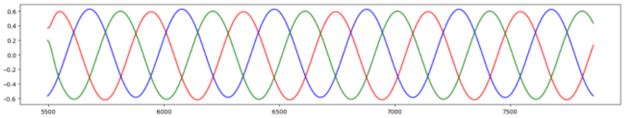


Fig. 8b: Voltage variation in no fault condition

4.3 Model Selections

In this project we have used logistic regression and SVM to train machine learning model. Both models were chosen for their proven efficacy in binary classification tasks like fault detection. Preprocessing steps involved handling missing values and standardizing features. During training, varied hyperparameters were explored for both models.

4.4 Model Training

The whole data set must be split into two sets train and test in order to train the model using the data.

a) Logistic regression: Logistic regression model is initialized using **LogisticRegression from Scikit-Learn package** [5].
set the Logistic Regression model to the training data by using **.fit()** method.

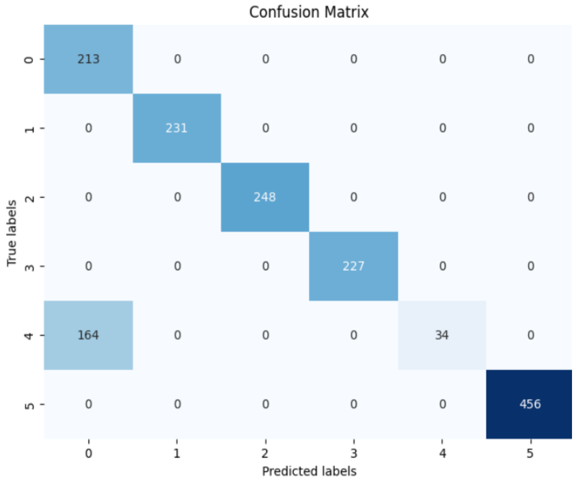


Fig.9: Confusion matrix for logistic regression

b) Support vector machine (SVM): SVM is initialized using **SVC from Scikit-Learn package** [5]**.**
Fit the SVM model to the training data using **.fit()** method.

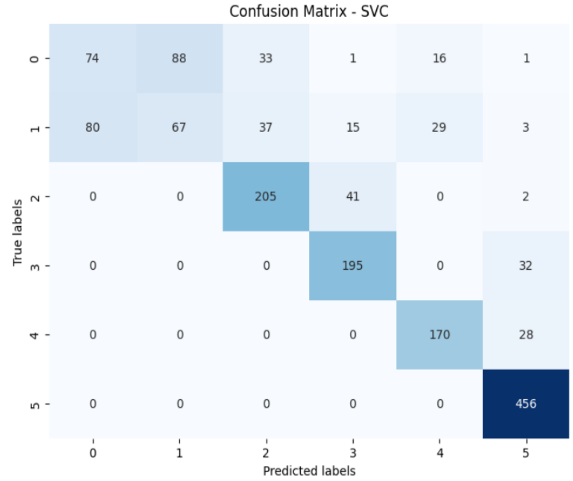


Fig.10: Confusion matrix for SVM

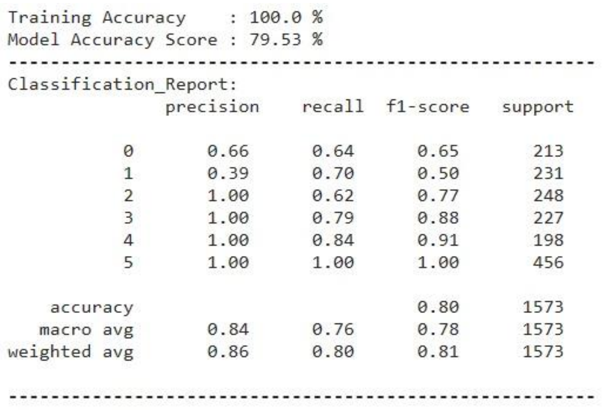
5 Results

The results for both models are summarized in Table 1

Model	**Training accuracy (%)**	**Model accuracy score (%)**
Logistic regression	90.19	89.57
SVM	100	79.53

6 Conclusion

To look into failure situations within a power system, logistic regression and SVM models were developed on given the data set. Both models showed to be able to perform fault prediction. The models were assessed using a range of performance indicators, including F1-score, recall, accuracy, and precision.



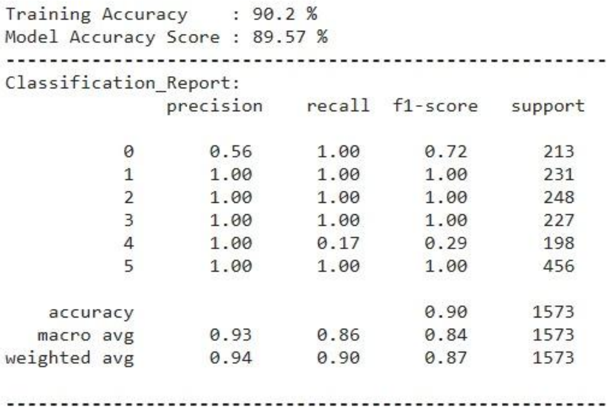
```
Training Accuracy     : 100.0 %
Model Accuracy Score : 79.53 %
--------------------------------------------------------
Classification_Report:
               precision    recall  f1-score   support

           0       0.66      0.64      0.65       213
           1       0.39      0.70      0.50       231
           2       1.00      0.62      0.77       248
           3       1.00      0.79      0.88       227
           4       1.00      0.84      0.91       198
           5       1.00      1.00      1.00       456

    accuracy                           0.80      1573
   macro avg       0.84      0.76      0.78      1573
weighted avg       0.86      0.80      0.81      1573

--------------------------------------------------------
```

Fig.11: F1-score, recall, accuracy, and precision for SVM



```
Training Accuracy     : 90.2 %
Model Accuracy Score : 89.57 %
--------------------------------------------------------
Classification_Report:
               precision    recall  f1-score   support

           0       0.56      1.00      0.72       213
           1       1.00      1.00      1.00       231
           2       1.00      1.00      1.00       248
           3       1.00      1.00      1.00       227
           4       1.00      0.17      0.29       198
           5       1.00      1.00      1.00       456

    accuracy                           0.90      1573
   macro avg       0.93      0.86      0.84      1573
weighted avg       0.94      0.90      0.87      1573

--------------------------------------------------------
```

Fig.12: F1-score, recall, accuracy, and precision for LR

[These results are also attached with code].

The comparative analysis between Logistic Regression and SVM models revealed that logistic regression performing well with big data with 90.19% training accuracy and 89.57% for model accuracy score on the other hand SVM with kernel RBF with **c = 10 and sigma = 0.1** showing training accuracy of 100% but model accuracy of 79.53 %. This study highlights the importance of machine learning models in analyzing faults in power systems. It provides valuable information on predicting faults, ensuring system dependability, and implementing actions to mitigate faults. Future work could explore ensemble methods or neural network architectures to further improve fault prediction accuracy.

References

1. Vishal, M., Kamal, A. and Viji, D., 2023, June. Electrical fault detection using machine learning algorithm. In AIP Conference Proceedings (Vol. 2782, No. 1). AIP Publishing.

2. Kim, M.C., Lee, J.H., Wang, D.H. and Lee, I.S., 2023. Induction Motor Fault Diagnosis Using Support Vector Machine, Neural Networks, and Boosting Methods. Sensors, 23(5), p.2585.

3. https://www.javatpoint.com/logistic-regression-in-machine-learning

4. https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/

5. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

6. Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller and Sarah Guido

7. Electric Power Systems: A Conceptual Introduction" by Alexandra von Meier

8. Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy.

9. Simon Hayekins, Neural Networks: A comprehensive foundation, published by Pearson education (Singapore) Pte. Ltd