Predictive Model for Classification of Power System Faults using Machine Learning

Tilottama Goswami Uponika Barman Roy Computer Science and Engineering Electrical and Electronics Engineering Anurag Group of Institutions BVRIT Hyderabad Hyderabad, India Hyderabad, India [tilottama.goswami@rediffmail.com](mailto:tilottama.goswami@rediffmail.com) [uponika@gmail.com](mailto:uponika@gmail.com)

*Abstract*—Power System is a combination of electric power generation, transmission, distribution and utilization systems. In brief, power system is the heart of any electrical system. In an electric power system, a fault or fault current is any abnormal electric current. As a consequence of such fault, the entire system may damage and eventually collapse. The aim of this work is to automatically classify the faults into one of the eleven faulty classes, which includes both balanced and unbalanced faults. The dataset of generated fault in overhead transmission lines is synthetic, which consists of 11 different faults for 100 kilometers. The simulation is done using MATLAB/Simulink software model. The task of classification of faults is implemented using supervised machine learning algorithms in Python and scikit-learn. Comparison is made using three commonly used classification algorithms - Decision Tree, K-Nearest Neighbor, Support Vector Machine (SVM). SVM performed excellent giving a performance with 91.6% test accuracy for the generated dataset. The predictive model will thus make the system more intelligent in bringing up reliable power supply.

Keywords—fault classification, power system, supervised machine learning

# Introduction

Nowadays the growth in the demand for electrical power is seamless all over the world. Transmission lines in a power system have their significant role in satisfying this ever increasing demand. Moreover transmission system has its own limitations in supplying power to its thermal and stability thresholds. Hence this condition leads to an unstable system [1]. Under normal conditions power system is balanced and it becomes unbalanced on occurrence of faults. A fault can be termed as any certain abnormality in voltage and current in normal power system that represents disturbance in a system. Fault occurs at any time and at any location with different magnitude of fault impedance. The point at which a fault occurs behaves as a sink point and as a result the voltage tends to become zero. Thus all the points those have higher potential than the faulty point start sending current to these faulty points and raise the fault level higher magnitude very higher than that at the normal level [2]. A fault can be symmetrical or unsymmetrical. When a fault involves all three phases at a time then is a symmetrical fault whereas a fault that involves one or more phases is an unsymmetrical fault. Any kind of fault it is, the effects to our power system are often damaging. This frequent problem faced by power system can be treated and has to be analyzed more preciously to maintain stable conditions and to lead the seamless operation in the system. Therefore, during fault conditions it is important to determine the faulty values of system voltage and current so that the unhealthy system can be identified and rectified. This process of evaluating power system voltages and currents under various types of conditions is called fault analysis [2]. In reality, power system consists of thousands of buses which complicate the task of calculating the necessary parameters without the use of computer software. Conventional algorithms based on deterministic computations and well-defined model of power lines, are resulting in the late detection and inaccurate results. Traditional distance relays consider power swing as a fault. Such malfunctioning could lead to serious disturbance of power system stability. Fast fault detection can help protect equipment by allowing the disconnection of the power transmission line before any damage occurs. Hence increasing cost savings and power transmission system efficiency and reliability [4].

Hence there is great need of accuracy and reliability in fault detection and location in order to maintain proper functioning of the system. Recent research began to introduce machine learning techniques to identify fault type, location detection and prediction. Techniques such as multilayer perceptron, random forests, support vector machines were used to locate and predict faults on transmission lines [3, 4]. The present scope of the paper focusses only on fault analysis and classification in the simulated dataset. The identification of faulty conditions from the good ones is not in the present scope. The paper is organized as follows. Section 2 mentions briefly the various types of faults and their causes. Section 3 describes the simulation of fault dataset generation. Section 4 briefly discusses the supervised machine learning algorithms used in this paper for classification. Section 5 is for result analysis. The paper finally concludes with conclusion and future work section.

# Types of Faults in power system

## Faults

Electrical power system has two types of faults which can occur on any transmission lines. They are balanced faults and unbalanced faults. The unbalanced faults can be classified further into single line-to-ground faults, double line faults and double line-to-ground faults.

* Line-to-ground fault: this type of fault occurs when one of the three transmission lines shorted with the ground. 70% of all transmission line faults are classified under this category.
* Line-to-line fault: when a phase comes in contact with the other due to heavy wind, it results the line-to-line fault. This type of fault is occurred as 15% of the all transmission lines faults.
* Double-line-to-ground: when two phases comes in contact with the ground potential it causes the double-line-to-ground fault. 10% of all transmission line faults fall under this category.
* Three phase faults: in this case, the entire system gets disrupted due to fatal failure of the towers, rupture of the transmission line and heavy wind flow. In reality, this type of fault not often exists which can be seen from its share of only 5% of all the transmission lines faults.

When a fault occurs, the use of powerful machine learning techniques in fault prediction can enhance the protection of power transmission system. In addition it diminishes the time required to clear the faults, especially for a long transmission line, hence increasing the overall power system reliability and efficiency [3].

## Causes of Faults

Electrical systems, machines and components often encounter different kinds of faults during its operation. During fault occurrence, the impedance values of the machines may vary from the original values to other values till the fault clearance. The fault occurrence involves the number of causes like falling tower, loss of insulation of the conductors, conducting path failure and natural disaster. All these results in short circuit of the lines and tremendous damage to the entire system.

At safe running conditions, the electric components in the power network work at rated voltage and default current. As the fault occurs in a circuit, voltage and current values get deflected from their nominal ranges.

* Open circuit faults are resulted due to disrupted conductor and error in circuit breaker the transmission phases.
* Short circuit faults are resulted due to system internal or system external effects such as
* Internal effects - include break of transmission lines or power system components, loss of insulation in generator, transformer and other electrical equipments, improper installations and inadequate design.
* External effects include overloading of power system, insulation failure due to surges and mechanical damage by public [1]

# Simulation of Fault Dataset Generation

The simulation modeling and the control of real plant are two-very often separated-fields of mechatronic practice, research and education. The deep gap between simulation world of controlled system and practical implementation in e.g. embedded hardware can be bridged through the use of special hardware and software directly in MATLAB/Simulink environment [5]. Simscape Electrical (formerly SimPowerSystems and SimElectronics) provides component libraries for modeling and simulating electronic, mechatronic, and electrical power systems. Simscape Electrical helps in developing control systems and in testing system-level performance.

The Simulink model is designed with all the blocks those constitute to a power system transmission line. Therefore the system can be properly analyzed and can be identified from various fault conditions. The different blocks that are used can be shown in a concise manner as in Fig 1. For better clarity, few blocks are shown separately in subsequent figures.

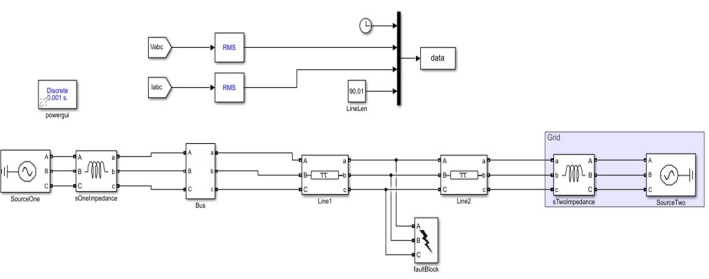


Fig 1: Fault Generation Model using Simulink

## Blocks of the Model

The Simulink model is designed with all the blocks those constitute to a power system transmission line. Therefore the system can be properly analyzed and can be identified from various fault conditions. The different blocks that are used can be shown in a concise manner as in Fig 1. For better clarity, few blocks are shown separately in subsequent figures.

Source Block: The Source block is a 3-phase balanced voltage source with internal impedance. The voltage sources are in Y connection with a neutral that can be grounded. The internal resistance and inductance of the block can be specified either directly by selecting R and L values or indirectly by adjusting the source inductive short-circuit level and X/R ratio as shown in Fig 2.

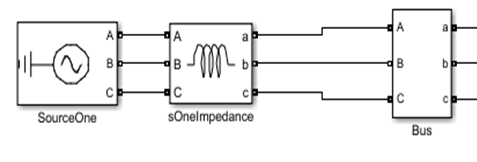


Fig 2: Source Block

Impedance Block: This block determines the impedance between two nodes of a linear circuit related to the frequency. It comprises of a current source Iz, connected between inputs one and two of the block, and a voltage measurement Vz, connected parallel to the current source.

Transmission Lines: This block models a 3-phase transmission line using the lumped-parameter pi-line model. This model involves the phase resistance, line-line mutual inductance, phase self-inductance and resistance, line-line capacitance and line-ground capacitance.

Bus Bar Block: A power system block having a schematic which indicates a significant location to measure and not to represent a physical component in the system. More than one physical signal lines are connected together directly without the use of a Bus Bar.

Fault Block: This block consists of 3-phase circuit breaker where the starting and ending times can be controlled either from a Simulink signal (external mode), or from an control timer (internal mode) as shown in Fig 3.

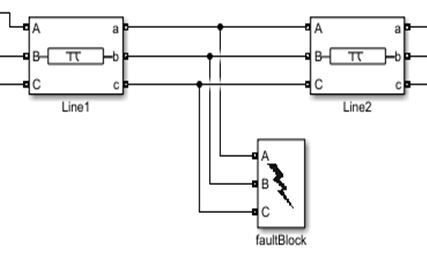


Fig 3: Fault Block

The 3-Phase Fault block uses Breaker blocks. It can be separately switched on and off to program phase-to-phase faults, phase-to-ground faults, or a combination of phase-to-phase and ground faults.

Data generation from fault block while running the Simulink: This block notes down the voltages and currents during the Simulink running and under different fault conditions. These values further greatly help us in analysis of the Fault occurred.

The Three-Phase voltages and currents thereby obtained are converted into RMS values for better accuracy. The timer sets the time span considered for the fault data capturing 5/50sec to 15/50 sec. The line length is set to 90.01 km as a required threshold of distance for fault consideration in our paper as shown in block diagram in Fig 4. The data block records the data generated in the form of columns and saves it in the MATLAB workspace.

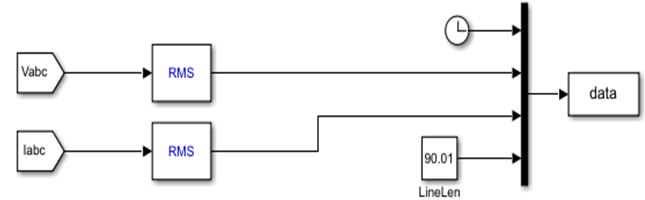


Fig 4: Data Block

## MATLAB files to make Synthetic faults – Fault generation

The power transmission line total length (90.01 Km) was divided into ten equal sectors, each sector length is 10.01Km of the line total length. The reason behind dividing the line to ten equal sectors is to attempt to increase the accuracy of fault location prediction by reducing the line length. For fault type detection, 11 faults are specified to be classified and predicted as listed below:

For fault type detection, 11 faults were specified to be classified and predicted as listed below:

* line A to Ground (AG)
* Line B to Ground (BG)
* Line C to Ground (CG)
* Line A and B (AB)
* Line B and C (BC)
* Line A and C (AC)
* Line A, B and Ground (ABG)
* Line B, C and Ground (BCG)
* Line A, C and Ground (ACG)

• Line A, B, C (ABC)

• Line A, B, C and Ground (ABCG).

The fault’s 3 phases’ line voltages Va, Vb, Vc and the short circuit currents Ia, Ib, Ic have been generated and normalized by simulating each fault condition. The data from the workspace is exported and written in Microsoft Excel sheets. This raw data in the workspace needs to be tabulated further in order to feed it to the machine learning classifier environment in a proper form for better analysis and greater accuracy. The snapshot of the data generated is shown in Fig 5. The data can be simulated in parallel at different distances. The paper reports result analysis for using the machine learning algorithm for classification with output of only fault types. For each of the 11 fault types, at least 1000 cases have been generated. Total 11300 instances are generated.

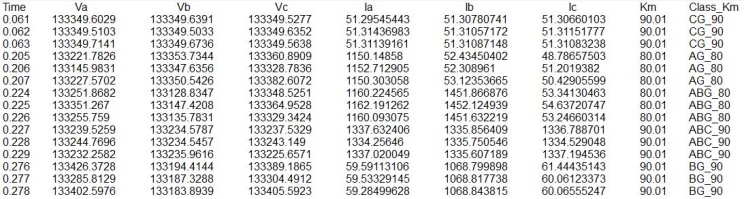


Fig 5: Fault Data Capture

# Machine Learning Task – Fault Classification

Machine Learning suggests that “machine learns”, which indicates that the technology makes a machine to learn or makes it intelligent. Thus it can be defined as the technology of making a system intelligent enough to take its own decision. Technically, it is an application where the system can take decisions, make predictions and improves its capability from the learning of its past experience “without being explicitly programmed”. Tom M. Mitchell [6] provided a formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". There are two types of learning – supervised and unsupervised.

* Supervised learning is where the data has input variables (x) and an output variable (Y) and an algorithm is used to learn the mapping function from the input to the output. The tasks that can be done on the supervised data are classification, regression. Each of the tasks has its own set of algorithms to perform the task.
* Unsupervised learning has only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. The tasks implemented on unsupervised data are clustering and association.

The methodology chosen is supervised classification algorithms for classifying the fault types for the dataset generated in the previous section.

## Dataset Preprocessing

#### The data considered for further processing consists of a part of the previously collected dataset.

#### Feature Selection : It consists of six input features, 3 set of RMS voltage values in volts (Va,Vb,Vc) and 3 set of RMS current values in amps (Ia, Ib,Ic) generated by simulation when a type of fault occurs at a particular location. The output feature is the fault type (Type) depicting the 11 fault types as mentioned in Section 3B. The csv is prepared accordingly and provided as input to the algorithm as shown in Fig 6.

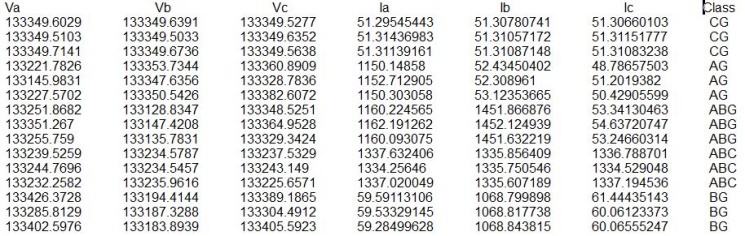


Fig 6: Snapshot of the dataset

* *Class Label Encoding:* In supervised machine learning for classification task, the output classes are sometimes labelled in the form of words, known as categorical data. For example, the output class ‘Type’ in the present work depicts faults - AG,BG,CG,AB,BC,AC,ABG,BCG,ACG,ABC,ABCG. These labels are human understandable in a particular domain. Machine Learning can perform in a better way when they are in the form of numbers. The process of converting the labels into numeric form is called label encoding and is an important pre-processing step for the structured dataset in supervised learning. The implementation is done using Python 3.0 and scikit-learn which is machine learning library[7]. The library has package sklearn.preprocessing which consists of utility function such as LabelEncoder to convert the categorical data , 'Type' in this case to numerical form.

## Classification Algorithms

Classification is a machine learning task that categorizes the class of an item to be tested. Based on the past observations during training, the classification algorithms identify the class of the new observation. The classifications are of the kinds like – binary classification, multi class classification. The paper has 11 different faults to be classified so it is a case of multi class classification. Once a new observation is given the classification algorithm should identify the kind of fault, if any, otherwise it will be identified as healthy. The paper implements the classification using the three commonly used algorithms KNN, SVM and Decision Tree as mentioned below. The subsequent section discusses the result analysis for each of the methods used.

### K-Nearest Neighbors (KNN): K nearest neighbors is a simple supervised algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN is a non-parametric and lazy learning algorithm. KNeighborsClassifier estimator object is available in sklearn.neighbors package,which has been used for experimentation using various values of K ranging from 1 to 25, considered emperically.

### Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both the classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features given) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the classes very well. In SVM, a hyper-plane is selected to best separate the points in the input variable space by their class, either class 0 or class 1. The hyper-plane is learned from training data using an optimization procedure that maximizes the margin. SVC estimator object is available in sklearn.svm package,which has been used for experimentation using non-linear kernel – Radial Basis Function (RBF).

### Decision Trees: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The tree is learned using a greedy algorithm on the training data to pick splits in the tree. Stopping criteria define how much tree learns and pruning can be used to improve a learned tree. While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories. As a result of which it turns up giving a very low accuracy. DecisionTreeClassifier estimator object is available in sklearn.tree package,which has been used for experimentation

## Performance Measure- Accuracy score

The learning done by machine for a particular task is evaluated using performance measure. There are various measures available depending on the estimators and domain. The paper reports results based on accuracy score - the most commonly used measure in supervised classification task with multi-labels. The predicted label should match with the true labels provided during supervised learning. The result is mentioned in percentage.

### 

# Result Analysis

The experiment is conducted on the synthetic dataset using the three classification algorithms. The whole dataset is split into two parts - one for training and another for testing. Scikit-learn provides *ModelSelection* package consisting of *train\_test\_split* class, which splits the dataset into various proportions. The dataset consists of 11300 instances. 80 percent of this is considered as training set and 20 percent as testdata. The comparison of the three algorithms based on test accuracy score is mentioned in Table 1 below. The average score is considered for kNN with k=1 to 25.

Table 1: Comparison of Classification Algorithms

|  |  |  |
| --- | --- | --- |
| SN | Algorithm | Accuracy (%) |
| 1 | Decision Tree | 84.2 |
| 2 | kNN | 86.15 |
| 3 | SVM | 91.06 |

# Conclusion And Future Scope

The paper focusses on classification of the overhead transmission line faults. The faulty data has been generated with the RMS values of three phase voltages and currents, by simulating each fault condition using MATLAB/Simulink. Total 11 types of faults have been considered. The data generated is downloaded in excel sheet. Data preprocessing steps are carried out such as feature selection, label encoder to classify the considered faults using classification algorithms of machine learning. Python and scikit learn library has been used for implementing classification algorithm. Three supervised classification algorithms have been used namely, Decision Tree, kNearestNeighbor and Support Vector Machine. The test accuracy score is used as comparison metric. SVM with Radial basis function as kernel outperformed the rest of the two. With more data instances the training can be made more robust and performance is expected to increase.

The work aims at extending its scope in identifying the exact location of the fault occurred. Fault identification is very significant for a large power system, but it is also important to know where exactly the fault occurred. This eases the effort of maintenance resulting in a more reliable power system.

##### References

[1] C L Wadhva, Electrical Power Systems, Seventh Edition, 2016.

[2] P. P. Pattanaik and C. K. Panigrahi, "Stability and fault analysis in a power network considering IEEE 14 bus system," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, pp. 1134-1138, doi: 10.1109/ICISC.2018.8398981, 2018.

[3] R. Malhotra and A. Jain, Fault Prediction Using Statistical and Machine Learning Methods for Improving Software Quality, Journal of Information Processing Systems, vol. 8, no. 2, pp. 241–262, doi: 10.3745/JIPS.2012.8.2.241, Jun. 2012.

[4] A. N. Hasan and B. Twala, Improving single classifiers prediction accuracy for underground water pump station in a gold mine using ensemble techniques, IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), Salamanca, pp. 1-7, 2015.

[5] Robert Grepl, Real-Time Control Prototyping in MATLAB/Simulink: Review of tools for research and education in mechatronics, IEEE International Conference on Mechatronics, doi: 10.1109/ICMECH.2011.5971238, 2011

[6] Tom M. Mitchell, Machine Learning, McGraw-Hill , Oct 1997

[7] Scikit-learn: Machine learning in Python, https://scikit-learn.org/