A REPORT

ON

Pattern Recognition Using Machine Learning and Deep Learning Techniques

BY

Name(s) of Student(s)	the	ID.No.(s)
Raghav Dhawan		2022B1A81010
Tarun Warrier		2022AAPS0122G

AT

Southern Regional Load Dispatch Centre(SRLDC)

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

May-July, 2024

A REPORT

ON

Pattern Recognition Using Machine Learning and Deep Learning Techniques

BY

Name(s) of the Student(s)	ID.No.(s)	Discipline(s)
Raghav Dhawan	2022B1A81010P	MSC. Biological Sciences and B.E. ENI
Tarun Warrier	2022AAPS0122G	B.E. ECE

Prepared in partial fulfilment of the
Practice School-I Course Nos.
BITS C221/BITS C231/BITS C241

AT

Southern Regional Load Dispatch Centre

A Practice School-I Station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

May-July, 2024

Acknowledgement

I would like to express my heartfelt gratitude to Dr. Selva Balaji, Faculty in Charge, Practice School Division, Birla Institute of Technology and Science (BITS), for providing me with the opportunity to work on pattern recognition in time series data as part of my Practice School program.

I am also immensely thankful to Mr. MVD Raghava, my mentor, and Mr. Muthu Kumar, Deputy General Manager, for their invaluable guidance, support, and encouragement throughout this experience. Their expertise and insights, as well as Mr. Muthu Kumar's efforts in ensuring the smooth allotment of project work and operations, have greatly contributed to my learning and professional growth.

This opportunity has significantly enhanced my practical knowledge and skills, and I am sincerely grateful to Dr. Selva Balaji, Mr. MVD Raghava, and Mr. Muthu Kumar for their support and mentorship during this pivotal phase of my academic journey.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI
(RAJASTHAN)
Practice School Division

Station: SRLDC Bengaluru

Centre: Bengaluru

Duration 28th May to 23rd July

Date of Start: 28th May, 2024

Date of Submission :23rd May 2024
Title of the Project: Pattern Recognition Using Machine Learning and Deep Learning Techniques

ID No./Name(s)/ Discipline(s)/of the student(s): Tarun Warrier (2022AAPS0122G), Raghav Dhawan (2022B1A81010P)

Name(s) and designation(s) of the expert(s): Mr. Raghava MVD(Project Mentor)

Name(s) of the PS Faculty: Dr. Selva Balaji

Key Words: Deep learning, Machine learning, Python

Project Areas: Time series voltage data

Abstract: This project focuses on developing a robust model for pattern recognition in time series data, enabling SRLDC to promptly identify and respond to anomalies and disturbances, thereby ensuring a stable and secure power transmission network across the Southern region of India.

Signature(s) Of Student(s)

Signature of PS Faculty

Date

Date

Table of Contents

Introduction:

Voltage time series data is essential for monitoring and stabilising power systems. Effective pattern identification in this data can greatly enhance fault detection and grid reliability. Traditional methods often fail to handle the complexity and dynamic nature of voltage variations, necessitating advanced approaches. This research aims to develop a comprehensive pattern identification system for voltage time series data using advanced machine learning techniques. Models such as Dynamic Time Warping (DTW) are employed to accurately detect and classify patterns and anomalies. The primary objectives are to monitor voltage levels in real-time, diagnose faults efficiently, and respond swiftly to anomalies, thus improving grid stability and security.

We explored various clustering methods, including k-means, k-medoids, hierarchical clustering, DBSCAN, and Gaussian Mixture Models. K-means clustering divides data into pre-specified K clusters based on the nearest mean, while k-medoids use actual data points as centres, handling outliers better but being more computationally intensive. Hierarchical clustering, both agglomerative and divisive, provides detailed structures but is computationally expensive for large datasets. DBSCAN groups closely packed points based on distance metrics, identifying arbitrary-shaped clusters resistant to noise and outliers. Gaussian Mixture Models use a probabilistic approach, adaptable but sensitive to initialization and computationally intensive.

Similarity measures such as cosine similarity and DTW similarity were also explored. Cosine similarity assesses the orientation similarity between two vectors, useful for high-dimensional data but less effective for voltage time series. DTW similarity measures resemblance between sequences that differ in speed or time, making it ideal for voltage time series analysis.

Machine Learning and Deep Learning Models:

- **Regression-Based Models:**
    - **Linear Regression:** Fits a linear relationship between input features and the voltage. Simple and computationally efficient but may not capture complex patterns.
    - **Support Vector Regression (SVR):** Extends SVM to regression, effective in high-dimensional spaces with versatile kernel functions, though parameter selection is crucial.
- **Deep Learning Models:**
    - **Long Short-Term Memory (LSTM) Networks:** Capture temporal patterns and long-term dependencies, effective for complex dynamics but computationally intensive and prone to overfitting without regularisation.
    - **Convolutional Neural Networks (CNNs):** Adapted for time series by treating temporal segments as spatial dimensions, effective for local patterns but require careful tuning.
    - **Hybrid CNN-LSTM Models:** Combine CNNs for feature extraction and LSTMs for sequence modelling, leveraging strengths of both but complex to design and train, and susceptible to overfitting.

Model Training and Evaluation:

- **Data Preparation:** Involves preprocessing, normalisation, feature extraction, and data augmentation to enhance diversity.
- **Training:** Models are trained on labelled data, associating input features with target outputs.
- **Validation and Testing:** Ensure models generalise well by evaluating performance on unseen data, using metrics like accuracy, precision, recall, and F1-score.

Problem Statement:

The primary objective is a pattern identification system for voltage time series data using machine learning or deep learning techniques. The objective are:

1. Accurately detect and classify patterns and anomalies in voltage time series data.
2. Segregation of unlabelled fault signatures

Fault Classification:

- Develop supervised learning models (e.g., SVM, Random Forest, XGBoost) to classify different types of faults based on historical labelled data.
- Train deep learning models (e.g., LSTM, GRU, TCN) to capture temporal dependencies and improve classification accuracy.

Pattern Recognition:

- Implement pattern recognition techniques to identify recurring patterns and anomalies in the voltage time series data.
- Use techniques like Dynamic Time Warping (DTW), Structural Similarity Index Measure (SSIM), and convolutional neural networks (CNNs) for effective pattern matching.

Brief view of Approach:

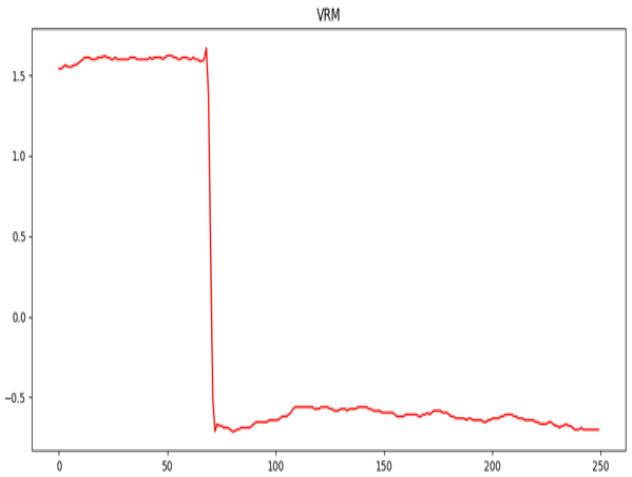
When starting with unlabeled data, supervised learning methods are not an option. Hence, we turned to unsupervised methods to effectively segregate the data.

Our approach can be broken down into several key steps:

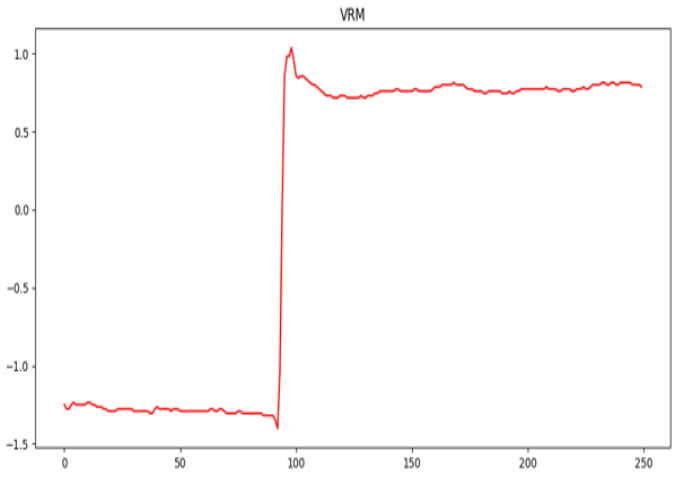
- **Feature Extraction:**When working with voltage time series data, feature extraction involves transforming raw signals into a meaningful set of features that highlight significant data points. This process requires understanding the behaviour, patterns, and anomalies in the voltage readings. By using a sliding window of size 30 and eliminating noise through cross-checking similar features in VRM, VYM, and VBM, we can reduce the impact of noise on accuracy, thereby enhancing performance.
    - **Peak Features:** Features like the number of peaks, and the width of the peaks at three levels namely, the topmost point of the peak, the middle region of the peak and the base point of the peak.
    - **Dip Features:**Features like the number of dips, and the width of the dips at three levels namely, the bottommost point of the dip, the middle region of the dip and the base point of the dip.
- **Clustering:** By using the above features, we clustered the data, grouping files based on the number of peaks and dips they contained. This method proved to be effective in identifying the majority of data patterns, for a cluster number of 16. It proved to be effective and returned certain outliers.
- **Fault Segregation:**
    - **One Dip Faults:** For faults characterised by a single dip, we further segregated them based on their widths. This allowed us to distinguish between Delayed clearance Faults (DCF) and Faults with trailing slope. However, Carrier-receiver Faults (CRF) could not be separated using this method.
    - **One Peak Faults:** We applied similar steps to faults with one peak, using the extracted features to cluster and segregate them accordingly as faults with a single peak or single dip are pretty similar.
    - **Tripping Faults Separation Using DTW:** To separate tripping faults, we utilised Dynamic Time Warping (DTW), an algorithm designed to measure similarity between two temporal sequences that may vary in speed. DTW aligns these sequences in a non-linear fashion to minimise the distance between them. By setting a threshold of around 1, we successfully separated tripping faults, finding this method to be effective.
- **Data Augmentation:** Data augmentation techniques are used to artificially increase the size and diversity of the dataset, which is particularly beneficial when dealing with limited data or imbalanced classes in voltage time series analysis. Some methods of data augmentation used are:

    - **Time Shifts:** Shifting the entire time series forward or backward by a random or fixed number of time steps. Introduces variability in temporal patterns and helps the model generalise better across different time shifts.
    - **Noise Injection:** Adding random noise to the voltage signal. Mimics measurement errors or variability in real-world data, making the model more robust to noise.
    - **Resampling:** Changing the sampling rate or interval of the voltage time series.Adjusts the temporal resolution of the data, potentially revealing different patterns or improving computational efficiency.
    - **Data Imputation:**Filling missing data points or segments in the time series.Enhances the completeness of the dataset, ensuring robustness in handling missing or incomplete data.
    - **Generating Synthetic data:** Generating based on certain known features to replicate data.
- **Conversion to heatmaps:** All the time series data is converted into heatmaps to be compatible with a hybrid CNN-LSTM model.
- **Hybrid CNN-LSTM model:** Finally in an effort to separate CRF files, we implemented a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model. Unfortunately, due to the limited dataset, the model tended to memorise the training data, resulting in poor performance on unseen data. This indicates a need for further improvement in this area.

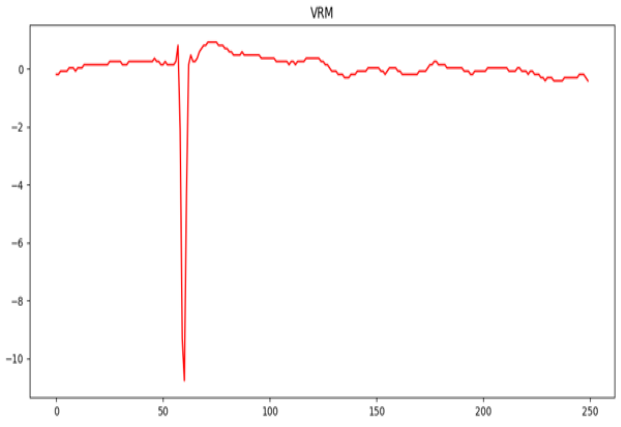
Identified Faults:



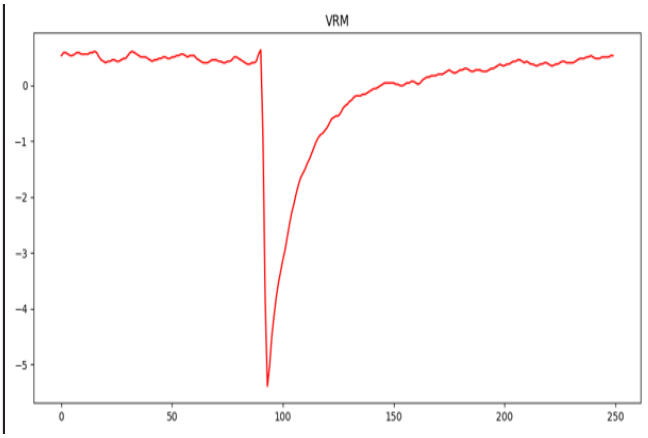
Fault 1



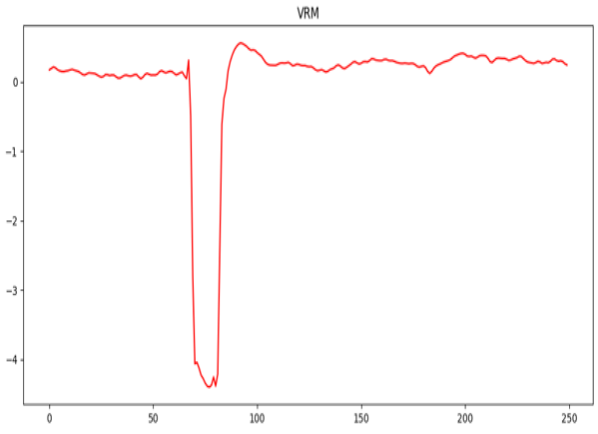
Fault 2



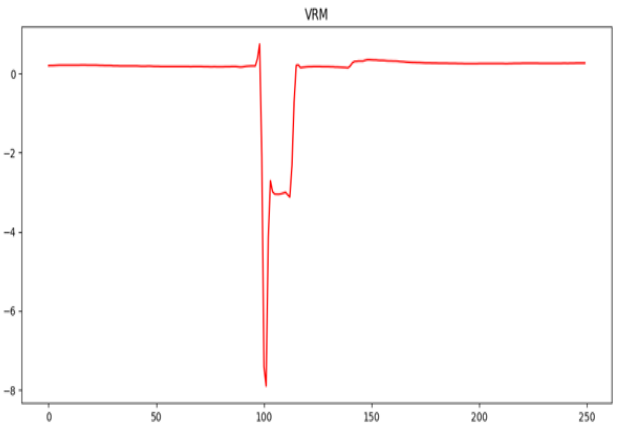
Fault 3



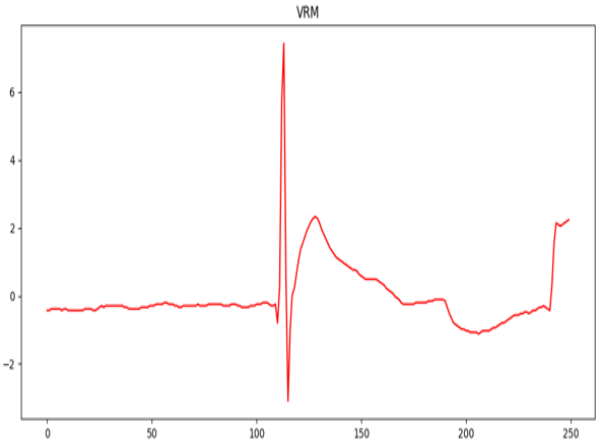
Fault 4



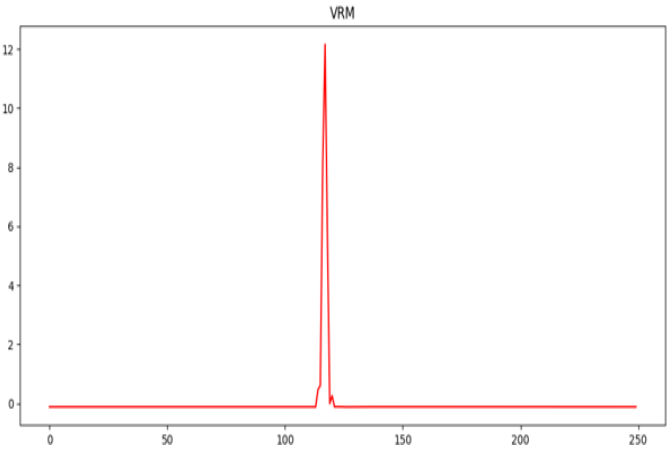
Fault 5



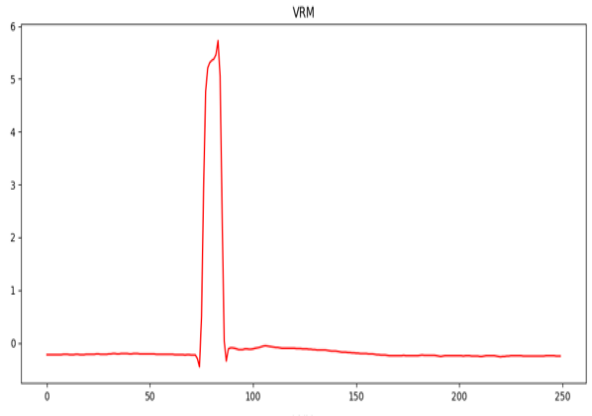
Fault 6



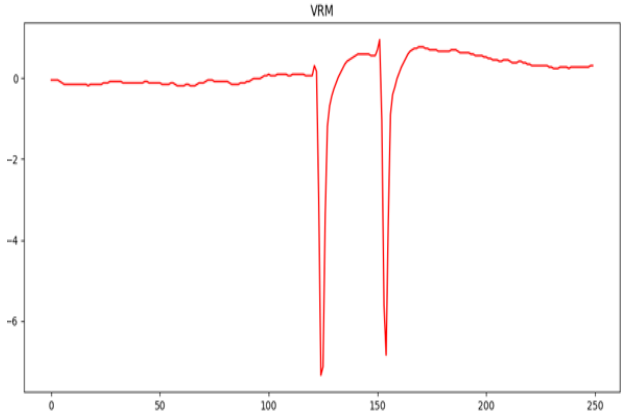
Fault 7



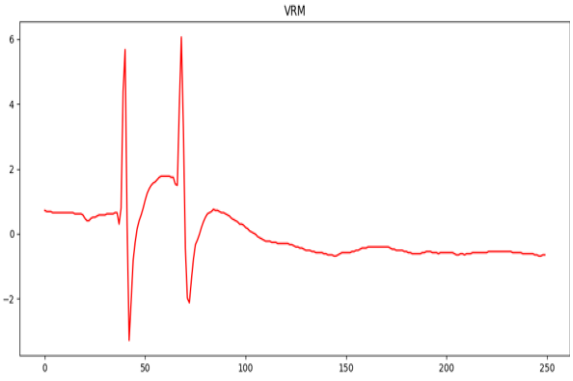
Fault 8



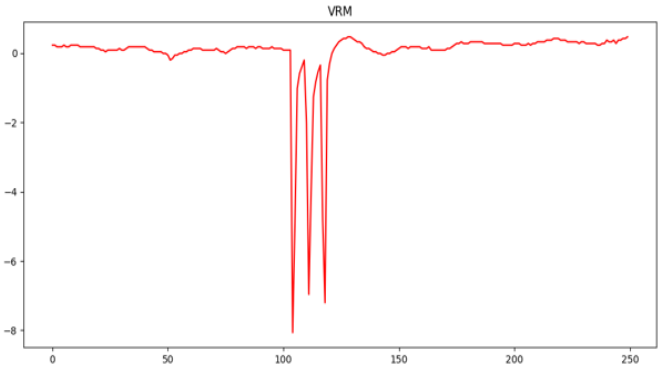
Fault 9



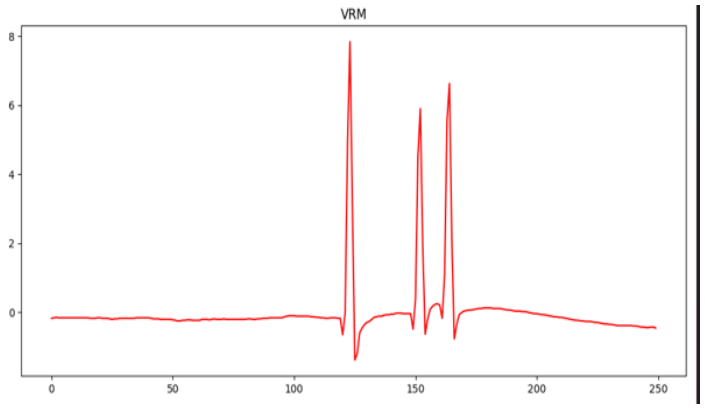
Fault 10



Fault 11



Fault 12



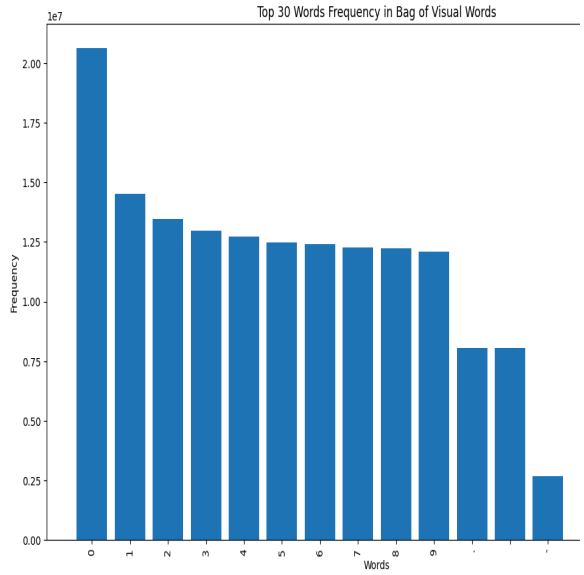
Fault 13

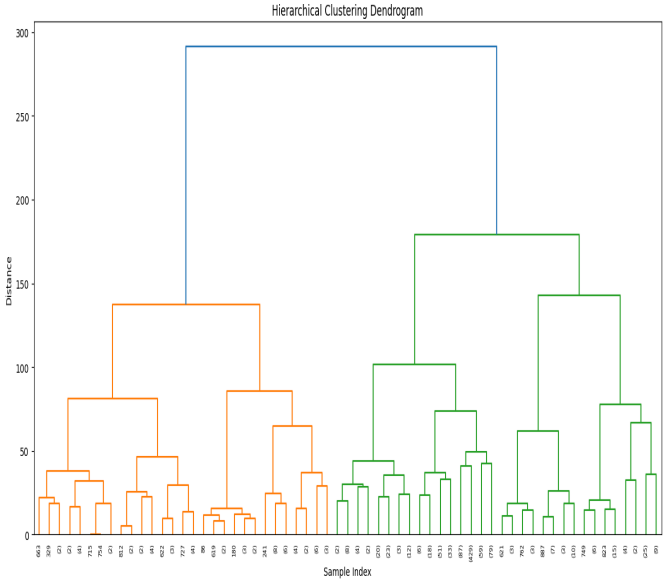
TOOLS AND PACKAGES USED :

- Python
- Sklearn
- tslearn
- Cv2
- PIL
- Tensorflow
- Keras
- Matplotlib
- Seaborn
- Pandas
- numpy

RESULTS:

Bag of Visual words and Hierarchical Clustering:





Regression classifier Models:(LightGBM,XGBoost,Decision Tree, Random Forest and Neural Network)

```
LightGBM Classifier - Accuracy: 0.6050, Precision: 0.5073, Recall: 0.6050
```

```
XGBoost Classifier - Accuracy: 0.5311, Precision: 0.5289, Recall: 0.5311
```

```
Decision Tree - Accuracy: 0.5210, Precision: 0.4839, Recall: 0.5210
```

```
Random Forest - Accuracy: 0.5210, Precision: 0.4787,
```

```
Neural Network - Accuracy: 0.6303, Precision: 0.5050, Recall: 0.6303
```

Hybrid CNN-LSTM:

```
8/8 - 19s - 2s/step - accuracy: 0.9918 - loss: 0.1342
Test accuracy: 0.9918367266654968
8/8 ━━━━━━━━━━━━━━━━━━━━ 30s 3s/step
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       125
           1       0.99      0.99      0.99       120

    accuracy                           0.99       245
   macro avg       0.99      0.99      0.99       245
weighted avg       0.99      0.99      0.99       245
```

Future scope:

Data Augmentation and Collection:

- Increasing Dataset Size: Collecting more comprehensive and diverse datasets from various grid locations and conditions will help improve the models' robustness and generalizability.
- Synthetic Data Generation: Developing methods for generating synthetic voltage time series data could augment training data, especially for rare fault conditions.

Advanced Feature Engineering:

- Automated Feature Extraction: Implementing advanced feature extraction techniques, such as autoencoders or deep feature synthesis, could capture more complex patterns and improve clustering and classification accuracy.
- Domain-Specific Features: Incorporating domain knowledge to engineer features that are particularly relevant to electrical grids could enhance model performance.

Enhanced Model Architectures:

- Ensemble Learning: Combining multiple machine learning and deep learning models to create ensemble methods could enhance prediction accuracy and reliability.
- Attention Mechanisms: Integrating attention mechanisms into LSTM or hybrid models could help focus on important time series segments, improving fault detection accuracy.

Explainability and Interpretability:

- Model Explainability: Developing methods to make the models more interpretable will help operators understand the decision-making process, increasing trust and facilitating adoption.
- Visualization Tools: Creating advanced visualization tools to display real-time data, model predictions, and fault diagnostics will aid in quicker and more informed decision-making.

CONCLUSION:

Finally, incorporating machine learning into fault categorization for sequence detectors represents a significant step forward in industrial and technological sectors. Organisations may improve the accuracy of issue detection and classification by leveraging ML algorithms, automate diagnosis processes to reduce downtime, and implement proactive maintenance strategies via real-time monitoring. These innovations not only increase operational efficiency, but also allow companies to respond quickly to changing fault patterns and dynamic surroundings. Looking ahead, continued research and development in ML interpretability, scalability across industries, and seamless integration with IoT and sensor networks will drive fault management innovation, ultimately leading to safer, more reliable, and efficient systems around the world.

References:

- **Lee, J.-H., Kang, J., Shim, W., Chung, H.-S., & Sung, T.-E. (2020).** Pattern Detection Model Using a Deep Learning Algorithm for Power Data Analysis in Abnormal Conditions. *Electronics*, 9(7), 1140. https://doi.org/10.3390/electronics9071140
- **Kang, S., Li, R., Zhao, L., & Li, Z. (2023).** Feature extraction based on time-series topological analysis for the partial discharge pattern recognition of high-voltage power cables. *Measurement*, 217, 113009. https://doi.org/10.1016/j.measurement.2023.113009
- **Uptime AI.** Pattern Recognition and Machine Learning. https://www.uptimeai.com/resources/pattern-recognition-and-machine-learning/
- **Stack Overflow.** Pattern detection in time series data. https://stackoverflow.com/questions/36549932/pattern-detection-in-time-series-data
- **Baeldung.** Pattern Recognition in Time Series. https://www.baeldung.com/cs/pattern-recognition-time-series