Comparative Analysis Between Fuzzy Theorem and KNN Methodology for Fault and Anomaly Detection

Ankit Dogra, Vinayak Kumawat, and Neetu Gupta(✉)

Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
{ankit.219302300,vinayak.219301086}@muj.manipal.edu, neetu.gupta@jaipur.manipal.edu

Abstract. Fault and anomaly detection systems are important to ensure the reliability of complex systems in various fields. This study presents a comprehensive comparative analysis between two main fault and anomaly detection strategies: fuzzy systems and k-nearest neighbour methods. Our study investigates the efficacy of each approach in identifying anomalies in a complex dataset. Our research aims to identify ideal data scenarios in which each technology excels, providing valuable insights across all technology sectors. By examining predefined parameters and input-based performance variations, we aim to guide domain-specific choices for error and anomaly detection systems. This research lays the foundation for identifying sectors that will benefit most from implementing these approaches, helping to improve the flexibility and reliability of complex systems.

Keywords: Fuzzy Theorem · K-Nearest Neighbours · Fault Detection · Anomaly Detection · Machine Learning

1 Introduction

The growing complexity of systems across various domains have heightened the significance of fault and anomaly detection. Detection of anomalies is pivotal in ensuring the reliability, safety, and optimal performance of systems, ranging from cybersecurity to industrial machinery [1]. As the reliance on data-driven approaches for anomaly detection continues to rise, researchers and practitioners are exploring diverse methodologies to enhance accuracy and efficiency.

Our research focuses on conducting a comparative analysis between two prominent methodologies: the Fuzzy Theorem, implemented through the Fuzzy C-Means (FCM) algorithm, and the K-Nearest Neighbours (KNN) methodology. Anomalies, often indicative of faults or irregularities, pose critical challenges in real-world applications. The selection of an appropriate methodology is crucial for achieving reliable and efficient anomaly detection.

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 25–38, 2025.
https://doi.org/10.1007/978-3-031-77075-3_3

The motivation behind this research stems from the need to comprehensively understand and compare the efficacy of Fuzzy Theorem and KNN methodologies in the context of fault and anomaly detection. While existing literature provides insights into individual methodologies, a direct comparison under varied conditions and datasets can illuminate their relative strengths and weaknesses.

The primary objectives of this research are as follows:

Conducting a Comparative Analysis: Evaluating and comparing the performance of Fuzzy Theorem (FCM) and KNN methodologies for anomaly detection under diverse conditions.

Exploring Feature Selection Methods: Investigating the impact of different feature selection methods, such as Mutual Information, SelectKBest, and Random Forest Classifier, on the performance of KNN models.

Providing Practical Insights: Offering practical insights and recommendations for selecting the most suitable methodology based on specific application requirements and dataset characteristics.

2 Literature Review

2.1 Fuzzy C-Means for Anomaly Detection

The approach employed in [2] leverages fuzzy logic to capture the uncertainty inherent in anomaly detection. The model employs decision trees to isolate anomalies efficiently, and the incorporation of fuzzy logic enhances its ability to handle complex and uncertain data patterns.

In [3], by incorporating fuzzy sets and rules, the model extends its ability to discern anomalies in situations where traditional methods may fall short. The study likely explores the nuanced decision-making process enabled by fuzzy logic.

Conducting a comparative study, the authors of [4] evaluate various anomaly detection methods for gross error detection problems. Focusing on their effectiveness in identifying and handling anomalies, the study provides insights into the comparative performance of different anomaly detection approaches.

2.2 FCM for Class Balancing and Anomaly Detection

This authors of [5] focus on the application of Fuzzy C-Means (FCM) in wireless sensor networks. By incorporating fuzzy logic, the proposed approach aims to achieve similarity-aware data aggregation, enhancing the efficiency of data processing and anomaly detection in wireless sensor networks.

Addressing the domain of transformer fault detection, this paper introduces an improved Fuzzy C-Means clustering algorithm. The authors of [6] explore enhancements to traditional FCM to better identify and classify transformer faults, thereby contributing to the reliability and performance of power systems.

2.3 Class Balancing and K-Nearest Neighbors

Focusing on power transformer operation data, [7] explores outlier detection and data filling using K-Nearest Neighbors (KNN) and Local Outlier Factor (LOF). The application of KNN, possibly complemented by fuzzy logic, contributes to accurate classification and management of power transformer operation data.

2.4 K-Nearest Neighbors for Classification

[8] introduces a data-driven heart disease prediction model through K-Means clustering-based anomaly detection. By leveraging the clustering capabilities of K-Means, the study contributes to the early identification of anomalies in heart disease datasets, demonstrating the potential for improved diagnostic accuracy.

3 Methodology

3.1 Preprocessing

We took the dataset and selected features V1 - V28, where we got –

Count of 0s: 284315

Count of 1s: 492

With such severe class imbalance, we will have to balance the target class. To do that we will be using Combining Oversampling and Undersampling. A hybrid approach that involves both oversampling the minority class and undersampling the majority class can sometimes be effective. This combination aims to balance the class distribution while retaining important data from both classes. Models we will be using for class balancing Combining oversampling and undersampling techniques can help mitigate class imbalance effectively. The idea is to oversample the minority class and undersample the majority class. One common method for doing this is to apply SMOTE to oversample the minority class and Tomek links to undersample the majority class.

Updated Output –

Count of 0s: 189878

Count of 1s: 189878

3.2 Data Scaling and Normalization

We perform Shapiro Wilk statistical test to visualize the distribution of our data.

Result of the test are shown in Figs. 1 and 2.

```
V1: Statistics=0.7069604396820068, p=0.0
V2: Statistics=0.8459051251411438, p=0.0
V3: Statistics=0.7683205604553223, p=0.0
V4: Statistics=0.9536792039871216, p=0.0
V5: Statistics=0.7399948239326477, p=0.0
V6: Statistics=0.9486317038536072, p=0.0
V7: Statistics=0.6992595195770264, p=0.0
V8: Statistics=0.5835127234458923, p=0.0
V9: Statistics=0.9311937689781189, p=0.0
V10: Statistics=0.833716630935669, p=0.0
V11: Statistics=0.9252386689186096, p=0.0
V12: Statistics=0.8450649976730347, p=0.0
V13: Statistics=0.9987892508506775, p=1.4566838780858506e-36
V14: Statistics=0.8777220249176025, p=0.0
V15: Statistics=0.9890421628952026, p=0.0
V16: Statistics=0.8347029685974121, p=0.0
V17: Statistics=0.7863301634788513, p=0.0
V18: Statistics=0.8606085181236267, p=0.0
V19: Statistics=0.9806700348854065, p=0.0
V20: Statistics=0.7257179021835327, p=0.0
V21: Statistics=0.5168805718421936, p=0.0
V22: Statistics=0.8532589077949524, p=0.0
V23: Statistics=0.507538914680481, p=0.0
V24: Statistics=0.9850587844848633, p=0.0
V25: Statistics=0.959875762462616, p=0.0
V26: Statistics=0.9844372272491455, p=0.0
V27: Statistics=0.7529212832450867, p=0.0
V28: Statistics=0.7895416021347046, p=0.0
Class: Statistics=0.6367088556289673, p=0.0
```

Since the p value of each feature is less than $0 = 0.05$ we can say that our data is not normally distributed.

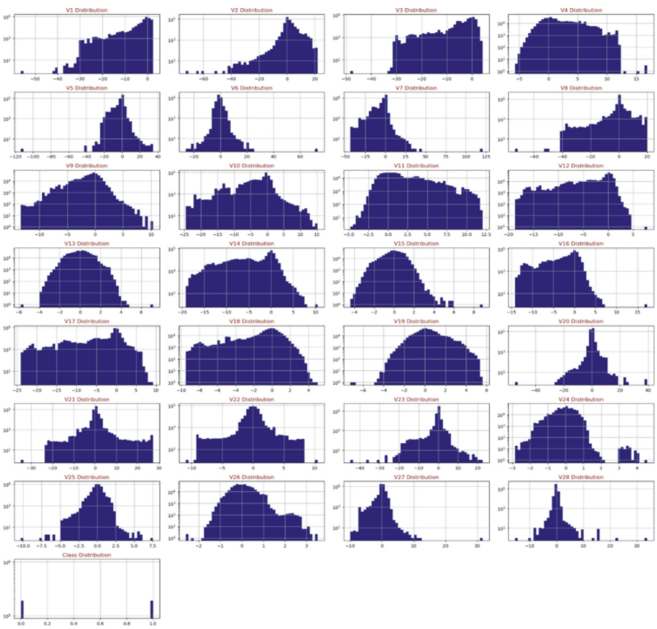

Fig. 1. Visualization of distribution of each feature using histograms. Normally distributed data should resemble a bell curve.

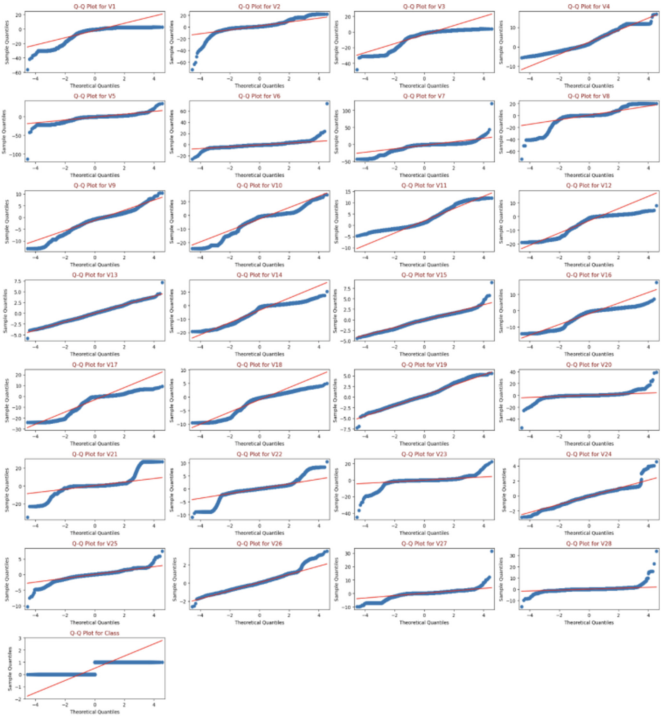

Fig. 2. Quantile-quantile (Q-Q) plots

Through these visualizations we can confirm that our data is not normally distributed, and we may need to normalize it but foremost we want to reduce the difference between the data points for better performance in KNN.

Standard scaling is a technique that adjusts each feature to have a value of 0 and a standard deviation of 1. This approach is beneficial for algorithms that are influenced by the magnitude of input features like optimization algorithms based on gradients used in machine learning models. Standardization aids in convergence of algorithms. Can enhance the performance of models such as support vector machines or k nearest neighbours [9, 10].

- Equal Weight to Features: Scaling ensures that all features contribute equally when calculating distances. Scaling features with scales might exert an influence on the distance calculation, which could potentially lead to biased outcomes.
- Improved Convergence: Scaling assists the algorithm in converging during distance-based calculations thereby enhancing its efficiency.
- Better Performance: The performance of k neighbors (KNN) can be affected by feature scales. By scaling the features, you create a playing field for all variables potentially improving the performance of the algorithm.

Results after normalization are shown in Figs. 3 and 4.

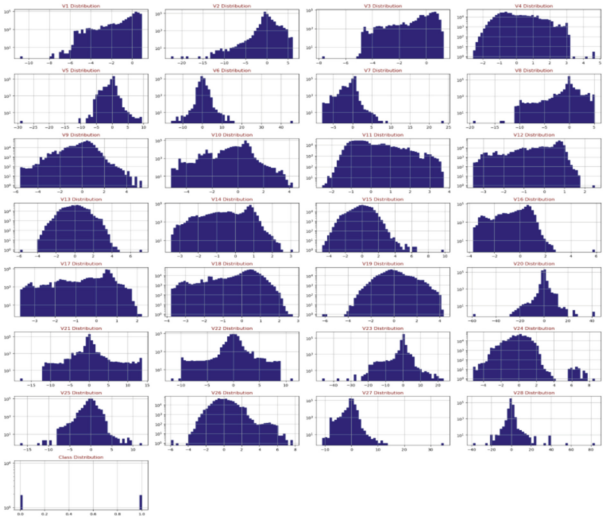

Fig. 3. Visualization of distribution of each feature using histograms after data normalization.

From the above plots we can see that there has been no change in data distribution hence we can say the data was already normalized.

3.3 Fuzzy Theorem Implementation

Fuzzy C-Means (FCM):

Fuzzy Clustering: FCM is a widely used fuzzy clustering algorithm that extends the traditional K-Means algorithm to accommodate fuzziness.

Fuzzy Sets: Membership values are fuzzy, representing the uncertainty or ambiguity in cluster assignments.

Objective Function: The goal of FCM is to minimize the following objective function:

$$J_m(U, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} ||x_i - c_j||^2 \quad (1)$$

where:

- U = $\{u_{ij}\}$ is the fuzzy membership matrix,
- C = $\{c_j\}$ is the set of cluster centroids,
- m is a fuzziness parameter, and
- n is the number of data points.

Algorithm Steps:

- Initialization
- Update Membership Matrix
- Update Cluster Centroids
- Repeat (from update membership matrix)

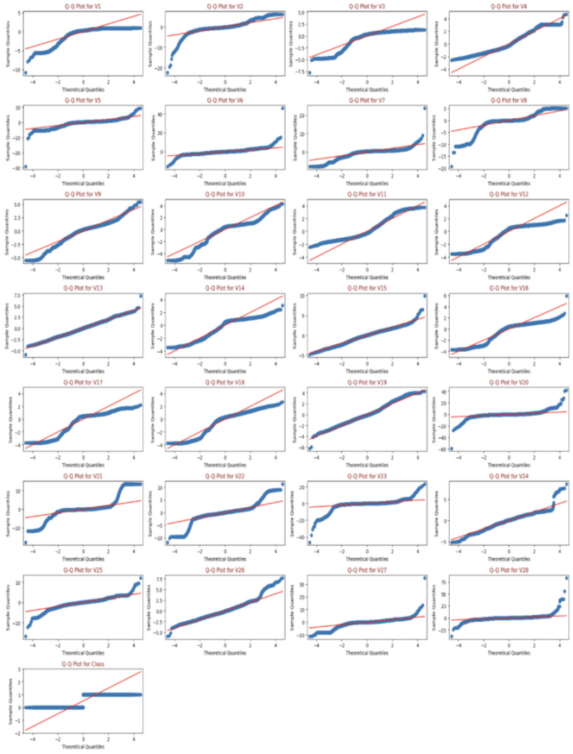

Fig. 4. (Q-Q) plot after data normalization

3.4 K-Nearest Neighbors (KNN) Implementation

KNN is a supervised machine learning algorithm used for classification and regression tasks. It classifies a data point by considering the majority class of its k nearest neighbors in the feature space [11, 12].

Distance Metric: The choice of distance metric, such as Euclidean distance, Manhattan distance, or others, determines how "closeness" is measured between data points.

Algorithm Steps:

- Training
- Prediction
- Identify Neighbors
- Majority Vote

Feature Selection: Feature selection is crucial for KNN. We used methods like Mutual Information, SelectKBest, and Random Forest Classifier to select informative features [13, 14].

4 Results

The following heat maps shown in Figs. 5 and 6 represent the clustering in the form of confusion matrix of both fuzzy c-means and KNN and these are used as foundation to evaluate the performance of both the approaches [15]. Roc curve comparison for FCM and KNN are represented in Figs. 7 and 8 respectively. Precision recall comparison for FCM and KNN technologies are shown in Figs. 9 and 10 respectively. Figures 11 and 12 represent Recall Comparison for various parameters and models of FCM and KNN respectively.

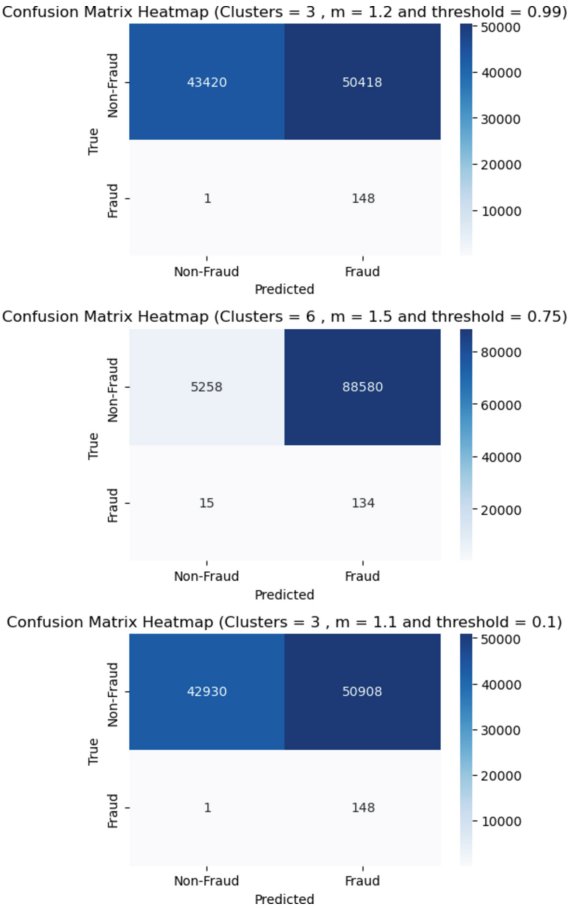

Fig. 5. Heatmap representation of Confusion matrices of FCM

The F2 score as shown in Figs. 13 and 14, is a measure that combines precision and recall, with a greater emphasis on recall. It is defined as follows:

$$F2 = \frac{(1+\beta^2).Precision.Recall}{(\beta^2.Precision) + Recall} \tag{2}$$

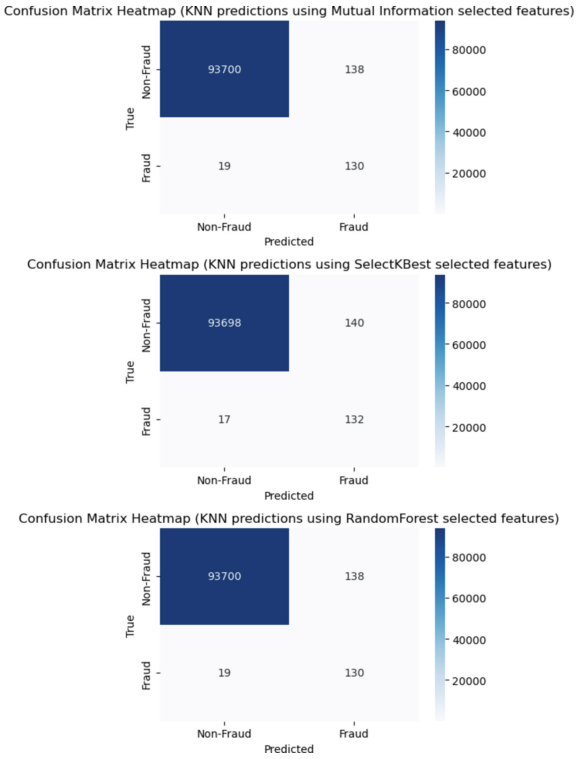

Fig. 6. Heatmap representation of Confusion matrices of KNN

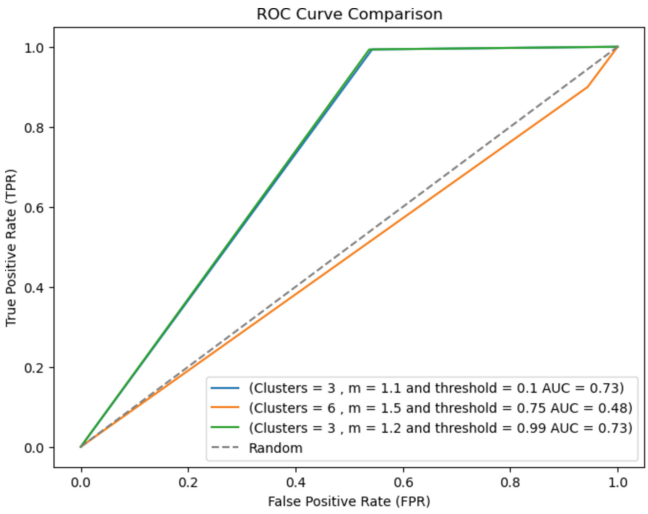

Fig. 7. ROC Curve Comparison of FCM

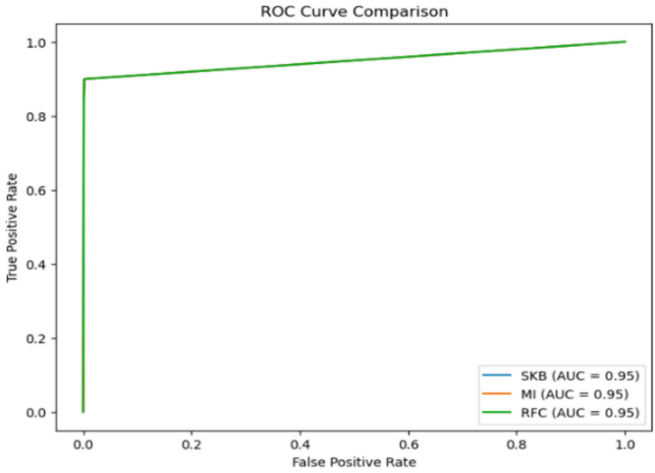

Fig. 8. ROC Curve Comparison of KNN

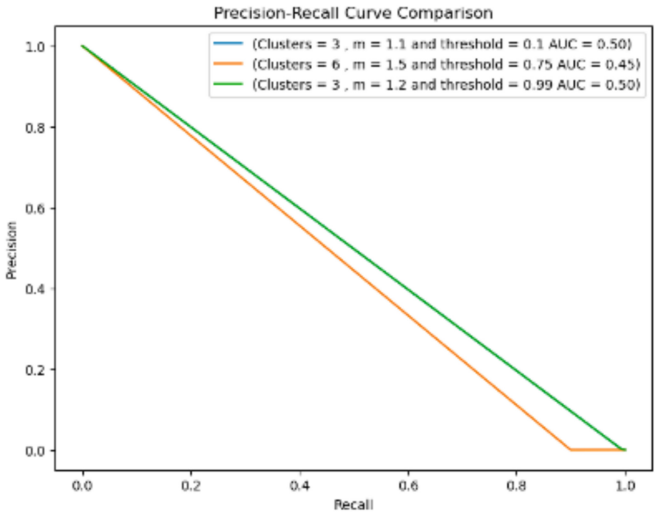

Fig. 9. Precision-Recall Curve Comparison of FCM

Here, β is a parameter that controls the weight of precision in the combined score. For the F2 score, β is set to 2, which means recall is given twice the weight of precision.

$$F2 = \frac{5.Precision.Recall}{4.Precision + Recall} \tag{3}$$

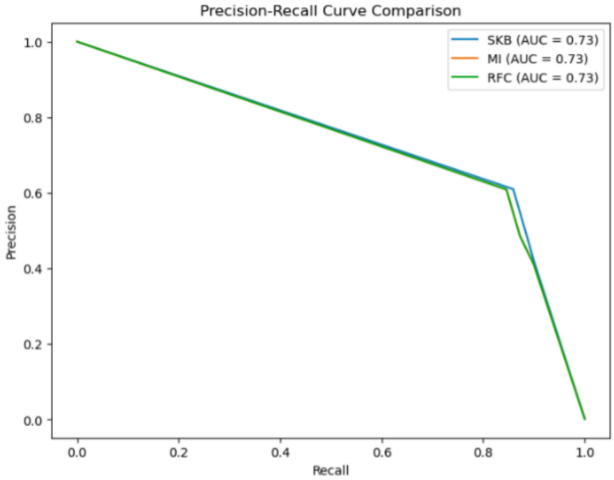

Fig. 10. Precision-Recall Curve Comparison of KNN

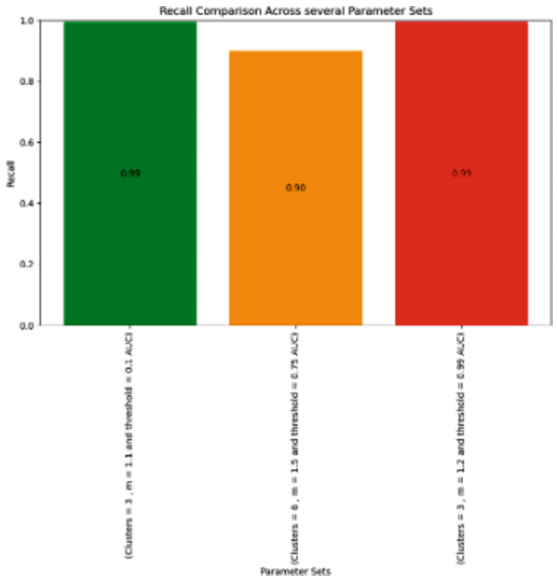

Fig. 11. Recall Comparison for various parameters and models of FCM

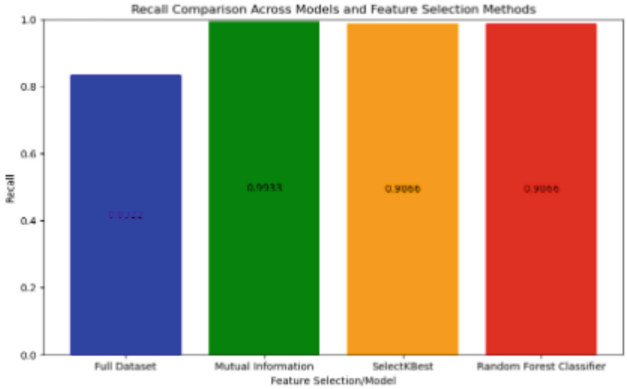

Fig. 12. Recall Comparison for various parameters and models of KNN

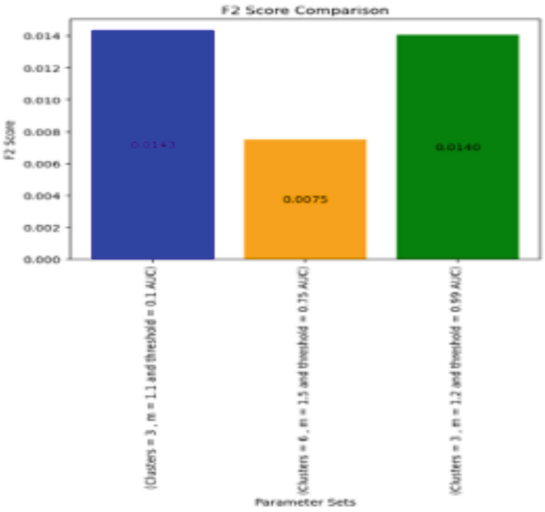

Fig. 13. F2 Score Comparison for various parameters and models of FCM

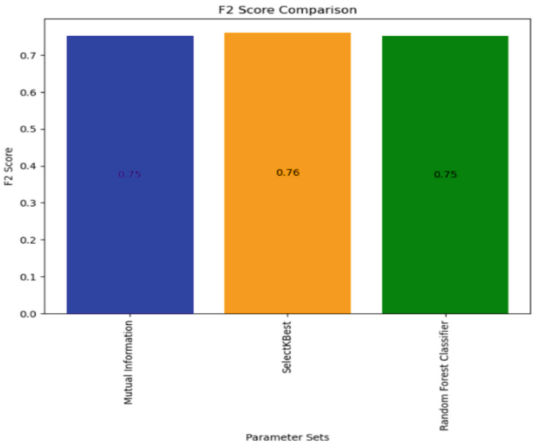

Fig. 14. F2 Score Comparison for various parameters and models of KNN

5 Conclusion

The results of FCM, when compared to KNN, revealed competitive performance. However, the Fuzzy Logic methodology introduced an interpretability aspect through cluster assignments, providing additional insights into the nature of anomalies within the dataset.

The comparative analysis between KNN and Fuzzy Logic methodologies highlighted their distinct strengths and weaknesses. KNN excelled in precision, recall, and accuracy, particularly when applied to datasets with carefully selected features. On the other hand, Fuzzy Logic, while slightly less accurate, introduced a nuanced understanding of anomalies through fuzzy cluster assignments.

The computational requirements for KNN heavily depend on the dataset size and the chosen feature selection methods. While KNN is generally fast during the prediction phase, the training phase, and the evaluation of distances for each prediction can become resource-intensive for large datasets.

Fuzzy Logic, particularly when applied through clustering algorithms like Fuzzy C-Means, introduces additional computational complexity. The iterative nature of the clustering process, where membership degrees are continuously updated, contributes to increased time complexity. The choice of parameters, such as the number of clusters and fuzziness coefficient, further influences the computational demand.

The choice between KNN and Fuzzy Logic methodologies depends on the specific requirements of the application. Both approaches demonstrate merit in fault and anomaly detection, providing researchers and practitioners with valuable tools for addressing real-world challenges.

References

1. Huang, B.: Detection of abrupt changes of total lest square models and application in fault detection. IEEE Trans. Control Syst. Technol. **9**(2), 357–367 (2001)
2. Chater, M., Borgi, A., Slama, M.T., Sfar-Gandoura, K., Landoulsi, M.I.: Fuzzy isolation forest for anomaly detection. Procedia Comput. Sc. **207**, 916–925 (2022)
3. Gladkykh, T., Hnot, T., Solskyy, V.: Fuzzy logic inference for unsupervised anomaly detection. In: IEEE First International Conference on Data Stream Mining & Processing (DSMP), pp. 42–47. IEEE (2016)
4. Dobos, D., et al.: A comparative study of anomaly detection methods for gross error detection problems. Comput. Chem. Eng. **175**, 108263 (2023)
5. Wan, R., Xiong, N., Hu, Q., Wang, H., Shang, J.: Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks. EURASIP J. Wirel. Commun. Netw. **2019**, 1–11 (2019)
6. Tang, S., Peng, G., Zhong, Z.: An improved fuzzy C-means clustering algorithm for transformer fault. In: China International Conference on Electricity Distribution (CICED), 1–5. IEEE (2016)
7. Zou, D., et al.: Outlier detection and data filling based on KNN and LOF for power transformer operation data classification. Energy Rep. **9**, 698–711 (2023)
8. Ripan, R.C., et al.: A data-driven heart disease prediction model through K-means clustering-based anomaly detection. SN Comput. Sci. **2**, 1–12 (2021)
9. Mollazade, K., Ahmadi, H., Omid, M.: An intelligent model based on data mining and fuzzy logic for fault diagnosis of external gear hydraulic pumps. Insight **51**, 594–600 (2009)
10. Chen, C.H., Shyu, R.J., Ma, C.K.: A new fault diagnosis method of rotating machinery. J. Shock Vibrat. **15**(6), 585–598 (2008)
11. Yang, B.S., Han, T., Hwang, W.W.: Fault diagnosis of rotating machinery based on multi-class support vector machines. J. Mech. Sci. Technol. **19**(3), 846–859 (2005)
12. Bagheri, B., Ahmadi, H., Labbafi, R.: Application of data mining and feature extraction on intelligent fault diagnosis by artificial neural network and k-nearest neighbor. In: Proceedings of the IEEE Electrical Machines Conference, pp. 1–7 (2010)
13. Ebrahimi, E., Mollazade, K.: Intelligent fault classification of a tractor starter motor using vibration monitoring and adaptive neuro-fuzzy inference system. Insight **52**(10), 561–566 (2010)
14. Zhang, L., Lin, J., Karim, R.: An angle-based subspace anomaly detection approach to high-dimensional data with an application to industrial fault detection. Reliab. Eng. Syst. Saf. **142**, 482–497 (2015)
15. Alippi, C., Roveri, M., Trova, F.: A self-building and cluster-based cognitive fault diagnosis system for sensor networks. IEEE Trans. Neural Netw. Learn. Syst. **25**(6), 1021–1032 (2014)