Machine Learning-Driven Soft Sensor Implementation for Real-Time Fault Detection in CDU of Oil Refinery

Mothena Fakhri Shaker AlRijeb
Advanced Lightning, Power and Energy Research (ALPER), Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang, Malaysia | Faculty of Engineering, Aliraqia University, Baghdad, Iraq
mothena.f@aliraqia.edu.iq (corresponding author)

Mohammad Lutfi Othman
Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia
lutfi@upm.edu.my

Aris Ishak
Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia
ishak@upm.edu.my

Mohd Khair Hassan
Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia
khair@upm.edu.my

Baraa Munqith Albaker
Faculty of Engineering, Aliraqia University, Baghdad, Iraq
baraamalbaker@ymail.com

Received: 30 November 2024 | Revised: 1 January 2025 | Accepted: 4 January 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.9781

ABSTRACT

Soft sensors in oil refineries provide operators with important insights into the behavior and performance of processes using real-time and historical data to generate predictions. This data-driven strategy makes it easier to make wise decisions for detecting faults, thus improving process optimization and control. The Crude Distillation Unit (CDU) imposes very harsh working environments for measuring instruments, imposing both the use of a very robust sensory system and periodic maintenance procedures, which are time-consuming and costly. Notwithstanding such precautions, faults in those measuring devices, such as temperature and pressure sensors, still occur, and the presence of a sensor fault deteriorates the efficiency, productivity, and reliability of the refinery process. Recent works focused only on some fault types (e.g., bias and drift), ignoring others. This study presents the design of a soft sensor to detect all possible fault types in the real-time processing of an oil refinery. This method used actual data collected from the Salahuddin oil refinery in Iraq, several preprocessing methods, and a machine-learning approach. The proposed soft sensor was designed using several stages, including data collection, preprocessing, clustering, and classification. In the classification stage, an approach based on a Bagged Decision Tree (BDT) and Support Vector Machine (SVM) was implemented to classify the detected faults. The proposed soft sensor was trained and tested using actual data, achieving a high fault detection and classification result of 99.96%.

Keywords-oil refinery; soft sensor; machine learning; BDT; SVM

I. INTRODUCTION

Distillation columns are crucial components in various industries, particularly in chemical and petrochemical processes, where they are used to separate mixtures into individual components based on differences in boiling points. The complexity of distillation columns arises from the interconnected variables involved in the process, which include temperature, pressure, flow rates, and feed composition [1, 2]. The achievement of specific objectives, such as product purity, operational continuity, and stable operating regimes, requires careful design and control of distillation processes. However, maintaining these objectives can be challenging due to various factors, including human error and equipment failure [3, 4]. Human error is a significant contributor to industrial accidents, accounting for more than 70% of such incidents [5, 6]. Errors in the operation, monitoring, and decision-making can cause disruptions in the distillation process, compromising safety and efficiency. Moreover, failures in physical components, such as sensors and actuators, can disrupt the operation of distillation columns [5-7]. These components are crucial for monitoring and controlling the various parameters within the column. Malfunctioning sensors or actuators can lead to inaccuracies in data collection and control actions, thereby impacting the overall performance of the distillation process. In addition to physical component failures, issues with data collection and monitoring systems can also affect the reliability of the distillation columns. Failures in these systems can result in a lack of real-time data and insights into the process, making it difficult to identify and address potential issues before they escalate [8-12].

To mitigate the risks associated with human errors and equipment failures, industries often implement robust safety protocols, regular maintenance schedules, and advanced monitoring and control systems. Additionally, ongoing training and education for operators can help reduce the likelihood of errors and improve the overall reliability of the distillation processes. However, in recent years, soft sensors have been used to address these failures, without human intervention. The integration of soft sensors has emerged as a pivotal paradigm for process monitoring and control. Unlike conventional sensors, soft sensors harness computational models to estimate and predict unmeasured or difficult-to-measure process variables, thereby contributing to enhanced system observability and control accuracy. This paradigm shift holds significant promise in various industries such as chemical engineering, manufacturing, and environmental monitoring. The reliance on soft sensors is rooted in the ever-growing demand for real-time, accurate, and cost-effective monitoring solutions. As traditional sensors face limitations in terms of robustness, maintenance, and adaptability to dynamic processes, soft sensor deployment addresses these challenges by leveraging advanced algorithms and machine-learning techniques. Integration of soft sensors marks a significant step towards intelligent and adaptive monitoring systems. The results of this research not only contribute to the academic understanding of soft sensors but also hold the promise of fostering innovation and optimization in industrial processes.

In the era of Industry 4.0, where data-driven decision-making is paramount, the role of soft sensors has become increasingly indispensable. The need for reliable and adaptive monitoring tools increases as manufacturing and industrial processes become more complex. Soft sensors not only bridge the gap in instances where physical sensors fail but also exhibit a remarkable capacity to adapt to evolving process dynamics. This adaptability is facilitated by their ability to assimilate and learn from large datasets, thereby enabling a proactive response to variations and disturbances in real time. The core of soft sensor technology lies in its ability to infer unmeasured variables by exploiting correlations within existing process data. Machine learning algorithms, including neural networks, Support Vector Machines (SVM), and Bayesian networks, form the foundation of these sensors, allowing them to decipher complex patterns and relationships. This study investigates the intricacies of these algorithms, shedding light on their applicability, limitations, and potential refinements in various industrial contexts. Furthermore, the economic implications of soft sensors cannot be easily estimated. The reduced dependence on physical sensors not only reduces upfront installation costs but also alleviates the burden of maintenance and calibration. As industries navigate toward sustainability and cost-effectiveness, the integration of soft sensors aligns seamlessly with these objectives, positioning itself as a sustainable solution for efficient process monitoring.

In light of these considerations, this study contributes to the existing body of knowledge on soft sensors by exploring new approaches, validating their performance through empirical studies, and addressing the challenges that can impede their widespread adoption. By synthesizing theoretical insights with practical applications, this study aims to foster a comprehensive understanding of soft sensors and catalyze their integration into mainstream industrial practices.

The landscape of soft sensors has garnered substantial attention in the recent literature, reflecting a collective effort to harness the potential of computational models in augmenting and, in some cases, supplanting traditional sensing methodologies. Previous studies categorized fault detection and diagnosis approaches into three main categories: model-based methods, knowledge-based methods, and data-driven methods [13]. In recent years, data-driven approaches have been increasingly employed because of the complex nature of chemical processes to construct a reliable and precise mathematical model without the need for information about the process or prior expert knowledge. In [14], the focus was on improving the performance of oil refinery processes by developing a soft sensor model that predicts crude oil cuts from the initial stage of the refining process. This predictive model combines Rough Set Theory (RST) and the Adaptive Neuro-Fuzzy Inference System (ANFIS). RST was used to handle uncertain and imprecise data by identifying essential features within a dataset. Compared to traditional Proportional-Integral-Derivative (PID)-based cascade control, the findings of the suggested ANFIS-based cascade control did not overshoot and provided an improvement of 26.65% and 84.63% in the rise and settling times, respectively. In [15], soft sensor data were analyzed in agricultural settings. Historical raw data was collected for cultivation using IoT-based soft sensor modules.

The raw data were preprocessed to eliminate missing values, normalize them, and eliminate noise from the image captured by the IoT module. A Weight-Optimized Neural Network with Maximum Likelihood (WONN_ML) was used to represent the features in the processed data.

The field of soft sensors continues to evolve rapidly. This study aimed to address specific gaps in the existing literature, such as focusing only on soft sensing and monitoring systems to detect, analyze, and isolate simple types of fault in the oil industry, but to develop prototypes in a simulation environment, explore algorithmic approaches, and validate their applicability in practical industrial scenarios. The synthesis of prior work and the proposed advances aim to contribute to the ongoing dialogue regarding the efficacy and scalability of soft sensors across diverse domains. This study addresses a real-world issue by proposing a soft sensor model with raw data from actual refinery operations. The model can be built to perform efficiently with inconsistent and erroneous data, increasing its precision and dependability in real-world situations. Emphasis on model interpretability is an important component of the proposed work. The goal is to clarify the model's decision-making process, establish trust among refinery employees, and enable well-informed decision-making by making the soft sensor work instead of the failure sensors until the problem is solved.

II. THE PROPOSED APPROACH

The proposed method was designed to detect all types of faults, which is a main challenge. It includes several main stages that work together to build the desired soft-sensor model. These stages are data collection, preprocessing, clustering, and classification. Figure 1 shows the main stages of the proposed method.

A. Data Collection

Real data were collected from the Salahuddin oil refinery in Iraq, as shown in Figure 2. The data was collected for seven months, from January 1, 2023, to July 31, 2023. These data were collected in 10 s intervals from the daily report of the unit's activities. Subsequently, the data were validated and exported from an Excel spreadsheet to a CSV file. The datasheet contains several types of data: temperature, pressure, data flow, Set Point (SP), Control Valve (CV), and Actual Value. The most effective values considered were Temperature (TE), and Pressure (PV) which are correlated with the SP using Pearson Correlation Coefficient Analysis (PCCA) to make the desired decisions for the CV [16]. The total number of collected data was 20 million. Table I presents a sample of collected data.

TABLE I. SAMPLE OF DATA COLLECTED FROM THE SALAHUDDIN OIL REFINERY

TE	PV	SP	CV
47.88	67.05	66	15.15
202.09	73.97	67	100
205.49	84.82	84.82	0
20.26	8.29	8.29	100
240.49	259.35	260	63.45
255.47	199.95	200	62.66
20.84	0	0	100

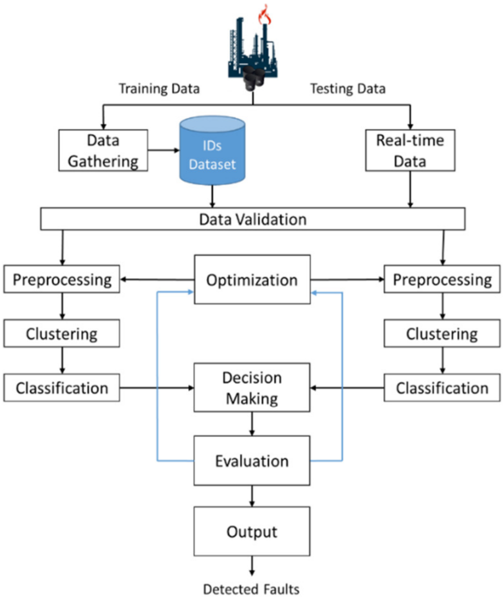

Fig. 1. The proposed approach.

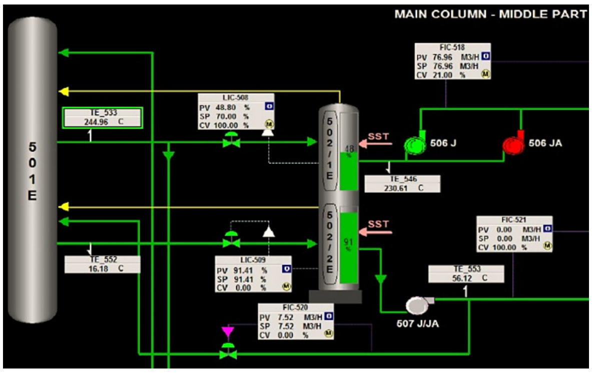
MAIN COLUMN - MIDDLE PART
501E
TE_533 244.96 C
LIC-508 PV 48.80 % SP 70.00 % CV 100.00 %
FIC-518 PV 76.96 M3/H SP 76.96 M3/H CV 21.00 %
502/1E
502/2E
SST
SST
506 J
506 JA
TE_546 230.61 C
FIC-521 PV 0.00 M3/H SP 0.00 M3/H CV 100.00 %
TE_553 56.12 C
TE_552 16.18 C
LIC-509 PV 91.41 % SP 91.41 % CV 0.00 %
FIC-520 PV 7.52 M3/H SP 7.52 M3/H CV 0.00 %
507 J/JA

Fig. 2. Salahuddin oil refinery.

The most effective values considered were temperature and pressure, which were correlated with SP using PCCA in [16]:

$$r(s_i, s_j) = \frac{Cov\,(s_i, s_j)}{\sqrt{Var(s_i)\,.Var(s_j)}} \tag{1}$$

where r is the Pearson correlation coefficient, $Var(s_i)$ and $Var(s_j)$ are respectively the mean values of the two variables, s_i denotes the individual values of one variable, and s_j denotes the individual values of the other variable. Table II lists the correlation coefficients for all sensors. The relevance between sensors is obvious due to cross-sensitivity. PCCA numbers between –1 and 1 measure the strength and direction of the relationship between two variables.

TABLE II. DATA CORRELATION

Correlation coefficient	TE	PV	SP	CV
TE	1	0.8227	0.8137	-0.2345
PV	0.8227	1	0.9996	-0.2281
SP	0.8137	0.9996	1	-0.2353
CV	-0.2345	-0.2281	-0.2353	1

B. Data Preprocessing

The second stage involved data preprocessing. The quality of historical data has a direct impact on the performance of soft sensors. Data collection in industrial processes can be hampered by issues such as missing data, sample time, and outliers. As a result, the data cannot be used for soft-sensor modeling. Thus, several steps were considered to address the aforementioned issues and prepare the data for the next stages. These steps included normalization, missing data imputation, outlier data, and data reduction.

1) Normalization

Normalization is used in the preprocessing step to remove significant imbalances between the features and balance their impact on the machine learning algorithm computations. The proposed method employs the min-max normalization approach. Each output data point has a value in the range of [0, 1]. Min-max normalization was performed using [17]:

$$X_{Scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (2)$$

where X_{Scaled} is the normalized value, X is the original value, X_{max} the maximum value of X, and X_{min} is the minimum value of X.

TABLE III. DATA NORMALIZATION

TE	PV	SP	CV
0.1174	0.2585	0.2538	0.1515
0.7730	0.2852	0.2576	1
0.7875	0.3270	0.3262	0
0	0.0319	0.0318	1
0.9363	1	1	0.6345
1	0.7709	0.7692	0.6266
0.0024	0	0	1

The collected data values' ties were maintained using min-max normalization. Smaller standard deviations in the data resulting from this constrained range can reduce the impact of outliers.

2) Missing Data Imputation

Missing data refers to the case when a variable in the data has no value recorded. This study used mean substitution [18], which preserves the sample mean for the variable by substituting the variable mean for any missing values.

TABLE IV. BEFORE DATA IMPUTATION

PV	CV
0.7970	0.3500
0.5812	**NaN**
NaN	0.3500
0.6470	0.3500

TABLE V. AFTER DATA IMPUTATION

PV	CV
0.7970	0.3500
0.5812	**0.3500**
0.6141	0.3500
0.6470	0.3500

3) Outliers Removal

An outlier is any observation that is abnormally distant from any other value. Outliers present a challenge for several statistical examinations, because they may either misrepresent the actual results or overlook important findings. This study used the K-Nearest Neighbor (KNN) to detect and remove outliers according to their distance from the other data, and the nearest non-outlier value was used to replace the outliers removed [19]. In Figure 3, the outlier is the highest point with a value of approximately 150, but the other values are between 115-120. Figure 4 shows the effect of removing the outlier by putting the data in the range 115-120, which is the normal range of the original data.

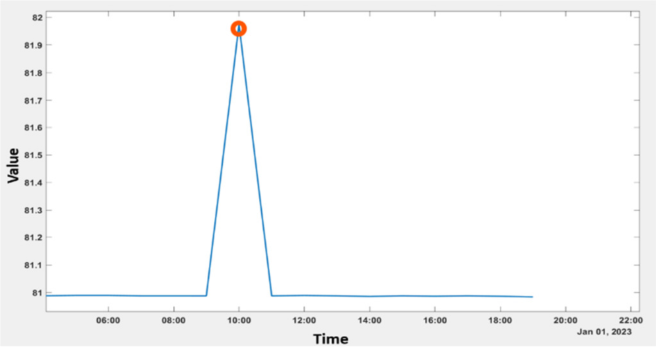

Fig. 3. Data outlier.

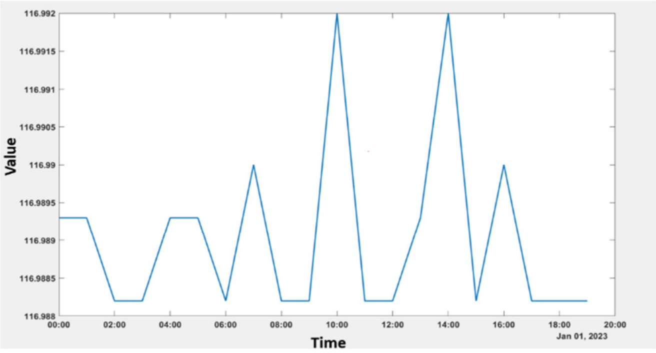

Fig. 4. Applying outliers' removal.

4) Data Reduction

Data reduction is the process of reducing some data elements. If the data are multidimensional, reduction may also occur in other areas, such as data dimensionality. Any data reduction typically results in a reduction in the data volume. Each time a request is made to access a redundant segment, it maintains only a single copy of that segment during storage. By combining a large number of variables into a smaller one, most of the information in the larger set is retained. Principal Component Analysis (PCA) is a dimensionality reduction method commonly used to reduce the dimensionality of large datasets [20]. Figure 5 shows a sample of the results of applying PCA to reduce the dimensions of the data.

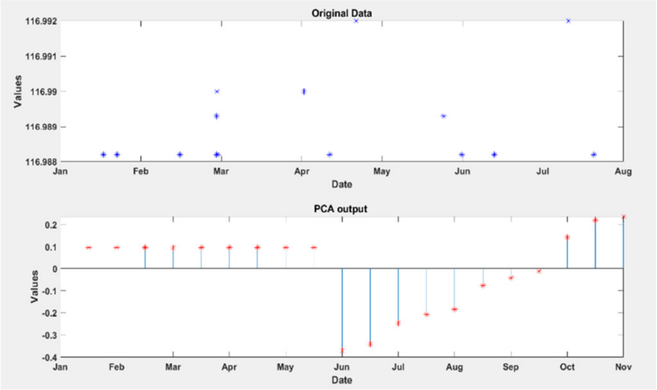

Fig. 5. Data reduction.

C. *Data Clustering*

Data clustering aims to divide a set of data into a certain number of groups or clusters that are best fit by a predetermined criterion function. Data clustering is an essential and supportive tool with a multitude of applications. Fuzzy C-Means Clustering (FCM) [21] was used to cluster the data into groups according to their characteristics. The output from this stage is then passed to the classification stage.

D. *Classification*

Classification is a method of separating and organizing data into relevant groups (classes) based on specific criteria. In this stage, a soft sensor was built using machine learning methods. A classification approach was proposed to build an efficient soft sensor based on a Bagged Decision Tree (BDT) using several decision trees [22] and an SVM with a polynomial kernel [23]. The input data were split into many sets that overlapped and then fed into the classification module. Several classification results were obtained from the proposed approach and then a voting function using SVM to select the most frequent results to consider as the final result. Figure 6 illustrates the proposed classification approach.

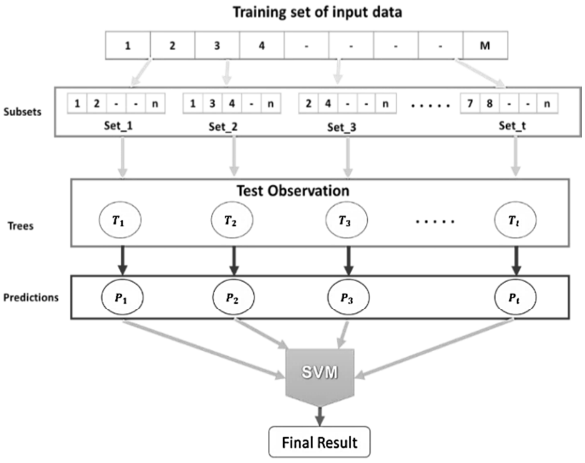

Fig. 6. Proposed classification approach.

The proposed approach classifies the data as normal or abnormal (fault) data. For normal data, this implies that the process is stable; otherwise, an error is considered. Each type of error has a unique data corruption response or a set of properties. Typical types of faults regarding operational circumstances and sensor malfunctions include the following:

- Bias: There is a continuous shift in the values received compared with the right value.
- Drift: Error levels in the data either increase or decrease with time.
- Precision Degradation (PD): Over time, the sensor plates may become worn down or unclean, which might cause inaccuracies in the received data from the sensors that resemble random noise around the normal values.
- Failure: The data obtained might be entirely random or continuous due to sensor failure or measurement limits.

Figure 7 shows all the fault types, where the blue line represents the normal data and the yellow stars represent the faults.

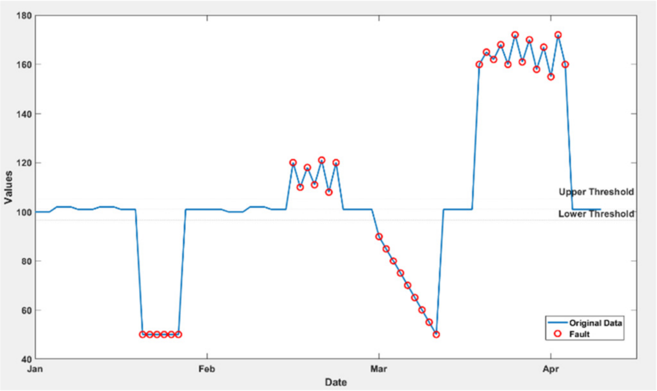

Fig. 7. Various types of gross errors.

E. *Decision Making*

After classifying the data using the proposed approach, a decision-making stage followed. The soft sensor maintains the occurrence of issues until they are fixed. The soft sensor works as a backup sensor to avoid interrupting the process based on the reference data obtained from the original collected data and addresses all possible types of fault that may occur. After this stage, an evaluation process was performed to evaluate the soft sensor output and then use it to optimize the incoming real-time data, thus obtaining better results. Algorithm 1 describes the main steps of the proposed method.

```
Algorithm 1: The proposed method
Input: IDs dataset
Output: Detected faults
Load the IDs dataset
Apply validation process on the input data
Split the data into training and testing
For all input data
  Perform preprocessing
```

```
    Normalization
    Missing Data Imputation
    Outliers Removal
    Data Reduction
Apply the clustering process using FCM
Apply the proposed approach for
classification BDT-SVM
Identify the fault using the decision-
making process
Return the detected fault type
```

III. RESULTS AND DISCUSSIONS

The proposed method was implemented using MATLAB 2023b in Windows 11. Several methods are used to evaluate the performance of soft sensors. The input dataset was divided into two sets: 70% for training and 30% for testing. Accuracy (3) was used to obtain system results. The input data was read by the system and the preprocessing steps were applied first.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

where TP denotes the true positives, TN denotes the true negatives, FP denotes the false positives, and FP denotes the false negatives. Table VI shows the effect of applying the preprocessing stage, indicating an improvement in accuracy by 3.6%.

TABLE VI. ACCURACY RESULTS BEFORE AND AFTER APPLYING THE PRE-PROCESSING

	Accuracy
Without preprocessing	96.33%
With preprocessing	99.96%

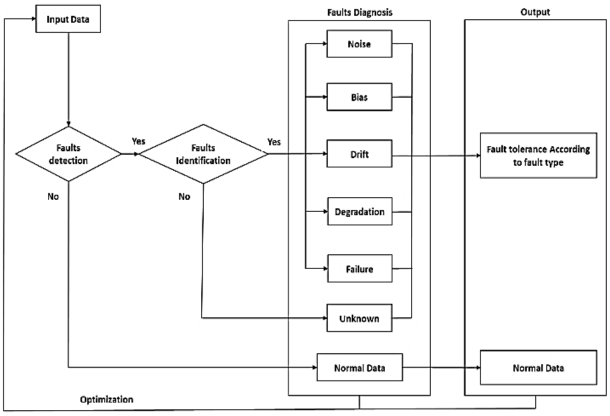

Fig. 8. Proposed soft sensor.

Applying data clustering enhances the results by grouping the data into several sets. These sets had the same characteristics and improved classification results. In the proposed method, the best clustering results were obtained by the Fuzzy C-Means (FCM) method, as shown in Figure 9.

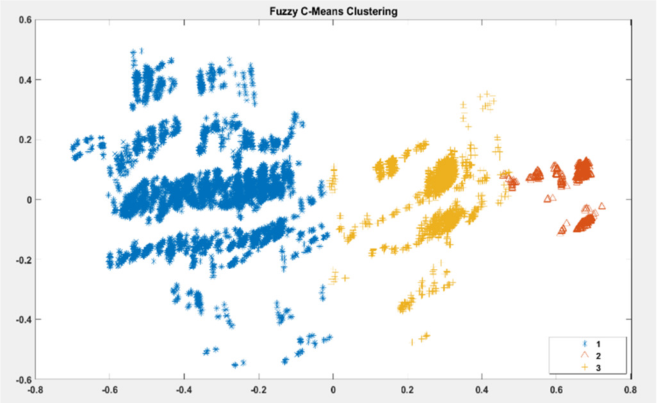

Fig. 9. Fuzzy C-means clustering.

Several methods have been used for classification in machine learning. Different machine learning classifiers were tested, and Table VII shows their results.

TABLE VII. CLASSIFICATION ACCURACY RESULTS

Classifier	**Accuracy**
KNN	94.7%
SVM	96.3%
DT	92%
BDT	97.2%
BDT-SVM	99.96%

The proposed approach (BDT-SVM) achieved 99.96% classification accuracy, which was higher than that of the other classifiers, offering an improvement in the obtained results by 5.26% compared to KNN and 7.96% compared to DT. In addition, the proposed approach achieved better results than the original BDT and SVM by 2.76% and 3.66%, respectively. Three SVM kernels were tested for better classification results, with the polynomial kernel achieving the best results, as shown in Table VIII.

TABLE VIII. ACCURACY RESULTS OF VARIOUS SVM KERNELS

SVM Kernel	**Accuracy**
Linear	98.6%
RBF	99%
Polynomial	99.96%

Moreover, the proposed classification approach achieved a minimum classification error of 0.2, which was eliminated after several training iterations using 50 trees, as shown in Figure 10. Figure 11 shows a sample of the detected faults after the classification stage. The faults (represented by yellow stars) are located above and below the original data represented by the blue line. After detecting the faults in the previous stage, the decision-making stage takes appropriate action to solve these faults. This stage functions as an output of the soft sensor to fix the possible faults that occur during the refinery process. The results of applying the decision-making stage are shown in Figure 12.

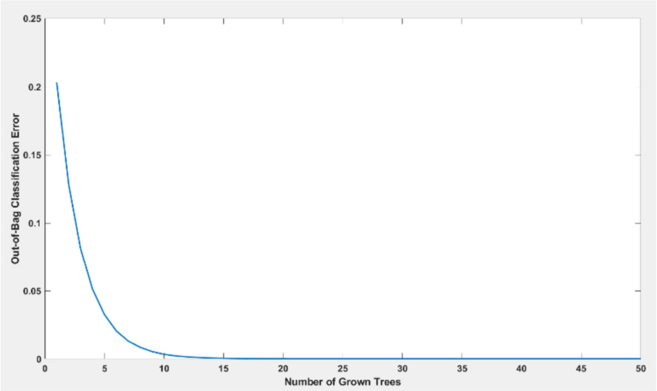

Fig. 10. Classification error.

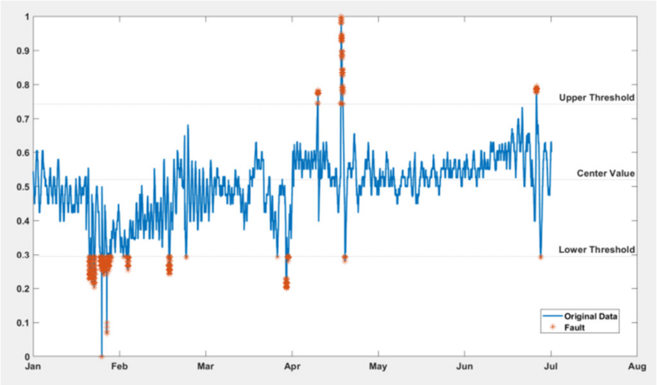

Fig. 11. Detected faults.

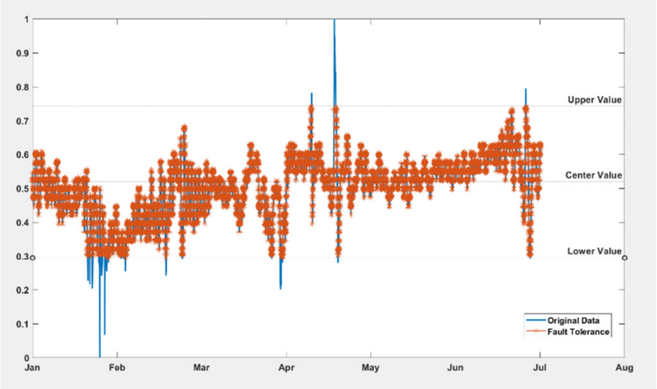

Fig. 12. Fault tolerance.

IV. CONCLUSION

This study proposed an approach for efficient soft sensors based on machine learning for real-time data obtained from the Salahuddin oil refinery in Iraq. Soft sensors can be deployed within the refinery control system or Supervisory Control and Data Acquisition (SCADA) systems to provide real-time predictions of process variables. Integration with existing control systems allows operators to effectively monitor and control refinery operations. The proposed method achieved high fault detection and classification results. The preprocessing stage prepared the data for the next stage via several steps, which made the data representation easy to deal with. The FCM clustering method was used to enhance the fault detection process by isolating normal data from other data containing errors. The proposed method involves a hybrid of BDT and SVM classifiers, testing all dataset attributes using BDT and then performing a voting process using SVM to select the best classification result. In the future, the proposed soft sensor can be employed to deal with other types of data.

ACKNOWLEDGMENT

The authors would like to thank Universiti Putra Malaysia (UPM), the Malaysian Ministry of Higher Education (MOHE), Al-Iraqia University, and the Iraqi Ministry of Higher Education and Scientific Research for their continuous support in this research. This work was supported in part by North Refineries Company and ALNUHA for General Contracts & Industrial Equipments Trading, Ltd.

REFERENCES

[1] A. S. Yamashita, A. C. Zanin, and D. Odloak, "Tuning the Model Predictive Control of a Crude Distillation Unit," *ISA Transactions*, vol. 60, pp. 178–190, Jan. 2016, https://doi.org/10.1016/j.isatra.2015.10.017.

[2] "Operating manual of Al Doura oil refinery," Aldoura Oil Refinery, Baghdad, Iraq, Technical Report, 2010.

[3] A. Raimondi, A. Favela-Contreras, F. Beltrán-Carbajal, A. Piñón-Rubio, and J. L. De La Peña-Elizondo, "Design of an adaptive predictive control strategy for crude oil atmospheric distillation process," *Control Engineering Practice*, vol. 34, pp. 39–48, Jan. 2015, https://doi.org/10.1016/j.conengprac.2014.09.014.

[4] V. T. Minh and A. M. Abdul Rani, "Modeling and Control of Distillation Column in a Petroleum Process," *Mathematical Problems in Engineering*, vol. 2009, no. 1, Jan. 2009, Art. no. 404702, https://doi.org/10.1155/2009/404702.

[5] T. Takahama and D. Akasaka, "Model Predictive Control Approach to Design Practical Adaptive Cruise Control for Traffic Jam," *International Journal of Automotive Engineering*, vol. 9, no. 3, pp. 99–104, 2018, https://doi.org/10.20485/jsaeijae.9.3_99.

[6] S. Kemaloğlu, E. Ö. Kuzu, D. Gökçe, and Ö. Çetin, "Model predictive control of a crude distillation unit," *IFAC Proceedings Volumes*, vol. 42, no. 11, pp. 880–885, 2009, https://doi.org/10.3182/20090712-4-TR-2008.00144.

[7] B. Shi, X. Yang, and L. Yan, "Optimization of a crude distillation unit using a combination of wavelet neural network and line-up competition algorithm," *Chinese Journal of Chemical Engineering*, vol. 25, no. 8, pp. 1013–1021, Aug. 2017, https://doi.org/10.1016/j.cjche.2017.03.035.

[8] L. Fortyna, S. Graziani, A. Rizzo, and G. Maria, *Soft Sensors for Monitoring and Control of Industrial Processes*. London, UK: Springer London, 2007.

[9] S. M. Jafari, M. Ganje, D. Dehnad, and V. Ghanbari, "Mathematical, Fuzzy Logic and Artificial Neural Network Modeling Techniques to Predict Drying Kinetics of Onion: Comparison of Modeling Techniques for Onion Drying," *Journal of Food Processing and Preservation*, vol. 40, no. 2, pp. 329–339, Apr. 2016, https://doi.org/10.1111/jfpp.12610.

[10] B. Bidar, J. Sadeghi, F. Shahraki, and M. M. Khalilipour, "Data-driven soft sensor approach for online quality prediction using state dependent parameter models," *Chemometrics and Intelligent Laboratory Systems*, vol. 162, pp. 130–141, Mar. 2017, https://doi.org/10.1016/j.chemolab.2017.01.004.

[11] C. Martin, H. Zhang, J. Costacurta, M. Nica, and A. R. Stinchcombe, "Solving Elliptic Equations with Brownian Motion: Bias Reduction and Temporal Difference Learning," *Methodology and Computing in Applied Probability*, vol. 24, no. 3, pp. 1603–1626, Sep. 2022, https://doi.org/10.1007/s11009-021-09871-9.

[12] C. A. Duchanoy, M. A. Moreno-Armendáriz, L. Urbina, C. A. Cruz-Villar, H. Calvo, and J. De J. Rubio, "A novel recurrent neural network soft sensor via a differential evolution training algorithm for the tire

contact patch," *Neurocomputing*, vol. 235, pp. 71–82, Apr. 2017, https://doi.org/10.1016/j.neucom.2016.12.060.

[13] P. Ilius, M. Almuhaini, M. Javaid, and M. Abido, "A Machine Learning–Based Approach for Fault Detection in Power Systems," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11216–11221, Aug. 2023, https://doi.org/10.48084/etasr.5995.

[14] A. H. H. Al Jlibawi, M. L. Othman, A. Ishak, B. S. Moh Noor, and A. H. M. S. Sajitt, "Optimization of Distribution Control System in Oil Refinery by Applying Hybrid Machine Learning Techniques," *IEEE Access*, vol. 10, pp. 3890–3903, 2022, https://doi.org/10.1109/ACCESS.2021.3134931.

[15] A. Wongchai, S. K. Shukla, M. A. Ahmed, U. Sakthi, M. Jagdish, and R. Kumar, "Artificial intelligence - enabled soft sensor and internet of things for sustainable agriculture using ensemble deep learning architecture," *Computers and Electrical Engineering*, vol. 102, Sep. 2022, Art. no. 108128, https://doi.org/10.1016/j.compeleceng.2022.108128.

[16] R. F. Tate, "Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[17] J. Pan, Y. Zhuang, and S. Fong, "The Impact of Data Normalization on Stock Market Prediction: Using SVM and Technical Indicators," in *Soft Computing in Data Science*, Singapore, 2016, pp. 72–88, https://doi.org/10.1007/978-981-10-2777-2_7.

[18] J. Zhu, Z. Ge, Z. Song, and F. Gao, "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annual Reviews in Control*, vol. 46, pp. 107–133, Jan. 2018, https://doi.org/10.1016/j.arcontrol.2018.09.003.

[19] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007, https://doi.org/10.1016/j.patcog.2006.12.019.

[20] L. F. A. Napier and C. Aldrich, "An IsaMill^{TM} Soft Sensor based on Random Forests and Principal Component Analysis," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1175–1180, Jul. 2017, https://doi.org/10.1016/j.ifacol.2017.08.270.

[21] U. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific databases," in *Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150)*, Olympia, WA, USA, 1997, pp. 2–11, https://doi.org/10.1109/SSDM.1997.621141.

[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, https://doi.org/10.1007/BF00994018.