Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v10. n.50, p. 17-27. November/ December., 2024.

DOI: https://doi.org/10.5935/jetia.v10i50.1289

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DETECTION OF SUBSTATION POLLUTION IN DISTRICT HEATING AND COOLING SYSTEMS: A COMPREHENSIVE COMPARATIVE ANALYSIS OF MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORK MODELS

[1] Emrah Aslan and [2] Yıldırım Özüpak

[1] Emrah ASLAN - Computer Programming Department, Silvan Vocational School, Dicle University, Diyarbakır, Turkey.
[2] Department of Electricity and Energy, Silvan Vocational School, Dicle University, Diyarbakır, Turkey.

[1]http://orcid.org/0000-0002-0181-3658, [2]http://orcid.org/0000-0001-8467-8702

Email: emrah.aslan@dicle.edu.tr, yildirim.ozupak@dicle.edu.tr

ARTICLE INFO

Article History
Received: September 9, 2024
Revised: October 16, 2024
Accepted: November 1, 2024
Published: November 30, 2024

Keywords:
Pollution Detection,
Grid Search Optimization,
Machine Learning,
CNN,
DHC.

ABSTRACT

This study analyzes the detection of substation fouling failures in District Heating and Cooling (DHC) systems using synthetic data. In the study, high, medium and low levels of contamination are considered and both machine learning and deep learning techniques are applied for the detection of these failure types. Within the scope of the analysis, machine learning algorithms such as K-Nearest Neighbors, XGBoost and AdaBoost are compared with the proposed Convolutional Neural Network (CNN) model. The machine learning algorithms and the Convolutional Neural Network model are trained to perform fault detection at different contamination levels. In order to improve the performance of the machine learning models, hyperparameter tuning was performed by Grid Search Optimization method. The results obtained show that the proposed Convolutional Neural Network model provides higher accuracy and overall success compared to machine learning methods. High performance measures such as Matthews correlation coefficient 0.944 and accuracy rate 0.972 were achieved with the CNN model. These findings reveal that contamination detection in substations can be done effectively with CNN-based approaches, especially for situations that require high accuracy. This study on fault detection in DHC systems provides a new and reliable solution for industrial applications.

I. INTRODUCTION

District Heating and Cooling (DHC) systems are a critical infrastructure component in modern urban energy management, providing efficient heating and cooling services to both residential and commercial buildings. These systems play an important role in improving energy efficiency and reducing operating costs. However, the reliability and performance of DHC systems can be affected by various failures, which can put the efficiency, safety and overall operational stability of the system at risk.

The development of effective fault detection and diagnosis (FDD) models for DHC systems is an important research area to ensure uninterrupted operation of the system and improve its reliability. Traditional fault detection methods are often based on manual checking and heuristic approaches, but these methods may not be sufficient to address the complexities of modern DHC systems. Recent advances in data-driven techniques, especially in the fields of machine learning (ML) and data analytics, offer promising solutions to these challenges. The lack of comprehensive datasets is a significant barrier to developing robust data-driven models. Existing studies emphasize that the lack of quality datasets limits the effectiveness of data-driven approaches [1]. To overcome this problem, researchers have focused on creating synthetic data sets through simulation or using open data sources. These data sets usually cover various system components such as generation units, distribution networks and storage facilities.

Numerous machine learning approaches for DHC system defect detection have been studied in recent research. For instance, because of its propensity to handle intricate patterns in data, XGBoost, Support Vector Machine (SVM), and Logistic Regression are often employed [2]. Studies have shown that these models can detect energy efficiency related faults with high accuracy, but there can be difficulties in detecting more subtle problems, e.g. thermal losses.

To improve the reliability of fault detection models, researchers are investigating integrating real-time data with simulation results and open data. This approach aims to increase the generalizability and robustness of the models, thus ensuring their applicability to a wide range of operational scenarios [3].

District Heating and Cooling (DHC) systems are increasingly being used to improve energy efficiency and reduce carbon emissions. DHC systems eliminate the need for individual heating or cooling systems by transmitting heat and cooling energy generated from a centralized source to buildings over a wide network. These systems play an important role in energy saving and environmental sustainability, especially in urban areas [4].

However, ensuring the long-term efficient operation of DHC systems poses a major challenge in terms of their maintenance and early detection of potential failures. Failures in substations, such as contamination, reduce the overall efficiency of the system and increase costs. In the past, the detection of such faults was mostly limited to physical checks or manual interventions in case certain critical limits were exceeded. Today, however, technological advances such as data analytics and machine learning offer the possibility to make these processes more automated and accurate [5].

In the past, fault detection processes in DHC systems were generally handled with a reactive approach. Physical inspections, periodic maintenance work and performance monitoring systems are among the classical methods that are activated after failures occur. These approaches are often time-consuming, lead to delays in fault detection and negatively affect the efficiency of the system. Furthermore, most of these methods are activated when there is a significant degradation in system performance, which often results in more costly repairs and system downtime. However, with the development of data collection and analysis techniques in recent years, methods such as machine learning and deep learning offer an important alternative for fault detection. By analyzing large amounts of data, these new methods provide the opportunity to detect signs and trends before failures occur [6]. Thus, early diagnosis of failures and implementation of preventive maintenance strategies become possible.

Machine learning and deep learning methods offer many advantages over traditional fault detection techniques [7]. Machine learning algorithms have achieved significant success in fault prediction by learning meaningful patterns from large datasets. Algorithms such as K-Nearest Neighbors (KNN), XGBoost and AdaBoost have the ability to automate fault prediction by training on data collected in the past. These algorithms have been successful in predicting system failures by analyzing correlations in the data and possible signs of failure. However, these methods are usually dependent on a more limited data structure and may be inadequate for complex or multidimensional data [8]. In contrast, deep learning methods, especially models such as Convolutional Neural Network (CNN), stand out with their capacity to process more complex data structures. CNN models provide higher success rates, especially in large and complex data sets, enabling the detection of previously undetected faults.

In this study, the failure conditions caused by contamination of substations in DHC systems are analyzed. Different machine learning algorithms (KNN, XGBoost, AdaBoost) and deep learning (CNN) models are compared using synthetic data for high, medium and low levels of contamination. Grid Search Optimization method was used to optimize the performance of the models and the best hyperparameters were selected. The results show that the CNN model outperforms the other models and achieves high accuracy rates.

The aim of this study is to develop a more accurate and efficient solution that goes beyond traditional methods for early detection of faults in DHC systems. In particular, it is aimed to detect complex failure types such as contamination of substations more effectively. The main advantage of the study is that higher accuracy rates are achieved with the CNN model and fault detection can be done at an earlier stage. This increases the overall efficiency of the system and reduces maintenance costs. However, the fact that deep learning models require large amounts of data and processing power is considered as a significant disadvantage of the study. Moreover, experiments with synthetic data need to be validated with real-world data, which is also one of the limitations of the study.

The study makes several important contributions to the literature. First, it demonstrates the applicability of deep learning methods for fault detection in DHC systems. Second, it proposes a solution for early detection of more complex fault types such as substation contamination. Finally, the high accuracy rates of the model proposed in the study provide a practical benefit for industrial applications. These findings not only contribute to the development of more efficient maintenance strategies in DHC systems, but also provide a new perspective on how deep learning algorithms can be applied in industrial processes.

Section 1 of this paper gives an overview of the problem. The following sections of the paper are as follows. Literature review related to the study is presented in Section 2. The implementation materials and methods are presented in Section 3, and the discussion and conclusions are presented in Section 4. Future work and conclusions are presented in Section 5.

II. THEORETICAL REFERENCE

In this study, problem diagnosis and detection techniques for District Heating and Cooling (DHC) systems are investigated. Using the IEA DHC Annex XIII as a framework, it offers a thorough study of typical DHC system faults. District heating systems have evolved from steam-based systems to ultra-low temperature networks, with future designs integrating distributed low-temperature sources and building-side heat pumps. A case study shows that while ultra-low temperature ring networks are 23% more expensive than 3rd generation systems, they are cost-effective when free waste heat is available [9].

France aims to decarbonize heating by expanding District Heating and Cooling Networks (DHCN) and identifying suitable areas for development. Using variables such as energy density, building age, and energy mix, the paper estimates the potential of DHN at 132 TWh/year and DCN at 7.8 TWh/year across France [10].

Decarbonizing energy sectors, especially heating, is crucial to combating climate change, as heating represents nearly half of global energy consumption. This paper reviews the role of heat pumps in reducing emissions and adding flexibility to renewable energy systems, but highlights economic, regulatory, and infrastructural challenges to their widespread adoption [11].

The article reviews 40 European thermal networks using distributed heat pumps and clarifies definitions of Fifth-Generation District Heating and Cooling (5GDHC). It finds that while 5GDHC systems in countries like Germany and Switzerland excel with renewable heat sources, they face higher pumping energy demands and variable control strategies compared to traditional district heating [12].

Even though the majority of research concur on a core set of typical failures, they frequently don't result in instant shutdowns and may thus go unreported. Although air-source chillers have

received a lot of attention, DHC systems are also susceptible to identical issues that have been seen in water-source systems. This study looks on ways to diagnose and find faults in District Heating and Cooling (DHC) systems.

District heating and cooling systems have evolved over a century, serving over 70 million people in Europe with an estimated energy consumption of over 450 TWh. This paper reviews the current state of these systems, highlighting variations across countries in terms of technology, market structure, and regulations, and introduces a new socio-demographic approach to create indicators for modeling future systems [13].

Fifth-generation district heating and cooling (5GDHC) systems, tested in Melbourne, show 9-29% cost savings and 25-58% GHG emissions reduction compared to traditional systems. They offer economic and environmental benefits, especially in mild climates, and could expand to other regions with similar conditions [14].

Temperature control significantly impacts Fifth Generation District Heating and Cooling (5GDHC) systems. This study finds that constant temperature control can be more effective but is sensitive to setting changes, while simple multi-stage controls may underperform. Proper strategy selection and coordination are crucial for optimal system efficiency [15].

District heating and cooling networks offer benefits by integrating renewable energies and local thermal resources, but effective design and optimization are key. This review evaluates the use of Life Cycle Assessment (LCA) for assessing the environmental impact of these networks, revealing a wide range of emission factors. It emphasizes the need for improved management practices and proposes future research for developing a universal LCA tool for network analysis [16].

Due to the limited availability of real-world data, the development of optimized synthetic data sets is investigated to improve the accuracy of three different ML models such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). The integration of real and synthetic data improved the identification of initial faults using ML algorithms and the quality of the synthetic data obtained was found to be superior to existing methods [17].

A simulation-based dataset was developed to evaluate various types of failures in District Heating and Cooling (DHC) systems and tested with five machine learning models. The tests showed that the dataset provides high performance in fault detection and the applicability of the models to real systems [18].

Traditional and innovative methods used to achieve goals such as peak shaving, demand response and fast fault detection in advanced district heating and cooling systems are examined. The advantages and disadvantages of modern approaches such as model predictive control and machine learning are detailed [19].

A machine learning method is proposed to detect leakage faults with data from flow and pressure sensors. Using a delayed warning algorithm, a leak signal is sent and the model identifies the faulty pipe based on this signal. The method was successfully tested with 85.85% accuracy and a macro-F1 score of 0.99786 [20].

A machine learning model has been developed that performs well and can distinguish between data sets containing faults. The model is trained with data from a district heating substation in Sweden and tested with various parameters. The results show that the model successfully models the substation behavior and has high fault detection capability [21].

A study was conducted to develop an automatic fault detection and diagnosis (AFDD) framework for district heating substations. With data from Denmark, common faults are analyzed and the potential of AFDD to reduce energy use is highlighted. Additional indicators were proposed and improvements were made to detect future anomalies [22]. A study was conducted to show how smart meter data can be used for fault detection and maintenance processes. Faults were detected in advance with machine learning algorithms and maintenance improvements were provided with performance indicators. The findings were validated by experts and the importance of data utilization for smart heating networks was emphasized [23].

A two-level model was developed for fault detection in district heating systems. This model, which distinguishes high-level system faults and low-level sub-faults, provided high accuracy and reliability in tests. The results show that the model provides an effective solution for real-time monitoring [24]. In [25], a machine learning method is proposed to detect leakage faults with flow and pressure data. The delayed warning algorithm and the model ensure accurate identification of leaks. The method was successfully tested with 85.85% accuracy and a macro-F1 score of 0.99786.

III. MATERIALS AND METHODS

III.1 DATASET

This dataset contains synthetic fault data related to contamination in substations of District Heating and Cooling (DHC) systems. The dataset was developed as part of the International Energy Agency's (IEA) DHC Annex XIII project “Artificial Intelligence Fault Detection and Prediction of Heat Production and Demand in District Heating Networks”. This project develops artificial intelligence methods for the prediction of heat demand and production and evaluates algorithms for fault detection. The experiments in the dataset are simulations covering a period of 28 days, during which faults that can occur at various time points are observed [18]. Failures can occur with different intensities, either sudden or gradual. The failure intensity can be interpreted differently depending on the simulation model used. This dataset provides a valuable resource for the development of new approaches to fault detection in DHC systems. Table 1 lists the input names and basic statistics of the dataset [26].

Table 1: Basic statistical information about the data used in the dataset.

Variable	Explanation	Min	Max
BC	Id of the boundary conditions used for this experimen	0	12
F1_type	Type of fault used in the experiment	0	1
F1_start	Start time of the fault, in hours	0	671
F1_stop	Stop time of the fault, in hours	0	672
F1_init	İnitial intensity of the fault, in the range [0-1]	0	1
F1_final	Final intensity of the fault, in the range [0-1]	0	1

Source: [26].

III.2 K-NEAREST NEIGHBOR REGRESSION

K-Nearest Neighbor (KNN) regression is one of the supervised learning methods and is based on the idea of predicting a target data point by looking at the values of its neighboring data points.

KNN regression has a similar approach to the classification problem, but focuses on the prediction of a continuous target variable. The predicted value for a data point is determined by the average or weighted average of the target values of the selected "k" number of nearest neighbors. Since KNN regression is a parameter-free method, the structure of the model is shaped according to the data set and provides flexibility, especially for complex data distributions. However, the increase in computational costs for high-dimensional data sets and the impact of distance metrics on the performance of neighbor selection are important factors to be considered [27].

III.3 EXTREME GRADIENT BOOSTING REGRESSION

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm often used in supervised learning problems such as regression and classification. XGBoost is an ensemble learning method based on decision trees and minimizes errors by successively adding weak learners (usually decision trees). In regression problems, the main goal of XGBoost is to linearly improve the model's predictions as much as possible and minimize the loss function. The algorithm's innovations on gradient boosting include optimizations that make tree structures faster and more efficient, regularization (L1 and L2) and missing data management. This allows XGBoost to achieve high performance on large datasets while providing superior results in terms of speed and accuracy compared to other methods. The fact that it does not require as much computational power as deep learning techniques and its nonparametric structure have made it a frequently preferred method in regression analysis [28].

III.4 ADAPTIVE BOOSTING REGRESSION

Adaboost (Adaptive Boosting) is an ensemble learning method used to improve the performance of weak learners. Basically, by emphasizing the errors of each weak learner, weight is given to the next learner. In the regression problem, Adaboost successively combines weak regression models to minimize prediction errors. At each step, different weights are assigned to the samples, taking into account the model's previous prediction errors. This weighting allows the model to focus specifically on data points that have struggled with previous predictions. Adaboost regression is often applied with weak learners such as decision trees and is known for providing high accuracy in regression as well as in classification problems. However, it can be sensitive to overfitting and noisy data, so it requires careful model selection and parameter tuning. Adaboost's strength is that it can successfully combine weak learners to improve the overall performance of the model [29].

III.5 CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are deep learning architectures structurally inspired by the visual cortex of the human brain. They are particularly effective in computer vision tasks such as image recognition, object detection and classification.

The basic building blocks of CNNs are convolution layers that focus on capturing local features in the data, thereby reducing the number of parameters of the model through dimensionality reduction. These layers extract meaningful features from the input data through filters, and these features are abstracted in deeper layers, allowing more complex structures to be learned. CNNs offer high performance and accuracy, especially on large data sets, thanks to parameter sharing and local connections. Their advanced architecture and level of performance have made CNNs a popular solution in various fields such as medical imaging, autonomous vehicles, and face recognition [30].

III.6 GRID SEARCH OPTIMIZATION

Grid Search Optimization is a method of hyperparameter tuning used to improve the performance of machine learning models. Machine learning algorithms are structured with various hyperparameters that affect the behavior of the model, and choosing the optimal values of these parameters greatly affects the accuracy and overall performance of the model.

The Grid Search method aims to systematically scan all possible combinations of the specified hyperparameters and select the combination that gives the best result. This is accomplished by evaluating the model against a specified performance metric. Although Grid Search is usually used for smaller datasets and in environments with limited computational power, the processing time can increase considerably with large datasets or a large number of hyperparameters. In this case, the Grid Search method can become costly in terms of time and resources. Nevertheless, by finding the optimal parameters, the generalizability and accuracy of the model can be improved. Due to these features, Grid Search Optimization is widely used in the process of improving the success rates of machine learning and deep learning models.

III.7 GENERATION OF SYNTHETIC FAILURE

Synthetic failure data was generated using Modelica-based open source simulation models. First, potential failures in the system were identified by Failure Modes, Effects and Criticality Analysis (FMECA) and different failure scenarios were modeled. Simulations were performed by applying different failure profiles (step and ramp type) under boundary conditions such as outdoor temperature, solar radiation and heat demand. These profiles are defined by parameters such as fault onset time, severity and development time. Each simulation recorded the responses of the system for the faulted and unfaulted cases, forming a synthetic data set [19].

The data set was diversified with various failure types and boundary conditions. The onset time and severity of the faults were randomly selected, resulting in a data set that matches real-world conditions. As a result, for each simulation, variables such as boundary conditions, system inputs and outputs, and fault conditions were recorded to create a data set. This data set is structured in accordance with the machine learning models to be used in fault detection and diagnosis. Failure profiles are defined by the parameters given in Figure 1 below:

- Profile type: Failure can occur as a step or ramp.
- Start time (t_0): The moment when the fault occurs.
- End time (t_x): In the ramp profile only, the moment when the fault reaches maximum severity.
- Start intensity (v_0): Always starts at 0.
- Final severity (v_x): varies between 0 and 1, 1 being the maximum fault severity.
- These parameters were used to model different failure scenarios in the simulations.

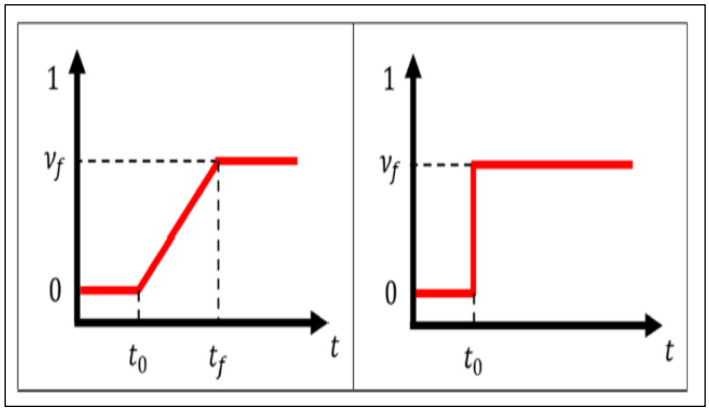

Figure 1: The simulation employed the following fault appearance profiles: ramp on the left for progressive faults and step on the right for sudden faults.
Source: [19].

The fault severity is represented by a number between 0 and 1, where 1 denotes the highest fault severity possible in the model. The model and type of fault will determine this.

IV. RESULTS AND DISCUSSIONS

In this section, the performance results and evaluation of the fault detection models developed for substation contamination are presented. The machine learning models KNN, XGBoost and AdaBoost along with the CNN model are tested and compared to detect faults at different contamination levels. Common performance metrics such as accuracy rate, Matthews correlation coefficient (MCC) and Accuracy are used to determine the performance of the models. The results show that the CNN model has a significant advantage over machine learning methods, especially in detecting more complex and low-level faults. In this section, the performance of each model is discussed in detail. The outstanding achievements of each model will be analyzed and the practical relevance of these findings for DHC systems will be evaluated.

The results of the KNN algorithm showed an accuracy rate of 86.4% and Matthews correlation coefficient performance of 72.3% for the detection of substation contamination faults. In order to improve the performance of the KNN model, hyperparameter optimization was performed with the Grid Search method. In this optimization process, important hyperparameters such as the number of neighbors (k) and distance metric for the KNN algorithm were selected from various values. In Table 2, the optimal hyperparameter values determined by the Grid Search method are presented in detail. These parameters aimed to improve the classification accuracy of the model, but it was observed that the performance of KNN was limited in complex data.

Table 2: KNN regression parameters.

Parameter	Value
n_neighbors	7
weight	distance
algorithm	auto
Leaf_size	30
p	2
n_jobs	10

Source: Authors, (2024).

The confusion matrix of the KNN algorithm is shown in Figure 2. In addition, Table 6 shows all performance metrics of the KNN algorithm.

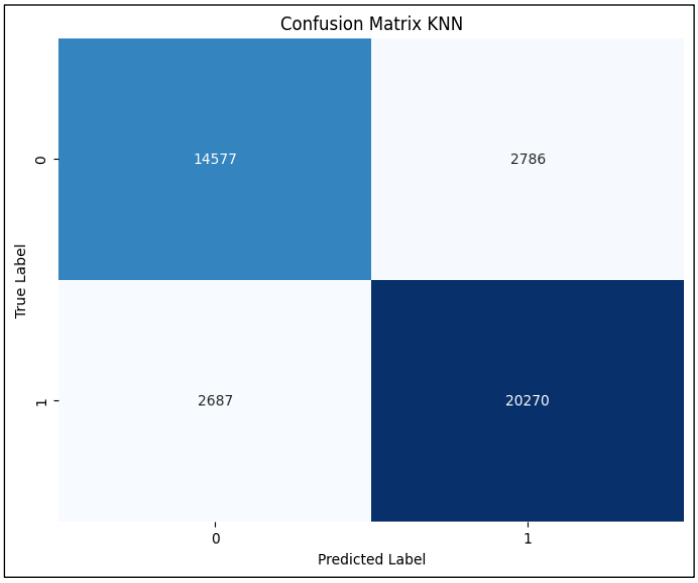

Figure 2: KNN regression confusion matrix.
Source: Authors, (2024).

Figure 3 presents the analysis for different levels of substation contamination in District Heating and Cooling (DHC) systems for the KNN algorithm. The graphs represent four different levels of contamination: very high fouling (75%), medium fouling (20% and 11%), and low fouling (5%). The expression UA [W/K] relates to the heat transfer coefficient in a heat exchange system or thermal system. There are two graphs for each level: The left-hand graphs show the probability of failure detection and the UA [W/K] values together. The graphs on the right-hand side show the correct and incorrect fault detections.

The case where the contamination in the system is very high is analyzed. In the left graph, the UA [W/K] values show a significant decrease with time, showing the effect of fouling and the loss of thermal performance of the system. The yellow dots represent the degree of fouling, while the blue line marks the point of failure. The right graph shows that the UA [W/K] value decreases continuously with time, indicating that the fouling is continuously increasing. In the case of moderate contamination, the spread of the yellow dots in the left graph is more regular and the UA [W/K] values are relatively more stable. However, after a certain point, there is a significant decrease in system performance with increasing fouling. In the right graph, the UA [W/K] value shows a later and slower decline compared to high contamination. It shows a similar pattern for a lower fouling level of 11%, which is lower than the medium level. In the left graph, the yellow dots are less diffuse and the UA [W/K] value decreases less over time. In the right graph, the decline in system performance is again observed, but it is later and less sharp.

Finally, for low levels of fouling, the graphs show a much more stable situation. In the left graph, the yellow dots are more concentrated at the upper levels and the UA [W/K] value remains stable for longer. In the right graph, it can be seen that the UA [W/K] remains largely stable and only slightly decreases.

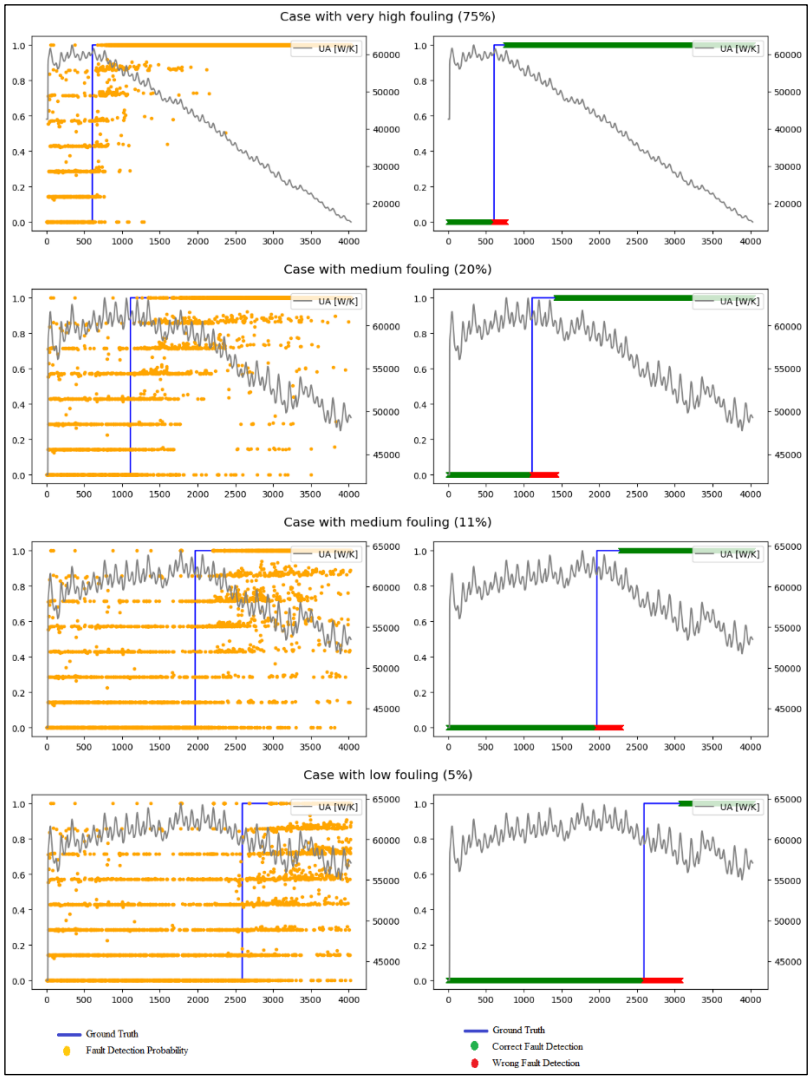

Figure 3: Analysis for different levels of substation contamination in DHC systems with KNN algorithm.
Source: Authors, (2024).

The results of the XGBoost algorithm were found to be highly successful for the detection of substation contamination faults. XGBoost achieved 96.1% accuracy rate and 92.0% Matthews correlation coefficient performance. In order to maximize the performance of the algorithm, hyperparameter tuning was performed using Grid Search Optimization. In this process, important hyperparameters such as learning rate and maximum depth (max_depth) were optimized.

Table 3: XGBoost regression parameters.

Parameter	Value
Base_score	0.7
booster	gbtree
Max_depth	5
Min_child_weight	5
Learning_rate	0.33
n_estimators	250
N_jobs	10
Random_state	5
Tree_method	approx

Source: Authors, (2024).

In Table 3, the optimal hyperparameters obtained as a result of this optimization process are presented in detail. The Grid Search optimization enabled the XGBoost model to detect fouling failures more accurately, resulting in an efficient model with high accuracy and short processing time. The confusion matrix of the XGBoost algorithm is shown in Figure 4. In addition, Table 6 shows all performance metrics of the XGBoost algorithm.

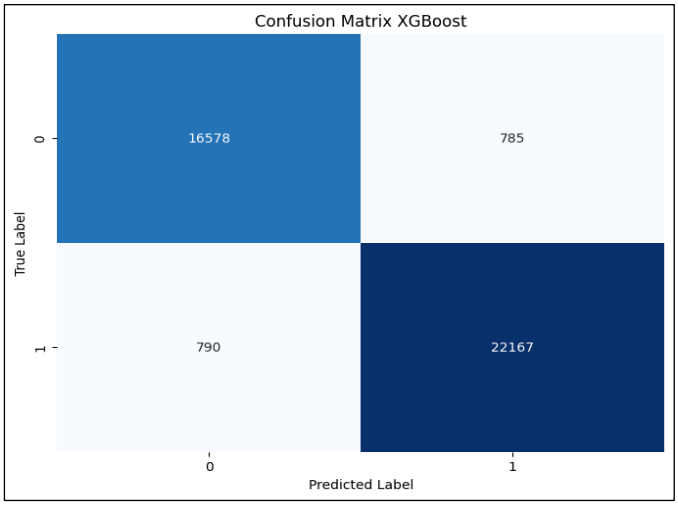

Figure 4: XGBoost regression confusion matrix.
Source: Authors, (2024).

Figure 5 presents the analysis for different levels of substation contamination in District Heating and Cooling (DHC) systems for the XGBoost algorithm. The left graph with high contamination shows the probability of fault detection and UA [W/K] values. In the case of very high contamination, the probability of fault detection increases rapidly from the point where the contamination starts and the system performance decreases significantly. In the right graph, the blue line represents the ground truth, the green area represents correct detections and the red area represents incorrect detections. In the case of very high contamination, the majority of correct failures are detected and the degradation of the system can be clearly observed. In the case of medium contamination, the left graph shows that at 20% contamination, the probability of fault detection fluctuates at first, but increases rapidly after a certain point. The UA [W/K] value shows a steady decrease over time. In the right graph, the correct detections are marked in green and it can be seen that the correct detection rate is high after the onset of contamination. However, false detections are observed for some points. At 11% contamination, the left graph shows that the probability of fault detection increases at a later stage compared to 20% contamination. The UA [W/K] value shows a less pronounced downward trend. In the right graph, correct detections are again indicated by green areas. At this level, correct detections are still predominant, but due to the low contamination, the system takes longer to detect a fault. At low contamination level, the left graph shows that the probability of fault detection remains low for a long time and only increases significantly in the later stages of contamination. The UA [W/K] values are more stable. In the right graph, the correct detections for this low level of contamination are highlighted in green, and the failure detection process is delayed compared to the other levels as the system performance is not degraded much.

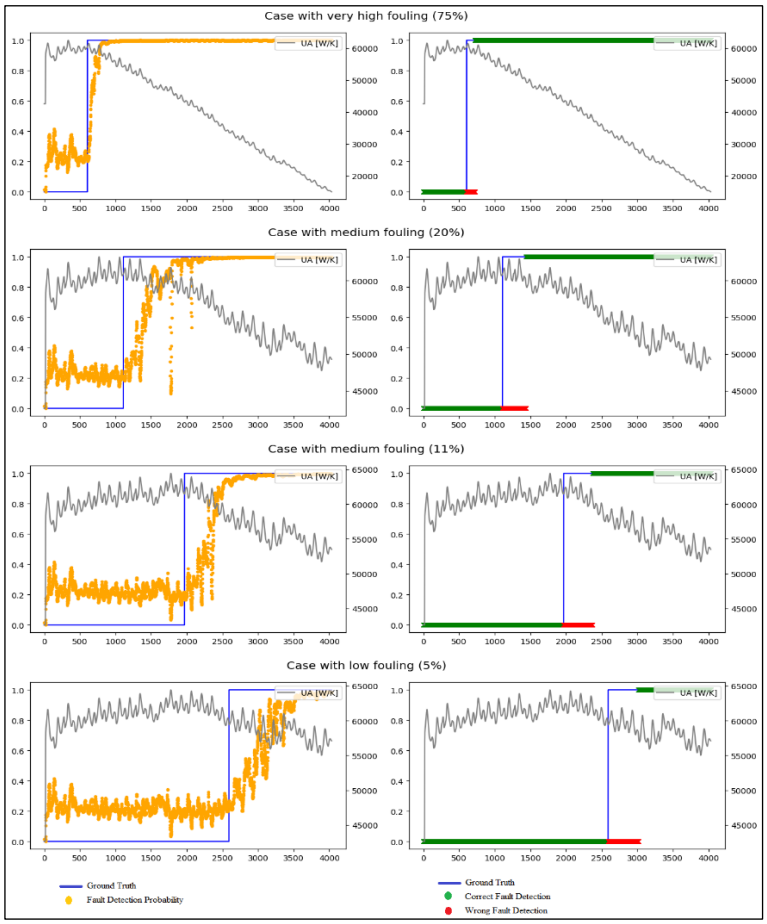

Figure 5: Analysis for different levels of substation contamination in DHC systems with XGBoost algorithm.
Source: Authors, (2024).

The AdaBoost model, a machine learning algorithm for the detection of substation contamination faults, achieved 91.3% accuracy and 82.6% Matthews correlation coefficient. Performance metrics such as model accuracy and Matthews correlation coefficient demonstrate the effectiveness of AdaBoost. In order to maximize the performance of the model, the hyperparameters were determined using the Grid Search Optimization method and the values are presented in Table 4. The confusion matrix of the Adaboost algorithm is shown in Figure 6. In addition, all performance metrics of the Adaboost algorithm are given in Table 6.

Table 4: Adaboost regression parameters.

Parameter	**Value**
Max_depth	5
Min_samples_leaf	5
Min_samples_split	10
n_estimators	250
Learning_rate	0.33
Random_state	5

Source: Authors, (2024).

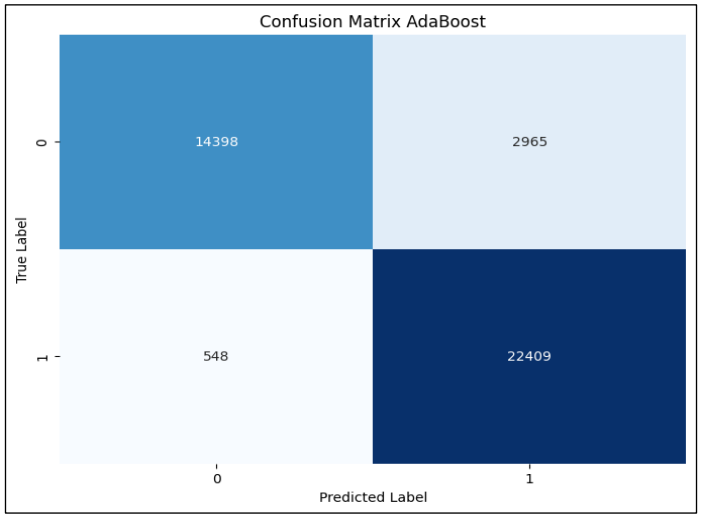

Figure 6: Adaoost regression confusion matrix.
Source: Authors, (2024).

Figure 7 shows the fault detection results for the Adaboost algorithm for different percentages of substation contamination levels. Very high contamination follows a curve with a very high probability of fault detection in the left-hand graph and a constant fault detection in a short time. There is a small difference between the onset of contamination and the point at which fault detection starts. The graph on the right shows that the system makes a correct detection in a very short time and there are almost no false detections. In the case of medium contamination, the probability of fault detection increases gradually and there is a steady detection trend after a value of about 1000. There is a slight delay in fault detection compared to the beginning of the contamination. The graph on the right shows that the system generally makes correct detections and there are few false detections. In the 11% contamination scenario, the probability of failure detection increases more slowly and stabilizes at a later time. This shows that the system is slower to detect lower contamination rates. The graph on the right shows that the correct detections are indicated by the green line and that these correct detections start later, as well as a few incorrect detections. At the lowest contamination level, the probability of fault detection increases significantly later and stabilizes over a longer period of time. This indicates that the system struggles to detect low contamination. The graph on the right shows that correct detections occur quite late and there are a relatively high number of false detections.

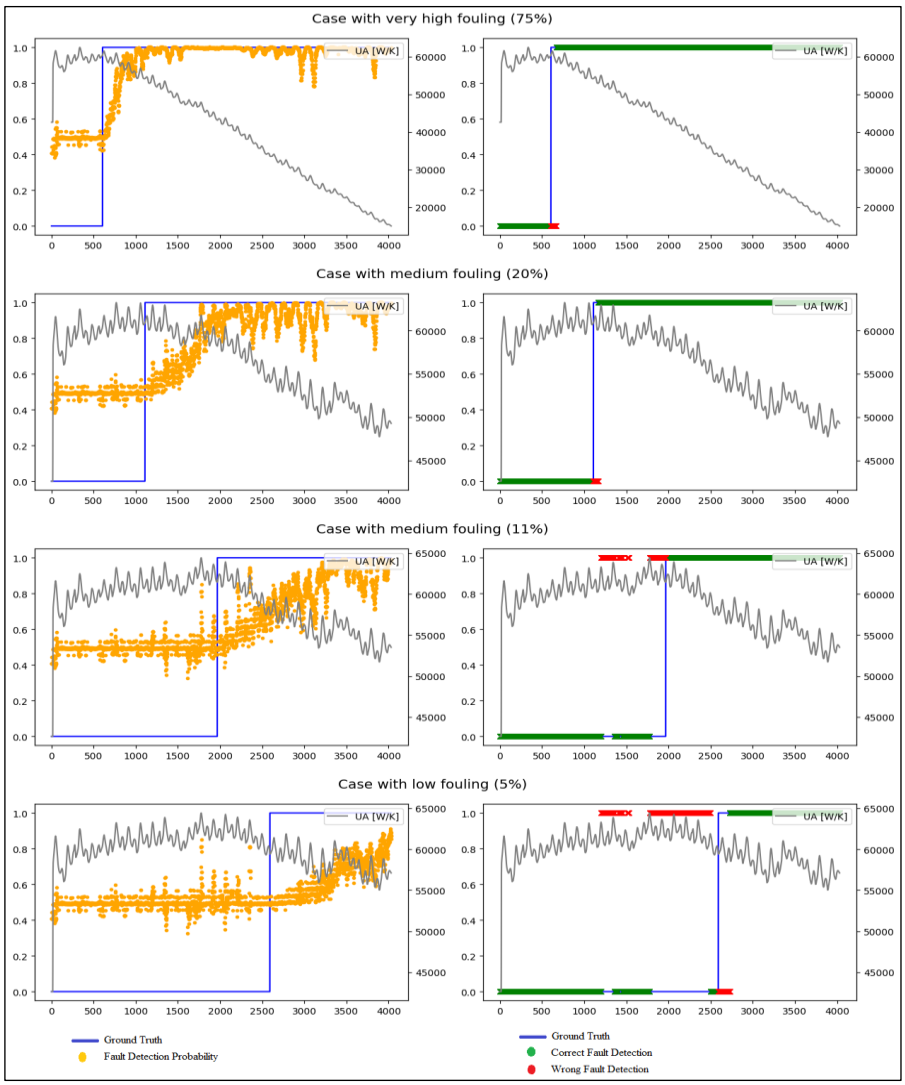

Figure.7: Analysis for different levels of substation contamination in DHC systems with Adaboost algorithm.
Source: Authors, (2024).

In the tests performed with the CNN algorithm, an accuracy rate of 97.2% and a Matthews correlation coefficient value of 94.4% were obtained.

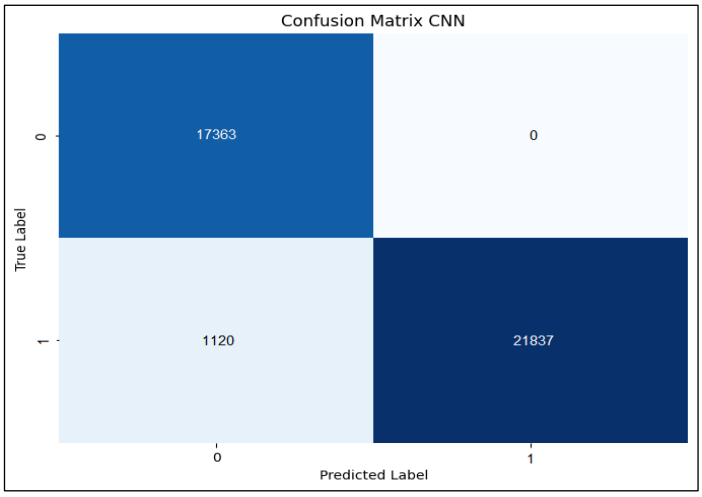

Figure 8: CNN algorithm confusion matrix.
Source: Authors, (2024).

A summary of the CNN model is given in Table 5. Figure 8 shows the confusion matrix of the CNN model. Table 6 shows the results obtained with the CNN algorithm.

Table 5: CNN model summary.

Layer	Output Shape	Param
dense	(None, 128)	1,280
batch_normalization)	(None, 128)	512
dropout	(None, 128)	0
dense	(None, 128)	16,512
batch_normalization	(None, 128)	512
dropout	(None, 128)	0
dense	(None, 64)	8,256
batch_normalization	(None, 64)	256
dropout	(None, 64)	0
dense	(None, 1)	65
Total params		**27,393**
Trainable params		**26,753**
Non-Trainable params		**640**

Source: Authors, (2024).

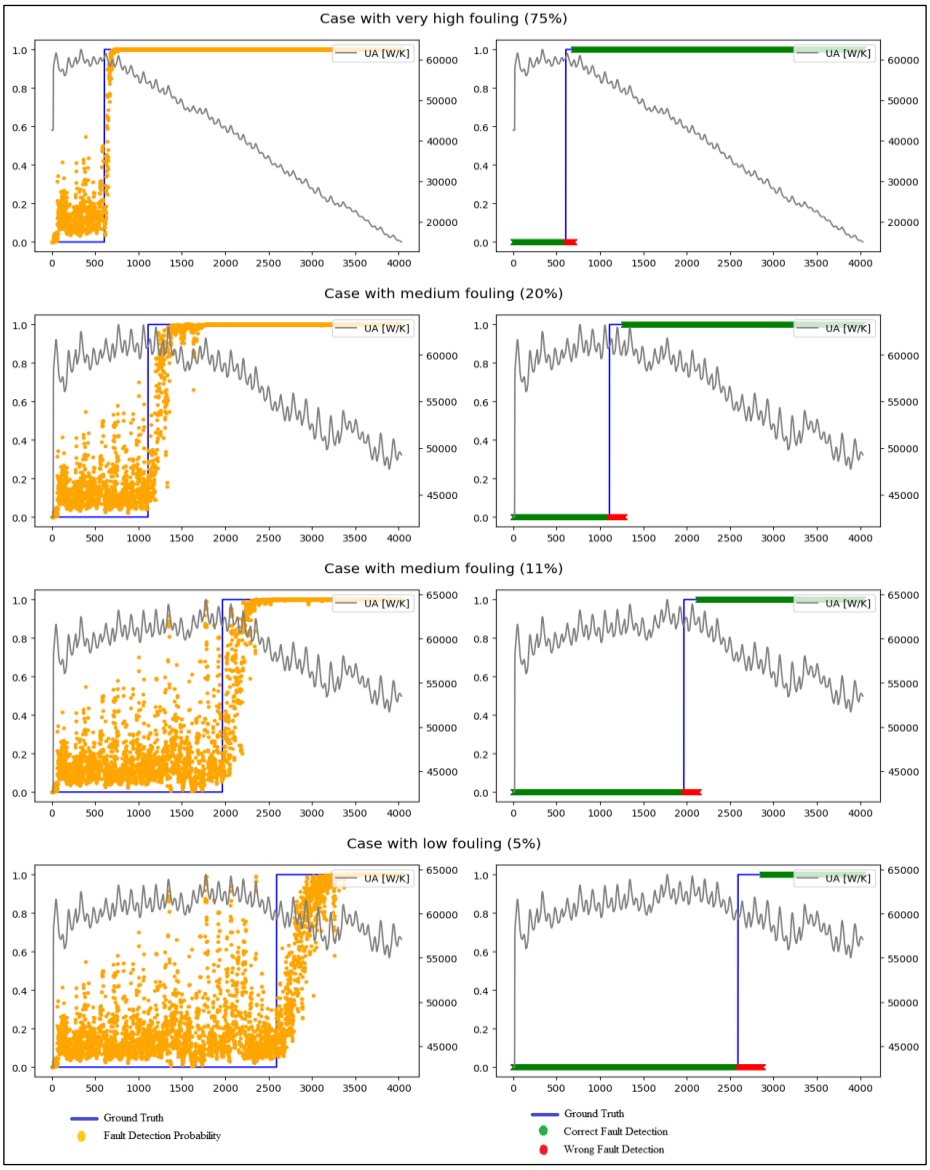

Figure 9: Analysis for different levels of substation contamination in DHC systems with CNN algorithm.
Source: Authors, (2024).

Figure 9 shows the fault detection results for the CNN algorithm for different percentages of substation contamination levels. In the case of very high levels of fouling, the graph on the left, the orange dots show the probabilities of fouling failure detection. The blue line shows the moment of actual fouling onset, while the gray line shows how the heat transfer coefficient (UA) value changes over time. Fouling was correctly detected with a high probability and at an early stage. In the graph on the right, the green and red horizontal lines represent correct and incorrect detections. The length of the green line indicates how long the model correctly detected contamination. In this case, the model has successfully detected very high contamination. The graph on the left shows that for medium contamination at 20%, the model made some incorrect detections but was generally correct. At the 11% contamination level, the detection success seems to be slightly lower, with a shorter green correct detection time and an increased number of incorrect detections. At the low contamination level, the model seems to have a lower probability of detection and a higher number of false detections. The blue line again shows the actual fouling time, while the gray line represents the heat transfer coefficient. In the graph on the right, the green correct detection line is quite short and the red false detection line is long. This indicates that the model struggles to detect the low fouling level.

In general, the graphs show that the accuracy and speed of fault detection varies depending on the contamination level. At very high contamination levels, the system makes fast and accurate detections, while at low contamination levels the detection time is longer and the number of false detections increases. This suggests that the system's ability to detect low levels of contamination is limited. These analyses highlight the importance of automatic fouling detection systems to improve the operational efficiency of DHC systems. Integrating systems with early warning mechanisms can play a critical role in saving energy and extending equipment lifetime.

Table 6: Comparison of algorithm results.

Algorithms	Accuracy	Matthews Corrcoef	Elapsed Time
KNN	86.4	72.3	67.55
XGBoost	96.1	92.0	26.96
AdaBoost	91.3	82.6	239.76
CNN Model	97.2	94.4	542.34

Source: Authors, (2024).

Table 6 shows the performance metrics of four different machine learning and deep learning algorithms for detecting substation contamination failures in District Heating and Cooling (DHC) systems. Accuracy, Matthews correlation coefficient (MCC) and Elapsed Time are evaluated for each algorithm. These metrics are important to understand the success and efficiency of the models in fault detection.

The KNN algorithm achieved 86.4% accuracy and 72.3% Matthews correlation coefficient. This shows that the model performs moderately well, but may misclassify some contamination cases. The KNN algorithm can perform well, especially when working with a small number of data points, but its performance may degrade with large datasets. The processing time was 67.55 seconds, which is a reasonable speed compared to other algorithms. The XGBoost algorithm performed very well with an accuracy of 96.1% and a Matthews correlation coefficient of 92.0%. These results show that XGBoost is capable of efficient classification and accurate detection at different contamination levels. Moreover, the processing time of the algorithm was very low at 26.96 seconds. The optimized nature of XGBoost makes it an effective option for users who want fast and accurate results.

The AdaBoost algorithm produced lower results than XGBoost, but better than KNN, with an accuracy of 91.3% and a Matthews correlation coefficient of 82.6%. AdaBoost uses the technique of strengthening weak classifiers to improve classification performance. Although it achieved a relatively high success rate, the processing time was considerably longer than the other algorithms at 239.76 seconds. This long processing time may limit AdaBoost's usefulness in scenarios with large datasets or fast results. The CNN model outperformed all other algorithms with the highest accuracy rate of 97.2% and the highest Matthews correlation coefficient of 94.4%. This high performance of the CNN model shows that deep learning models can work more effectively with large and complex datasets. However, the CNN model has the longest processing time of 542.34 seconds, indicating that the model requires high computational power and therefore runs for longer periods of time.

Comparing the performance of all algorithms, the CNN model gives the best results in terms of accuracy and Matthews correlation coefficient, but this superior performance is achieved at the cost of a longer processing time. XGBoost, on the other hand, produces almost as good results as CNN, but with a much shorter processing time, which makes it advantageous in practical applications. AdaBoost performs well in terms of accuracy, but is at a disadvantage in terms of processing time. The KNN algorithm, although one of the fastest algorithms, has lower accuracy and Matthews correlation coefficient compared to the other models. In conclusion, each algorithm offers different advantages and disadvantages in terms of accuracy, speed and computational cost.

V. CONCLUSIONS

In this study, various machine learning and deep learning algorithms are investigated for the detection of substation contamination faults in District Heating and Cooling (DHC) systems. The algorithms used include K-Nearest Neighbors (KNN), XGBoost, AdaBoost and Convolutional Neural Network (CNN) models. Contamination failures were analyzed at high, medium and low levels and hyperparameter optimization for the detection of these failures was performed by Grid Search method. The results show that the CNN model offers the best performance with 97.2% accuracy and 94.4% Matthews correlation coefficient. However, CNN has the longest processing time (542.34 seconds). XGBoost stood out as a fast and efficient alternative with 96.1% accuracy and short processing time (26.96 seconds). AdaBoost, despite having an accuracy of 91.3%, was quite slow with a processing time of 239.76 seconds. KNN, on the other hand, showed the lowest performance with an accuracy of 86.4%, and although it works fast, it is insufficient for complex data. In future studies, it is recommended to test these models on real-world datasets, optimize the processing time of CNN models, and investigate different types of failures more comprehensively.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Emrah Aslan and, Yıldırım Özüpak
Methodology: Emrah Aslan and, Yıldırım Özüpak
Investigation: Emrah Aslan and, Yıldırım Özüpak
Discussion of results: Emrah Aslan and, Yıldırım Özüpak
Writing – Original Draft: Emrah Aslan and, Yıldırım Özüpak
Supervision: Emrah Aslan and, Yıldırım Özüpak
Approval of the final text: Emrah Aslan and, Yıldırım Özüpak

VIII. REFERENCES

[1] Z. Ren, T. Lin, K. Feng, Y. Zhu, and Z. Liu, "A systematic review on imbalanced learning methods in intelligent fault diagnosis," IEEE Transactions on Instrumentation and Measurement, vol. 72, 2023. doi:10.1109/TIM.2023.3246470.

[2] L. Jin, D. K. Kim, K. Y. Chan, and A. Siada, "Deep Machine Learning-Based Asset Management Approach for Oil Immersed Power Transformers Using Dissolved Gas Analysis," IEEE Access, vol. 12, 2024. doi: 10.1109/ACCESS.2024.3366905.

[3] S. Zhou, Z. O'Neill, and C. O'Neill, "A review of leakage detection methods for district heating networks," Applied Thermal Engineering, vol. 137, pp. 567–574, 2018. doi: 10.1016/j.applthermaleng.2018.04.010.

[4] J. Zheng, Z. Zhou, J. Zhao, and J. Wang, "Effects of the operation regulation modes of district heating system on an integrated heat and power dispatch system for wind power integration," Applied Energy, vol. 230, pp. 1126–1139, 2018. doi: 10.1016/j.apenergy.2018.09.077.

[5] S. Månsson, I. L. Benzi, M. Thern, R. Salenbien, K. Sernhed, and P. J. Kallioniemi, "A taxonomy for labeling deviations in district heating customer data," Smart Energy, vol. 2, 2021. doi: https://doi.org/10.1016/j.segy.2021.100020

[6] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 3125–3148, 2018. doi: 10.1109/TSG.2018.2818167

[7] D. Qu, W. Luo, Y. Liu, B. Fu, Y. Zhou, and F. Zhang, "Simulation and experimental study on the pump efficiency improvement of continuously variable transmission," Mechanism and Machine Theory, vol. 131, pp. 137–151, Jan. 2019. doi: https://doi.org/10.1016/j.mechmachtheory.2018.09.014.

[8] A. Rodler, S. Guernouti, M. Musy, and J. Bouyer, "Thermal behaviour of a building in its environment: Modelling, experimentation, and comparison," Energy and Buildings, vol. 168, pp. 19–34, Jun. 2018. doi: https://doi.org/10.1016/j.enbuild.2018.03.008.

[9] M. Jebamalai, K. Marlein, and J. Laverge, "Design and cost comparison of district heating and cooling (DHC) network configurations using ring topology – A case study," Energy, vol. 258, 2022, Art. no. 124777. [Online]. Available: https://doi.org/10.1016/j.energy.2022.124777.

[10] R. Patureau, C. T. Tran, V. Gavan, and P. Stabat, "The new generation of District heating & cooling networks and their potential development in France," Energy, vol. 236, 2021, Art. no. 121477. [Online]. Available: https://doi.org/10.1016/j.energy.2021.121477.

[11] A. S. Gaur, D. Z. Fitiwi, and J. Curtis, "Heat pumps and our low-carbon future: A comprehensive review," Energy Research & Social Science, vol. 71, 2021, Art. no. 101764. [Online]. Available: https://doi.org/10.1016/j.erss.2020.101764.

[12] S. Buffa, M. Cozzini, M. D'Antoni, M. Baratieri, and R. Fedrizzi, "5th generation district heating and cooling systems: A review of existing cases in Europe," Renewable and Sustainable Energy Reviews, vol. 104, pp. 504-522, 2019. [Online]. Available: https://doi.org/10.1016/j.rser.2018.12.059.

[13] V. Munćan, I. Mujan, D. Macura, and A. S. Anđelković, "The state of district heating and cooling in Europe - A literature-based assessment," Energy, vol. 304, 2024, Art. no. 132191. [Online]. Available: https://doi.org/10.1016/j.energy.2024.132191.

[14] K. Gjoka, B. Rismanchi, and R. H. Crawford, "Fifth-generation district heating and cooling: Opportunities and implementation challenges in a mild climate," Energy, vol. 286, 2024, Art. no. 129525. [Online]. Available: https://doi.org/10.1016/j.energy.2023.129525.

[15] Y. Zhang, M. Liu, Z. O'Neill, and J. Wen, "Temperature control strategies for fifth generation district heating and cooling systems: A review and case study," Applied Energy, vol. 376, Part A, 2024, Art. no. 124156. [Online]. Available: https://doi.org/10.1016/j.apenergy.2024.124156.

[16] K. Autelitano, J. Famiglietti, M. Aprile, and M. Motta, "Towards Life Cycle Assessment for the Environmental Evaluation of District Heating and Cooling: A Critical Review," Standards, vol. 4, no. 3, pp. 102-132, 2024. [Online]. Available: https://doi.org/10.3390/standards4030007.

[17] A. J. Patil, R. Naresh, R. K. Jarial, and H. Malik, "Optimized Synthetic Data Integration with Transformer's DGA Data for Improved ML-based Fault Identification," IEEE Transactions on Dielectrics and Electrical Insulation, 2024. doi: 10.1109/TDEI.2024.3421915.

[18] M. Vallee, T. Wissocq, Y. Gaoua, and N. Lamaison, "Generation and evaluation of a synthetic dataset to improve fault detection in district heating and cooling systems," Energy, vol. 283, p. 128387, 2023. doi: 10.1016/j.energy.2023.128387.

[19] S. Buffa, M. H. Fouladfar, G. Franchini, I. Lozano Gabarre, and M. Andrés Chicote, "Advanced Control and Fault Detection Strategies for District Heating and Cooling Systems—A Review," Applied Sciences, vol. 11, p. 455, 2021. doi: 10.3390/app11010455.

[20] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, and J. Liu, "Machine learning-based leakage fault detection for district heating networks," Energy and Buildings, vol. 223, p. 110161, 2020. doi: 10.1016/j.enbuild.2020.110161.

[21] S. Månsson, P.-O. J. Kallioniemi, K. Sernhed, and M. Thern, "A machine learning approach to fault detection in district heating substations," Energy Procedia, vol. 149, pp. 226-235, 2018. doi: 10.1016/j.egypro.2018.08.187.

[22] D. Leiria, K. H. Andersen, S. P. Melgaard, H. Johra, A. Marszal-Pomianowska, M. S. Piscitelli, A. Capozzoli, and M. Z. Pomianowski, "Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth," in Proceedings of Building Simulation 2023: 18th Conference of IBPSA. doi: https://doi.org/10.26868/25222708.2023.1576.

[23] F. Theusch, P. Klein, R. Bergmann, W. Wilke, W. Bock, and A. Weber, "Fault Detection and Condition Monitoring in District Heating Using Smart Meter Data," PHM Society European Conference, vol. 6, no. 1, p. 11, 2021. doi: 10.36001/phme.2021.v6i1.2786.

[24] M. Li, W. Deng, K. Xiahou, T. Ji, and Q. Wu, "A Data-Driven Method for Fault Detection and Isolation of the Integrated Energy-Based District Heating System," IEEE Access, vol. 8, pp. 23787-23801, 2020. doi: 10.1109/ACCESS.2020.2970273.

[25] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, and J. Liu, "Machine learning-based leakage fault detection for district heating networks," Energy and Buildings, vol. 223, p. 110161, 2020. doi: 10.1016/j.enbuild.2020.110161.

[26] M.Valle, "DHC substation fouling - synthetic faults," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/mathieuvallee/ai-dhc-substation-fouling/data. Accessed: July. 29, 2024.

[27] B. Said, L. Mazouz, T. NAAS, Özüpak Yildirim, and R. Mohammedi, "Broken magnets fault detection in pmsm using a convolutional neural network and SVM", JETIA, vol. 10, no. 48, pp. 55-62, Jul. 2024. doi: https://doi.org/10.5935/jetia.v10i48.1185.

[28] P. Dimri, A. Nath, and P. Dimri, "Flood prediction and management", JETIA, vol. 10, no. 47, pp. 95-103, Jul. 2024. doi: https://doi.org/10.5935/jetia.v10i47.1103.

[29] E. Aslan, "Temperature prediction and performance comparison of permanent magnet synchronous motors using different machine learning techniques for early failure detection," Eksploatacja i Niezawodność – Maintenance and Reliability, 2024. doi: https://doi.org/10.17531/ein/192164.

[30] S. Gonjari, R. Pawar, R. Pawar, S. Kshirsagar, and R. Kokare, "Real time emotion recognition and classification for diverse suggestions using Deep Learning", JETIA, vol. 10, no. 48, pp. 14-20, Jul. 2024. doi :https://doi.org/10.5935/jetia.v10i48.1007.