The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 2, April 2025
https://doi.org/10.33022/ijcs.v14i2.4674

A Novel Leak Detection Algorithm Based on SVM-CNN-GT for Water Distribution Networks

Giresse Komba[1], Topside E. Mathonsi[2], Pius A. Owolawi[3]
ghimco@gmail.com[1], mathonsite@tut.ac.za[2] ,
owolawipa@tut.ac.zamthulisi.velempini@ul.ac.za[3]
[1,3] Department of Computer Systems Engineering, Tshwane University of Technology, Pretoria, South Africa
[2] Department of Information Technology, Tshwane University of Technology, Pretoria, South Africa

Article Information

Received : 31 Jan 2025
Revised : 27 Apr 2025
Accepted : 30 Apr 2025

Keywords

Water Distribution Networks, pipeline leak detection, Sensor Placement, Machine Learning, SVM-CNN-GT algorithm.

Abstract

Water Distribution Networks (WDNs) suffer substantial water losses due to pipeline leaks, resulting in economic ramifications and exacerbating global water scarcity concerns. This paper seeks to improve the precision of leak detection and the identification of leak locations within WDNs. The pervasive issue of leaks in WDNs poses significant challenges with economic and environmental implications for water utilities. Traditional leak detection methods are time-consuming, resource-intensive, and susceptible to inaccuracies and false alarms due to the random placement of sensors. The detection of concealed background leaks, invisible to the naked eye and situated beneath the surface, presents a particular challenge. This situation complicates efforts for their real-time identification and subsequent repairs. To address these challenges, this paper introduces the SVM-CNN-GT algorithm, an advanced ensemble supervised Machine Learning (ML) approach that incorporates Support Vector Machines (SVM), Convolutional Neural Network (CNN), and Graph Theory (GT). By combining multiple ML algorithms, the SVM-CNN-GT model takes into account various factors that influence leak detection and localization, resulting in more precise and reliable assessments of leak presence and location. The algorithm leverages automatic feature extraction and heterogeneous dual classifiers to accurately assess leaks based on data related to flow rate, pressure, and temperature.

Furthermore, a combination probability scheme enhances leak detection efficiency by integrating diverse classifier models with distinct prediction outputs. Through the EPANET performance evaluations, the SVM-CNN-GT algorithm outperforms CNN and SVM algorithms, demonstrating remarkable proficiency with the highest average leak detection accuracy of 98%, followed by CNN at 82% and SVM at 78%.

A. Introduction

Water Distribution Networks (WDNs) play a crucial role in providing clean and safe water to communities around the world [1, 2]. However, these networks face significant challenges, with leakage being one of the most pressing issues [3, 4]. Leakage refers to the loss of water from the distribution system due to pipe failures, cracks, or other infrastructure problems [5, 6]. This problem not only leads to a waste of precious resources but also poses financial burdens on water utilities [7]. The extent of water loss due to leakage is staggering. As per the World Bank's data, developing countries experience a daily water loss of approximately 45 million cubic meters, which translates to an annual economic loss exceeding US$3 billion [8, 9]. For instance, according to a study conducted by the Water Research Commission (WRC) of South Africa, it is estimated that around 35% of the country's treated water is lost due to leakage [10, 11]. Non-Revenue Water (NRW), also known as water loss or unaccounted for water, refers to the volume of water that is pumped into WDNs but is not billed or consumed by customers [12, 13]. In South Africa, NRW translates to an annual financial loss of approximately R9.9 billion [14, 15]. This loss occurs due to various factors such as leaks, theft, meter inaccuracies, and unauthorized consumption [16].

Falkenmark's indicator water availability threshold states that a country is considered to be under water scarcity if its per capita water consumption is below 1000 m3 [17, 18]. Additionally, a per capita water consumption of less than 1700 m3 is classified as a water stress situation [19]. In 2025 African countries will either be in a state of water scarcity, stress or vulnerable to changes in water supply as the projections indicate (see Figure. 1) [20, 21]. The higher the value of the indicator, the greater the water stress and the need for conservation efforts.

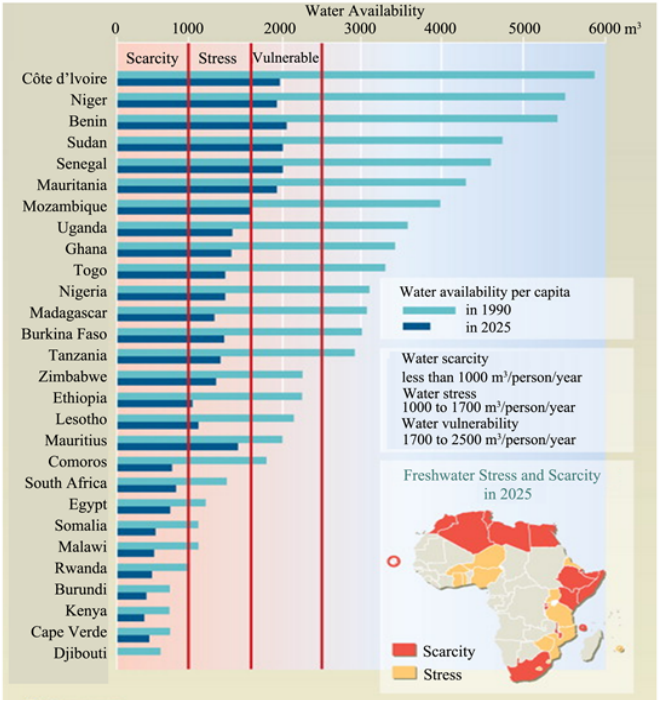

Figure 1. The world water scarcity by 2025 [21]

Figure. 1. Within WDNs, there exist three distinct categories of leakage: reported, unreported, and background leakages as illustrated in Figure 2. Reported and unreported leakages stem from structural pipe failures or bursts, characterized by sudden drops in network water pressure, which render them detectable and are frequently reported by either the public or water utility personnel [22-24].

Figure 2. WDNs leakage type (a) burst leakage (b) background leakage

In contrast, background leakages manifest when there are minor cracks, holes, deteriorated joints, or fittings in the pipes, resulting in a continuous and subtle water outflow [25, 26]. Unlike reported and unreported leakages, background leakages in WDNs remain hidden and do not immediately or significantly reduce water pressure [27, 28]. As a result, they often go unnoticed and can persist for longer periods [29, 30]. However, it is crucial to acknowledge that these background leakages still contribute to the overall volume of water loss in the network [31, 32]. Approximately 90% of water loss in WDNs is attributed to these leaks. Nonetheless, due to the sensitivity of leakage to network pressure, reducing the network pressure has been acknowledged to be a valuable intervention tool in minimizing background leakage as well as reducing the frequency of pipe bursts.

A multitude of techniques have been proposed for detecting leakages in WDNs. However, there is a significant uncertainty when it comes to detecting background leakage, rendering these techniques ineffective. Background leakages tend to increase with the internal pressure of pipes. To mitigate water losses caused by leaking pipes, it is worthwhile to reduce excessive pressure at strategic areas in the WDNs. However, due to the complexity of WDNs, identifying the specific areas or nodes in the network and the exact leaking pipelines connected to them, where pressure control measures can be implemented, poses a challenging task.

The main contributions of this paper are summarized as follows:

1) This paper presents an ensemble Machine Learning (ML) model called the SVM-CNN-GT algorithm, which combines Support Vector Machines (SVM), Convolutional Neural Network (CNN), and Graph Theory (GT). The algorithm is designed for water leakage detection in WDNs and utilizes automatic feature extraction and heterogeneous dual classifiers. By considering flow rate, pressure, and temperature data, the SVM-CNN-GT algorithm accurately evaluates leaks. Additionally, a combination probability scheme is proposed to integrate diverse classifier models with varying prediction outputs.

2) To improve the accuracy of classification and decrease the time required for learning, we have incorporated optimal learning parameters into our ensemble SVM-CNN-GT model. These parameters include water pressure, flow rates, temperature, and leak information. By utilizing these parameters effectively, we are able to minimize the total number of free parameters used.
3) To enhance the accuracy of location estimation, a novel GT-based local search algorithm incorporating a virtual node scheme is being introduced. The proposed algorithm aims to minimize errors in distance estimation, repair costs, and time, thereby providing significant advantages for large-scale location-aware applications.
4) To effectively identify faults and locate their sources, the proposed method employ a range of sensors such as pressure sensors, vibration sensors, and flow sensors. Thus, this adaptable approach can also be extended to the detection of gas leaks.

The remaining sections of this paper are organized as follows: Section B provides a comprehensive review of previous approaches used to detect and locate pipeline leaks within WDNs. Section C delves into the architecture and formulations of the proposed algorithm. In Section D, we present simulation results, describe the experimental setup, provide details about the dataset used, and conduct a performance analysis. Finally, Section E concludes with a summary of the findings and suggests potential avenues for future research.

B. Related Work

This section provides an in-depth analysis of various methodologies used to detect and locate pipeline leaks within WDNs. Moreover, this section explores the advantages and constraints associated with each approach.

Porwal et al. [33] proposed a weighted-sample SVM algorithm for leak detection in WDNs, which enhanced classification accuracy and mitigated noise and outliers. Their algorithm outperformed other methods like acoustic signal analysis, transient signal analysis, and temperature variation analysis in leak detection accuracy. The proposed SVM-CNN-GT algorithm builds upon this approach by considering optimal sensor placement, optimizing detection, and reducing costs. Unlike their approach of randomly placing sensors, the SVM-CNN-GT algorithm uses a more strategic and efficient method for sensor placement.

Zhou et al. [34] proposed a leak detection algorithm for WDNs that combined SVM and Kernel Principal Component Analysis (KPCA). The objective of their approach was to improve the accuracy of leak detection in complex networks by utilizing KPCA to extract relevant features from flow data and SVM for classification purposes. Through evaluation using Flowmaster software, the KPCA algorithm demonstrated superior performance compared to Support Vector Data Description (SVDD) and k-means in terms of leak detection accuracy. However, their method failed to consider the significance of minimizing distance errors, which led to delays in leak repairs and increased costs. Inefficient leak localization could necessitate multiple attempts for accurate detection, resulting in additional expenses related to excavation, repair, and maintenance.

Lang et al. [35] proposed a leak detection and location technique that combined the Least Squares Support Vector Machine (LS-SVM) and Local Maxima

Decomposition (LMD). Their method was evaluated using Flowmaster software on both simulated and real-world WDNs data. The results showed that their approach outperformed wavelet transform and particle swarm optimization in terms of accuracy. However, their algorithm which was designed by integrating LS-SVM and LMD algorithms lacked strategic sensor placement, resulting in excessive sensor deployment. In addition, the assimilation of LS-SVM and LMD algorithms increased the computational complexity, leading to higher overhead and compromising leak detection accuracy in WDNs.

Lučin et al. [36] proposed an ML approach for identifying leak locations within WDNs using a Random Forest classifier. They conducted simulations to test the effectiveness of their method, considering various leak scenarios with Monte Carlo-generated parameters and demand fluctuations. By analyzing raw pressure sensor data collected over 24 hours, their approach accurately detected and located leaks by utilizing EPANET simulations. The researchers employed Scikit-learn and high-performance computing to achieve reliable leak detection even with sparse sensor placement. However, the absence of strategic sensor placement consideration in their approach resulted in increased installation costs, system design complexities, and modeling efforts. Additionally, the lack of location estimation in their algorithm posed challenges for precise leak detection, leading to water loss and potential infrastructure damage.

Rajabi et al. [37] proposed a Conditional Deep Convolutional Generative Adversarial Networks (CDcGANs) algorithm for leak detection and localization in WDNs. Their method aims to overcome the limitations of traditional leak detection techniques by leveraging the power of deep learning and generative adversarial networks. CDcGANs are trained on a dataset of simulated pressure measurements to learn the underlying patterns and characteristics of leaks in the network. The trained model is then used to detect and localize leaks in real-world scenarios, achieving high accuracy and efficiency. However, the researchers failed to consider the strategic placement of sensors, which led to various negative consequences. Firstly, this oversight resulted in increased installation costs as the sensors were not optimally positioned. Strategic sensor placement is crucial for efficient and cost-effective monitoring systems, as it ensures that the sensors are located in areas where they can effectively detect leaks or other anomalies. Without considering this aspect, the researchers incurred unnecessary expenses during the installation process.

Liu [32] introduces an innovative approach to improve the efficiency of water pipeline leakage detection by combining ML with Wireless Sensor Networks (WSNs). The core of this method is the utilization of SVM as a classifier for identifying pipeline leaks. Wireless sensors are strategically placed on water pipelines to collect data, which is then transmitted via a 4G network. To reduce energy consumption in the WSNs, a "leakage-triggered networking" mechanism is employed, extending the system's operational life. To enhance the precision of leak detection, the algorithm leverages techniques like the intrinsic mode function, approximate entropy, and principal component analysis, creating a set of signal features. The evaluation of this method was carried out using OPNET Modeler 14.5 as the simulation tool, involving the design and conFigureuration of different network layers, including node, process, and network layers. ZigBee nodes were

chosen for this purpose, and simulation parameters were adjusted to meet the research requirements.

Although this method shares similarities with the SVM-CNN-GT algorithm, both relying on SVMs for pattern recognition and classification, a limitation in their study was the absence of strategic sensor placement. This led to increased installation costs, as sensors were not optimally positioned. Proper sensor placement is crucial for cost-effective monitoring systems, ensuring they are in locations suitable for efficient leak and anomaly detection.

Furthermore, the SVM-CNN-GT algorithm integrates a CNN model that employs one-dimensional convolutions for sequence processing. Its primary goal is to detect pipeline leaks in WDNs, enabling real-time data processing, anomaly identification, and leak prediction. This approach contributes to minimizing water loss and reducing environmental impacts through feed-forward and backpropagation techniques.

C. Proposed Leak Detection And Architecture

The architecture of Internet of Things (IoT) based WSNs for WDNs, as depicted in Figure. 3 has the potential to bring about a revolutionary change in the way we detect and localize leaks in pipelines. By strategically deploying sensors using GT principles, the SVM-CNN-GT algorithm achieves remarkable accuracy in leak detection. Optimized sensor placements enhance coverage, facilitate data collection, and enable precise predictions, thereby reducing false alarms. Leveraging the GT, the SVM-CNN-GT algorithm maximizes the utilization of sensor data, leading to improved leak detection compared to conventional methods. This is accomplished through the integration of pressure, vibration, and flow sensors, which collaborate to measure leaks and analyze water flow patterns.

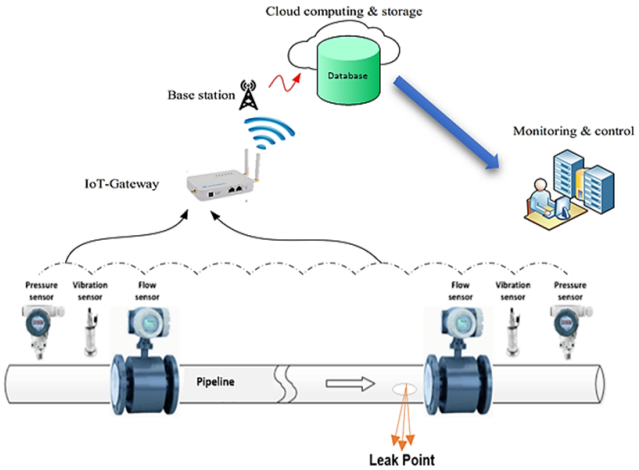

Figure 3. Pipeline leak detection and Location architecture

This architecture allows for immediate analysis and response to leaks by transmitting sensor data from the IoT gateway. The use of cloud platforms enhances data analysis and management, with archived data offering valuable insights into pipeline performance and problems. Proactive maintenance is

achieved through the identification of trends that cause leaks, optimization of detection methods, and overall improvement in pipeline management.

A. Graph Theory

GT examines entity relationships through nodes and edges, essential for network optimization [38, 39]. This paper utilizes GT to reduce leak localization errors and costs, optimizing sensor placement in WDNs for accurate leak localization and focused repairs. Efficiency benefits extensive networks, with virtual node search limits representing WDNs pipelines as graphs.

Consider the graph $G(V,E)$ which serves as the representation of the pipeline network, where V represents the collection of nodes (pipe junctions and measurement points with sensor nodes), E represents the set of edges connecting nodes. The edge weights, denoted by $W(u,v)$, represent the pipe lengths between nodes u and v . We assume the presence of a continuous water leak at a specific location within the water pipe network, the algorithm employs equations (1) and (2) in conjunction with Dijkstra's algorithm to estimate the temporary location of the leak along the graph's edge. The nearest linked node is represented as the starting virtual node (V).

The arrangement of virtual nodes is determined by considerations such as accuracy in resolution and computational resource demands. The cost function is responsible for measuring the minimal error in distance or time difference between the actual and virtually generated leakage signals and can be represented mathematically through the utilization of equation (1).

$$Cost(X_i) = \sum \left(k \in SkN\right)((t_j, t_k) - \left(v_{ij} - v_{ik}\right)^2 \quad (1)$$

Where t_j is the timestamp of the actual leakage signal at node j, t_k is the timestamp of the actual leakage signal at neighboring node k in the neighborhood SkN of node j, v_{ij} is the virtually generated leakage signal at node j corresponding to the virtual node X_i, SkN represents the set of neighboring nodes of node j, and the summation is taken over all neighboring nodes k in the neighborhood SkN of node j.

Additionally, the expression for the virtual node X_i can be represented by equation (2):

$$V_i = \text{argmin}\left(\cos t(X_i)\right) \quad (2)$$

Where V_i represents the virtual node chosen as the starting location, and argmin denotes the function that finds the argument (in this case, X_i) that minimizes the cost function Cost (X_i). This process allows the algorithm to identify the virtual node with the least error, resulting in a more accurate estimation of the leakage location within the WDNs.

B. Support Vector Machines

SVM is a powerful supervised learning model that leverage specialized learning algorithms to identify patterns in data and make predictions [40]. Thus, this model are widely used in data analysis for both classification and regression tasks [41, 42]. SVM work by creating a hyperplane or a set of hyperplanes in a high-dimensional space [43, 44]. These hyperplanes are used to separate different classes or to approximate the relationship between input variables and output values, as illustrated in Figure. 4 The goal of SVMs is to find the optimal hyperplane(s) that maximally separate the data points or minimize the error in regression analysis [45, 46].

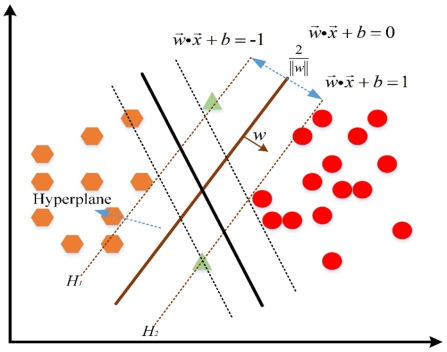

Figure 4. Classification Process of SVM algorithm [29].

Figure. 4 provides a visual representation of the SVM algorithm for classification. This schematic depiction illustrates how the algorithm separates two classes within a dataset using a hyperplane. The hyperplane acts as the solution for classifying the data points into their respective classes.

The SVM algorithm has been widely utilized in the field of water management to detect and predict water losses caused by leakage in WDNs. By analyzing various data variables such as water pressure, flow rates, temperature, and leak information, the SVM algorithm has proven to be effective in accurately predicting the location and severity of leaks in WDNs.

1) Data Variables

The input data variables used for training and prediction are represented by a data matrix $X = [x_1, x_2, x_3 \ldots x_n]$, where x_i represents a specific variable (e.g., water pressure, flow rate, temperature, leak information). There are n total data variables. This data matrix X can be represented by equation (3).

$$X = [x_1, x_2, x_3 \ldots x_n] \quad (3)$$

equation (1) serves as the foundation for defining the input data matrix X where each row x_i corresponds to a data point with n features (data variables). This data matrix is a crucial component used in the SVM algorithm for performing tasks such as leak detection and severity estimation in WDNs.

2) Labels

The equation $Y = [y_1, y_2, y_m]$ represents the label vector for the training data. Each y_1 denotes the presence (1) or absence (-1) of a leak in the WDNs data point x_i. There are m total data points in the label vector Y. This is represented by equation (4):

$$Y = [y_1, y_2, y_3, \ldots y_m] \quad (4)$$

The label vector Y plays a critical role in training the SVM algorithm. It provides the ground truth information needed for the SVM to learn and distinguish between leak and non-leak instances in the WDNs data. By associating the input data matrix X with the corresponding label vector Y, the SVM algorithm can accurately perform tasks like leak detection and severity estimation in WDNs.

3) SVM for Binary Leak Detection

The standard SVM formulation for binary classification, which aims to separate the data points into two classes: one representing leaks and the other non-leak situations. This can be represented by equation (5):

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \max(0, 1 - y_i(w.x_i - b)) \quad (5)$$

Where, is the weight vector of the hyperplane, is the bias term (intercept), is the regularization errors. It is a hyperparameter set by the user. The objective is to find the optimal hyperplane that best separates leak and non-leak data points.

4) SVM for Multiclass Leak Detection

In some cases, there might be multiple classes representing different levels of leaks (e.g., small leaks, medium leaks, large leaks). In such cases, you can use a multiclass SVM approach, such as one-vs-rest or one-vs-one. The objective function is similar to the binary case, but it extends to multiple classes and is represented by equation (6):

$$\min_{w_k}, b_k \frac{1}{2}\|w_k\|^2 + C\sum_{i=1}^{m} \max(0, 1 - \delta_{y_i,k}(w_k.x_i - b_k)) \quad (6)$$

Where, w_k and b_k are the weight vector and bias term for class k, $\delta_{y_i,k}$ is an indicator function that equals 1 if y_i is equal to class k and 0 otherwise.

The objective is to find multiple hyperplanes that separate each class from the rest.

5) SVM for Leak Severity Prediction

If the SVM is employed to predict the severity of leaks, it can be considered a regression problem. In this case, the objective function is represented by equation (7):

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} y_i - (w.x_i) \quad (7)$$

Where, and are the regression parameters (weight and bias).

The objective is to find the optimal regression line that best fits the data points, minimizing the regression error.

C. Convolutional Neural Network

CNN, short for Convolutional Neural Network, is a type of deep feed-forward Artificial Neural Network (ANN) that has gained significant recognition for its exceptional ability to generalize well when compared to networks that utilize fully connected layers [47, 48]. CNNs have brought about a revolution in various fields, particularly in computer vision tasks like image classification and object detection [49, 50]. They have proven to be highly efficient in extracting abstract features from objects, especially when dealing with spatial data [51]. This deep CNN model consists of layers dedicated to processing, enabling the learning of diverse input data features, such as images, at multiple levels of abstraction [52]. Figure. 5 illustrates the CNN as a deep learning architecture, encompassing convolutional and fully connected Multilayer Perceptron (MLP) layers. By automatically extracting features from raw data and employing multiple classifiers, the CNN shows potential for higher efficacy in classifying time-series signals.

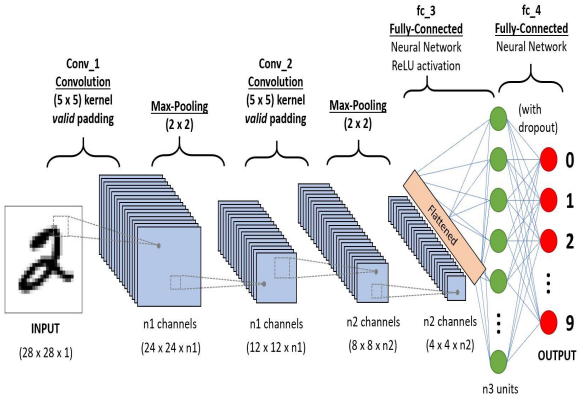

Figure 5. Convolutional Neural Network Architecture with Fully Connected Multilayer Perceptron [31]

The inclusion of Rectified Linear Unit (ReLU) in activation layers accelerates training and addresses vanishing gradients in neural networks. A systematic approach combines convolutional and pooling layers, enhancing training and data representation. Pooling involves subsampling for local responses. This paper introduces a CNN framework that employs one-dimensional (1D) convolutions for sequence processing, aiming to detect pipeline leaks in WDNs. It processes sensor data, identifies anomalies, and predicts leaks in real time, minimizing water loss and environmental impact through feed-forward and backpropagation.

Here are the mathematical equations involved in the prediction and detection process:

1) ReLU Activation Function

The ReLU is a type of activation function that imparts non-linear properties to the CNN. It applies the function $f(x)$ element-wise to the input tensor x. If the

input value x is positive, the function outputs x unchanged; otherwise, it returns zero. This non-linearity, which can be represented by equation (8) helps the CNN learn complex patterns and accelerates training by preventing the vanishing gradient problem during backpropagation.

$$f(x) = \max(0, x) \qquad (8)$$

2) 1D Convolution Operation for Feature Extraction

In the context of leak detection and localization in WDNs, $X[i]$ represents the 1D input tensor (sensor data sequence), and Y $[i]$ denotes the output feature map after applying 1D convolution. The operation involves sliding a 1D kernel of size K (denotes by $W[k]$) along the input sequence and computing the dot product between the kernel and the local region of the sequence at each step. The bias term *b* is added to each output element. This convolution operation, which can be represented by equation (9) helps to extract local patterns and relevant features from the sensor data, potentially indicating the presence of pipeline leaks.

$$Y[i] = \sum_{k=0}^{K-1} X[i+k].W[k] + b \qquad (9)$$

3) Max Pooling Operation

After each convolutional layer, a pooling layer is applied to reduce spatial dimensions and computational complexity. In leak detection within WDNs, $X[i, j]$ represents the 2D feature map, and $Y[i, j]$ denotes the pooled output feature map. The pooling operation subsamples the feature maps by selecting the maximum value within localized 2x2 neighborhoods. Equation (10) represents a powerful feature selection process that retains the most salient features while discarding irrelevant information, thereby enhancing leak detection and localization.

.

$$Y[i, j] = \max\begin{pmatrix} X[2i, 2j],\ X[2i, 2j+1], \\ X[2i+1, 2j], X[2i+1, 2j+1] \end{pmatrix} \qquad (10)$$

4) Average Pooling Operation

Alternatively, average pooling is used to reduce spatial dimensions. The average pooling operation calculates the average of the values in localized 2x2 neighborhoods of the feature maps. It smooths the representations and further assists in detecting and localizing leaks by focusing on important regions. This operation is represented by equation (11):

$$Y[i, j] = \frac{1}{4}\begin{pmatrix} X[2i, 2j] + X[2i, 2j+1] + \\ X[2i+1, 2j] + X[2i+1, 2j+1] \end{pmatrix} \qquad (11)$$

5) Backpropagation for Weight Updates

Backpropagation is used during training to update the weights W and biases b of the CNN. The gradients $\frac{\partial Low}{\partial W}$ and $\frac{\partial Loss}{\partial b}$ represent the sensitivity of the loss

function with respect to the weights and biases, respectively. They are calculated based on the loss $(Loss)$and the output feature map (Y). By iteratively adjusting the weights and biases using these gradients, the CNN learns to better predict and detect pipeline leaks based on the input sensor data. This update process is represented by equation (12):

$$\frac{\partial Loss}{\partial W} = \frac{\partial Loss}{\partial Y} \cdot \frac{\partial}{\partial W}$$
$$\frac{\partial Loss}{\partial b} = \frac{\partial Loss}{\partial Y} \cdot \frac{\partial Y}{\partial b} \qquad (12)$$

The presented mathematical equations play a pivotal role in the pipeline leak detection and localization process when utilizing CNNs. They facilitate the network's ability to process sensor data with efficiency, extracting pertinent features, and making real-time predictions regarding the occurrence and location of leaks. As a result, these contributions lead to enhanced operational efficiency and reduced water loss in WDNs.

D. SVM-CNN-GT Algorithm

The proposed SVM-CNN-GT algorithm optimizes leak detection and localization accuracy by integrating SVM, CNN, and GT. In addition, the proposed algorithm minimizes distance errors and optimizes sensor placement in WDNs (see Algorithm 1).

Algorithm 1: Proposed SVM-CNN-GT algorithm

1. **Input:**
   Define the graph representing the pipeline network with nodes and edges.
   Set the edge weights to represent the pipe lengths between nodes.
   Define the set of neighbouring nodes for each node.
   Define the data matrix (X) containing input data variables for SVM training and prediction.
   Define the label vector (Y) containing labels for SVM training.
2. **Output:**
   Estimated leak locations and starting virtual nodes from the graph-based localization algorithm.
   SVM models for binary/multiclass leak detection and leak severity prediction.
   CNN model for classifying time-series signals to detect pipeline leaks.
3. **For each** node in the graph, **do**:
4. **For each** neighbouring node in the neighbourhood of the current node, **do**:
5. Calculate the cost function (Eq. 1) for the current virtual node using timestamps and virtually generated leakage signals.
6. Find the virtual node (starting location) that minimizes the cost function using argmin (Eq. 2).
7. Implement Dijkstra's algorithm to estimate the temporary leak location on the graph's edge using the selected virtual node.
8. Set the nearest linked node to the temporary leak

location as the starting virtual node (V).
9. Prepare the data matrix (X) with input data variables for SVM training and prediction.
10. Prepare the label vector (Y) representing the presence (1) or absence (-1) of leaks in the WDN data.
11. Train the SVM model for binary leak detection (Eq. 5) using the data matrix (X) and label vector (Y).
12. Train the SVM model for multiclass leak detection (Eq. 6) using the data matrix (X) and label vector (Y) if applicable.
13. Train the SVM model for leak severity prediction (Eq. 7) using the data matrix (X) and label vector (Y) if applicable.
14. Define the CNN architecture with convolutional and pooling layers.
15. Use ReLU activation function (Eq. 8) in the CNN to introduce non-linearity and prevent vanishing gradients.
16. **For each** data point in the input sequence, do:
17. Apply 1D convolution (Eq. 9) with sliding filters to extract local patterns and features.
18. After each convolutional layer, apply max pooling (Eq. 10) or average pooling (Eq. 11) to reduce spatial dimensions.
19. Implement backpropagation (Eq. 12) during training to update CNN weights and biases for better leak detection performance.
20. Real-time Leak Detection and Localization:
21. Use the graph-based localization algorithm to estimate leak locations with selected virtual nodes.
22. Employ the SVM models for binary/multiclass leak detection and leak severity prediction using input sensor data.
23. Use the trained CNN model to classify time-series signals and detect pipeline leaks.
24. End

The proposed SVM-CNN-GT algorithm is a comprehensive solution aimed at improving the accuracy of leak detection and localization in WDNs. By integrating SVM, CNN, and GT algorithms, this approach offers a holistic approach to address these challenges. The proposed algorithm incorporates graph-based localization techniques to provide precise estimation of leaks and strategically places sensors for effective monitoring. Furthermore, the SVM model is trained to detect binary/multiclass leaks and predict their severity, while the CNN-based signal classification enables real-time detection of pipeline leaks.

By combining these techniques, the proposed algorithm significantly enhances leak detection accuracy, reduces distance errors, and optimizes sensor placement within WDNs. This ultimately leads to improved water conservation, cost efficiency, and positive environmental impact.

D. Simulation Results

The paper proposes the SVM-CNN-GT algorithm, a novel and efficient solution for leak detection in WDNs that achieves superior precision, optimal sensor placement, and reduced energy consumption. This algorithm merges SVM, CNN, and GT to attain these goals effectively. SVM within the SVM-CNN-GT algorithm executes binary and/or multiclass leak detection and severity prediction

tasks. By pinpointing and categorizing leaks based on input data variables, SVM augments the algorithm's overall effectiveness.

Meanwhile, CNN undertakes a crucial role in classifying time-series signals, particularly in leak identification. Despite its visual data focus, CNN adeptly processes sequential data, extracting local patterns and leak-related features.

GT serves to estimate leak locations using a graph representation of the pipeline network. By evaluating a cost function for virtual nodes using timestamps and generated leakage signals, the algorithm identifies neighboring nodes and gauges temporary leak positions on the graph's edges through Dijkstra's algorithm. This strategy optimizes sensor placement, curtails distance errors, and reduces computational load, ensuring precise and efficient leak localization in WDNs.

The SVM-CNN-GT algorithm provides a holistic solution to address diverse facets of leak detection and optimization in WDNs. Its efficacy is validated via multiple EPANET simulations, where average outcomes affirm its effectiveness.

EPANET is selected due to its widespread adoption as WDNs simulation platform. As a favored open-source software, EPANET's event-driven approach, multiple interfaces, and efficient C++ protocol implementation offer notable advantages for WDN simulations. Its tailored Graphic User Interface (GUI) and functionalities make it an invaluable resource for ML model training. The user-friendly setup of EPANET-2 further sets it apart from other network simulators like NS-2, OMNET, and OPNET.

Table I provides the values for various simulation parameters used in the implementation of the proposed algorithm. The simulation parameters include the MAC Protocol IEEE 802.11ah, the number of 68 nodes, the total simulation duration is 3600 seconds, the accuracy tolerance of 0.001, flow change tolerance of 0.0001, pressure units is PSI, quality time step is 1 second, head loss formula is Hazen-Williams, and reporting step is 30 seconds.

Table 1. Simulation Parameters

Parameters	Values
MAC Protocol	IEEE 802.11ah
Number of Nodes	68
Total simulation duration in seconds	3600 seconds
Accuracy Tolerance	0.001
Flow Change Tolerance	0.0001
Pressure Units	PSI
Quality Time Step	1 Second
Flow Change Tolerance	0.0001
Head Loss Formula	Hazen-Williams
Reporting steps in seconds	30 seconds
Architecture	IoT-based WSNs for WDNs
Sensor Types	Pressure, Vibration, Flow
Sensor Data Analysis	Real-time
IoT Gateway	Enables data transmission
Cloud Platforms	Efficient data analysis and management

In Figure. 6, you can observe the schematic diagram of the WDNs case study, which is denoted as case study network 1 in this investigation. This network is composed of one supply node (referred to as the tank node) and 45 demand or

load nodes (referred to as non-tank nodes). Specifically, node 1 corresponds to the supply node, while nodes 2 through 46 represent the load nodes. These nodes are interconnected by pipes of varying lengths and diameters. For detailed data regarding each pipe and node within this network, consult Tables II.

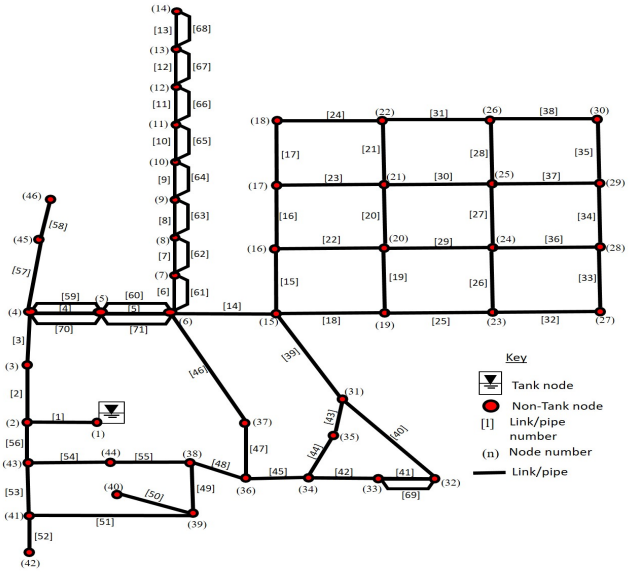

Figure 6. The schematic diagram of the case study network 1 [32]

Table 2. Data Overview of Pipes and Nodes in the Network, including Timestamp, Flow Rate, Pressure, Temperature, and Leak Status

Timestamp	Flow rate (GPM)	Pressure (PSI)	Temperature (°C)	Leak Status
2023/03/05 09:28	10.8	35.9	50.2	1
2023/03/05 09:29	11.1	36.4	49.5	0
2023/03/05 09:30	11.7	35.7	50.0	0
2023/03/05 09:31	12.0	36.0	50.5	0
2023/03/05 09:32	11.3	36.2	50.8	1
2023/03/05 09:33	10.5	35.8	49.8	1
2023/03/05 09:34	10.2	36.3	49.5	0
2023/03/05 09:35	9.9	36.1	50.0	0
2023/03/05 09:36	10.3	35.9	50.5	1
2023/03/05 09:37	11.0	36.4	50.2	1
2023/03/05 09:38	11.5	35.7	50.0	0
2023/03/05 09:39	12.2	36.1	50.8	0
2023/03/05 09:40	11.5	36.3	51.0	1
2023/03/05 09:41	10.8	35.9	50.2	1
2023/03/05 09:42	11.1	36.4	49.5	0
2023/03/05 09:43	11.7	35.7	50.0	0
2023/03/05 09:44	12.0	36.0	50.5	0
2023/03/05 09:45	11.3	36.2	50.8	1
2023/03/05 09:46	10.5	35.8	49.8	1
2023/03/05 09:47	10.2	36.3	49.5	0
2023/03/05 09:48	9.9	36.1	50.0	0
2023/0/-05 09:49	10.3	35.9	50.5	1
2023/03/05 09:50	11.0	36.4	50.2	,1
2023/03/05 09:51	11.5	35.7	50.0	0

Table 2 presents the data associated with pipes and nodes in the network, including information such as timestamp, flow rate, pressure, temperature, and leak status. It provides a comprehensive overview of the network's hydraulic behavior and potential anomalies.

A. Water Loss Volumes

In Figure. 7, the leakage profiles of nodes within case study network 1 are visualized, encompassing the supply node. Noteworthy are nodes 5, 30, and 40, which manifest the highest leakage outflows, signifying their pivotal role in the network's criticality. To tackle this concern, employing pressure control strategies is advisable. The objective is to mitigate leakage outflows from these nodes and curtail overall network leakage. The essential pipes connected to these pivotal nodes are as follows: node 5 is linked to pipes 4, 5, 59, 60, 70, and 71; node 6 is associated with pipes 5, 6, 14, 46, 60, 61, and 71; and node 41 connects to pipes 51, 52, and 53.

For pipes connected to these nodes exhibiting a notably high leakage flow rate, it is advisable to consider implementing pressure control measures at one or both of their endpoints. This approach will help mitigate the leakage problem and enhance the overall performance of the network.

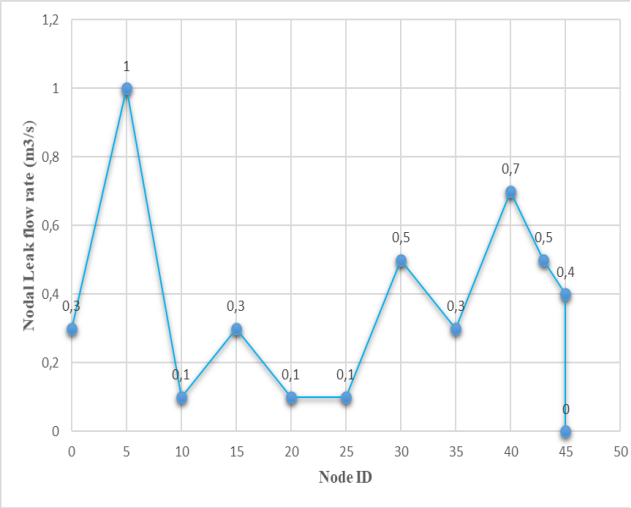

Figure 7. Water Loss Volumes per Pipe in the Case Study Network

B. Leak detection accuracy

Water leak detection accuracy as illustrate is the measure of a system's ability to reliably and accurately identify the presence and location of leaks in WDNs. In a study, the results of three algorithms, namely SVM-CNN-GT, CNN, and SVM, were compared. The average leak detection accuracies achieved by these algorithms were 98%, 82%, and 78% respectively, as illustrated in Figure 8. Notably, the SVM-CNN-GT algorithm outperformed both the SVM and CNN algorithms in terms of leak detection accuracy.

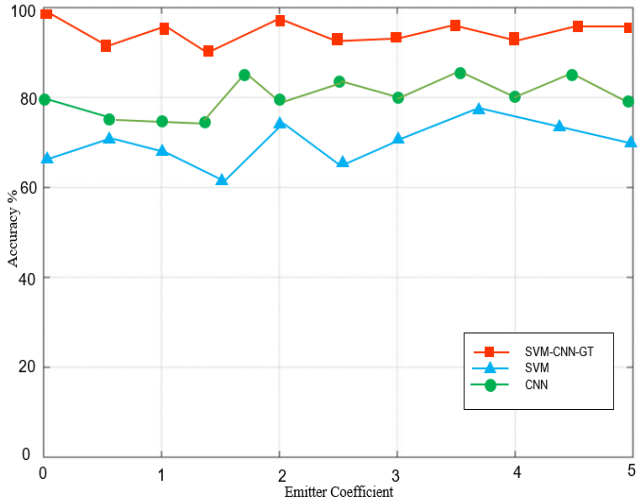

Figure 8. Accuracy of Leak detection for the Case Study Network 1

The SVM-CNN-GT algorithm's strategic sensor placement, which incorporates GT principles, is responsible for this exceptional performance. . In contrast, the SVM and CNN algorithms, lacking the ability to strategically place sensors, achieved inferior water leak detection accuracy.

E. Conclusion

The SVM-CNN-GT algorithm is proposed in this paper with the objective of improving leak detection accuracy while optimizing sensor placement in WDNs. The proposed algorithm uses SVM for leakage detection and prediction by analyzing various data variables such as water pressure, flow rates, temperature, and leak information. Furthermore, CNN is utilized for time-series signal classification to effectively identify pipeline leaks in WDNs. In addition, the proposed algorithm uses GT, which optimizes leak location estimates and sensor placement, enabling the identification of areas prone to leaks. This optimization process not only reduces installation costs but also minimizes potential infrastructure and environmental damage. EPANET simulations have demonstrated the effectiveness of the proposed algorithm in reducing sensor placement while improving leak detection accuracy. By strategically placing sensors in areas with high uncertainty, the proposed algorithm is able to minimize the number of sensors needed while maximizing the accuracy of leak detection. This can lead to significant cost savings and improved efficiency for water distribution systems. Future work may involve testing the algorithm with non-ideal sensor conditions, such as sensor failures or malfunctions, to further enhance the accuracy and reliability of the algorithm. Additionally, exploring advanced data analytics tools, such as deep learning, may provide even more accurate and efficient leak detection and localization.

F. Acknowledgment

The authors extend their appreciation to Tshwane University of Technology for their financial support in conducting this research. Furthermore, they affirm that there are no conflicts of interest associated with the publication of this paper.

G. References

[1] Organization, W.H., Water, sanitation, hygiene and health: a primer for health professionals. 2019, World Health Organization.

[2] Organization, W.H., Manganese in drinking water: background document for development of WHO guidelines for drinking-water quality. 2021, World Health Organization.

[3] Komba, G.M., T.E. Mathonsi, and P.A. Owolawi. Water Pipeline Leak Detection and Localisation in Water Distribution Networks. in 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC). 2023. IEEE.

[4] Ali, H. and J.-h. Choi, A review of underground pipeline leakage and sinkhole monitoring methods based on wireless sensor networking. Sustainability, 2019. 11(15): p. 4007.

[5] Ociepa, E., M. Mrowiec, and I. Deska, Analysis of water losses and assessment of initiatives aimed at their reduction in selected water supply systems. Water, 2019. 11(5): p. 1037.

[6] Zaman, D., et al., A review of leakage detection strategies for pressurised pipeline in steady-state. Engineering Failure Analysis, 2020. 109: p. 104264.

[7] Velayudhan, N.K., et al., IoT-enabled water distribution systems-a comparative technological review. IEEE Access, 2022.

[8] Murugan, S.S. and S. Chandran, Assessment of Non-revenue water in a water distribution system and strategies to manage the water supply. Assessment, 2019. 6(04).

[9] Murugan, S.S. and P. Selvakumar, Management of non revenue water in a water Distribution system for a Municipality in tamilnadu. Water and Energy International, 2021. 63(11): p. 8-12.

[10] Letley, G. and J. Turpie, Water Research Commission. 2023.

[11] du Plessis, A. and A. du Plessis, Evaluation of Southern and South Africa's freshwater resources. Water as an Inescapable Risk: Current Global Water Availability, Quality and Risks with a Specific Focus on South Africa, 2019: p. 147-172.

[12] Al-Washali, T., et al., Monitoring nonrevenue water performance in intermittent supply. Water, 2019. 11(6): p. 1220.

[13] Gebrehiyot, T., Assessing Water Supply Coverage and Water Losses in Distribution System: A Case Study of Debre Birhan Town, Ethiopia. Unpublished MSc Thesis, 2015.

[14] Gumbi, N. and M. Rangongo. Factors that Hinder Effective Management and the Supply of Clean Potable Water at eThekwini Municipality in KwaZulu-Natal. 2018. International Conference on Public Administration and Development ….

[15] Poona, V.A., Non-revenue water reduction programmes funded by the private sector to solve under staffing at Kwa-Zulu Natal's municipalities. 2018.

[16] Vlasa, I., et al., Smart metering systems optimization for non-technical losses reduction and consumption recording operation improvement in electricity sector. Sensors, 2020. 20(10): p. 2947.

[17] Musse, S.A., Exploring the cornerstone factors that cause water scarcity in some parts of Africa, possible adaptation strategies and a quest in food security. 2018.
[18] Organization, W.H., Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs. 2021.
[19] WHO, U., UNFPA, World Bank Group, UNPD. Trends in maternal mortality 2000 to 2017. UK: World Health Organization (WHO), United Nations Children's Fund (UNICEF), United Nations Population Fund. 2019, UNFPA), World Bank Group, United Nations Population Division.
[20] Leal Filho, W., et al., Understanding responses to climate-related water scarcity in Africa. Science of the Total Environment, 2022. 806: p. 150420.
[21] McGranahan, G., Demand-side water strategies and the urban poor. 2002: IIED.
[22] Nsanzubuhoro, R., Pressure-based leakage characterisation of bulk pipelines. 2019.
[23] Chan, T.K., C.S. Chin, and X. Zhong, Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. Ieee Access, 2018. 6: p. 78846-78867.
[24] Cody, R., Acoustic Monitoring for Leaks in Water Distribution Networks. 2020.
[25] Adedeji, K.B., Y. Hamam, and A.M. Abu-Mahfouz, Impact of pressure-driven demand on background leakage estimation in water supply networks. Water, 2019. 11(8): p. 1600.
[26] Nie, X., et al., Big data analytics and IoT in operation safety management in under water management. Computer Communications, 2020. 154: p. 188-196.
[27] Wan, X., et al., Literature review of data analytics for leak detection in water distribution networks: A focus on pressure and flow smart sensors. Journal of Water Resources Planning and Management, 2022. 148(10): p. 03122002.
[28] Daniel, I., et al., A sequential pressure-based algorithm for data-driven leakage identification and model-based localization in water distribution networks. Journal of Water Resources Planning and Management, 2022. 148(6): p. 04022025.
[29] Chew, A.W.Z., et al., Generalized Acoustic Data Analysis Framework for Leakage Detection and Localization in Field Operational Water Distribution Networks. Journal of Water Resources Planning and Management, 2023. 149(11): p. 04023056.
[30] Al-Ali, M.S., Non-Revenue Water: Methodological Comparative Assessment. 2022.
[31] Hu, Z., et al., Review of model-based and data-driven approaches for leak detection and location in water distribution systems. Water Supply, 2021. 21(7): p. 3282-3306.
[32] Liu, Y., et al., Water pipeline leakage detection based on machine learning and wireless sensor networks. Sensors, 2019. 19(23): p. 5086.
[33] Porwal, S., S. Akbar, and S. Jain. Leakage detection and prediction of location in a smart water grid using SVM classification. in 2017 International

Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). 2017. IEEE.
[34] Zhou, M., et al., An integration method using kernel principal component analysis and cascade support vector data description for pipeline leak detection with multiple operating modes. Processes, 2019. 7(10): p. 648.
[35] Lang, X., et al., Leak detection and location of pipelines based on LMD and least squares twin support vector machine. IEEE Access, 2017. 5: p. 8659-8668.
[36] Lučin, I., et al., Data-driven leak localization in urban water distribution networks using big data for random forest classifier. Mathematics, 2021. 9(6): p. 672.
[37] Rajabi, M.M., et al., Leak detection and localization in water distribution networks using conditional deep convolutional generative adversarial networks. Water Research, 2023. 238: p. 120012.
[38] Majeed, A. and I. Rauf, Graph theory: A comprehensive survey about graph theory applications in computer science and social networks. Inventions, 2020. 5(1): p. 10.
[39] Komba, G.M., O.P. Kogeda, and T. Zuva. A new gateway location protocol for mesh networks. in Proceedings of the World Congress on Engineering and Computer Science. 2014.
[40] Sen, P.C., M. Hajra, and M. Ghosh. Supervised classification algorithms in machine learning: A survey and review. in Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018. 2020. Springer.
[41] Muthukumar, V., et al., Classification vs regression in overparameterized regimes: Does the loss function matter? The Journal of Machine Learning Research, 2021. 22(1): p. 10104-10172.
[42] Otchere, D.A., et al., Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. Journal of Petroleum Science and Engineering, 2021. 200: p. 108182.
[43] KUMAR, S., SUPPORT VECTOR MACHINE AS A CLASSIFIER FOR FEATURE-BASED CLASSIFICATION: A TECHNICAL NOTE. Methodologies and Applications for Analytical and Physical Chemistry, 2018: p. 347.
[44] Kumar, H.H. A novel approach of SVM based classification on thyroid disease stage detection. in 2020 third international conference on smart systems and inventive technology (ICSSIT). 2020. IEEE.
[45] Mohan, L., et al. Support vector machine accuracy improvement with classification. in 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN). 2020. IEEE.
[46] Omar, I., M. Khan, and A. Starr, Suitability analysis of machine learning algorithms for crack growth prediction based on dynamic response data. Sensors, 2023. 23(3): p. 1074.
[47] Khan, A., et al., A survey of the recent architectures of deep convolutional neural networks. Artificial intelligence review, 2020. 53: p. 5455-5516.
[48] Abiodun, O.I., et al., Comprehensive review of artificial neural network applications to pattern recognition. IEEE access, 2019. 7: p. 158820-158846.

[49] Agarwal, S., J.O.D. Terrail, and F. Jurie, Recent advances in object detection in the age of deep convolutional neural networks. arXiv preprint arXiv:1809.03193, 2018.
[50] Alzubaidi, L., et al., Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 2021. 8: p. 1-74.
[51] Kattenborn, T., et al., Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 2021. 173: p. 24-49.
[52] Liu, L., et al., Deep learning for generic object detection: A survey. International journal of computer vision, 2020. 128: p. 261-318.