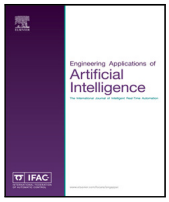
Engineering Applications of Artificial Intelligence 126 (2023) 107137

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



P^{2}CA-GAM-ID: Coupling of probabilistic principal components analysis with generalised additive model to predict the $k-$barriers for intrusion detection

Abhilash Singh [a,*], Jaiprakash Nagar [b,1], J. Amutha [c], Sandeep Sharma [d,*]

[a] Fluvial Geomorphology and Remote Sensing Laboratory, Indian Institute of Science Education and Research Bhopal, India
[b] Subir Chowdhury School of Quality and Reliability, Indian Institute of Technology Kharagpur, India
[c] Department of Artificial Intelligence and Data Science, E.G.S. Pillay Engineering College, TamilNadu, India
[d] Department of Higher and Technical Education, Government of Jharkhand, Ranchi, India

ARTICLE INFO

Keywords:
Machine learning
Intrusion detection
WSNs
Probabilistic PCA
GAM

ABSTRACT

Drastic advancement in computing technology and the dramatic increase in the usage of explainable machine learning algorithms provide a promising platform for developing robust intrusion detection algorithms. However, the development of these algorithms is constrained by their applicability over specific scenarios of Wireless Sensor Networks (WSNs). We introduced a hybrid framework by combining Probabilistic Principal Component Analysis (P^{2}CA) and Generalised Additive Model (GAM), which is performing well for all the scenarios of WSNs. To demonstrate our framework's broad applicability, we evaluated its performance over three publicly available intrusion detection datasets (*i.e.,* LT-FS-ID, AutoML-ID, and FF-ANN-ID), each from different scenarios. Our findings highlight that the presented framework can accurately predict the number of $k-$barriers for all three datasets. Furthermore, we conducted a comprehensive performance comparison between our proposed framework and benchmark algorithms, which revealed that our approach outperforms all of them. Additionally, we evaluated the framework's versatility by testing its performance on datasets unrelated to intrusion detection, specifically ALE datasets. Notably, our approach accurately predicted the response variable in these datasets and exceeded the performance of its primary algorithm, further demonstrating its robustness and adaptability.

The implications of this research are substantial. By developing a robust intrusion detection framework that performs well across diverse WSN scenarios, we address a critical need for reliable network security in various domains, including industrial IoT, smart cities, and environmental monitoring. Our findings not only enhance the understanding of intrusion detection in WSNs but also pave the way for developing more sophisticated and adaptable systems to safeguard sensitive data and critical infrastructure.

1. Introduction

We live in a world facing political instability and unrest, causing insecurity among the people. The hunger for political power, geographically essential regions, and control over others has made people usurp. Therefore, securing national boundaries against any potential attack of enemy forces, intrusion, or unauthorised entry is one of the governments' critical issues, requiring the concerned authorities' immediate attention. A country may share borders with neighbouring nations, extending over thousands of kilometers, necessitating continuous monitoring. Several countries do not have regular armies to guard their borders or inhabitants in the proximity of their international boundaries. In addition, no country can establish checkpoints at every location along the border area; thus, a vast region between the checkpoints and the boundary lines between the two countries remains unguarded (Singh et al., 2022c). Further, patrolling methods are conventional, limited, and periodic, resulting in unattended borders for the long haul. These inadequate security measures would invite enemy forces apparently to capture and control some geographically significant regions. Further, enemies may intrude in restricted regions to steal highly secret information or demolish some crucial military or civilian establishments, which may substantially harm the country. Therefore, intrusion detection at borders and around some crucial establishments has become a country's top priority (Singh et al., 2022b).

Wireless Sensor Networks (WSNs) can address this problem effectively. A WSN may comprise thousands of minuscule, affordable Sensor Nodes (SNs) which do not demand any pre-installed foundation

* Corresponding authors.
E-mail addresses: abhilash.iiserb@gmail.com (A. Singh), sandeepsvce@gmail.com (S. Sharma).
[1] Now at Eurecom, SophiaTech, Saint Antipolish, 06410, Biot, France.

https://doi.org/10.1016/j.engappai.2023.107137
Received 30 January 2023; Received in revised form 31 August 2023; Accepted 9 September 2023
Available online 22 September 2023

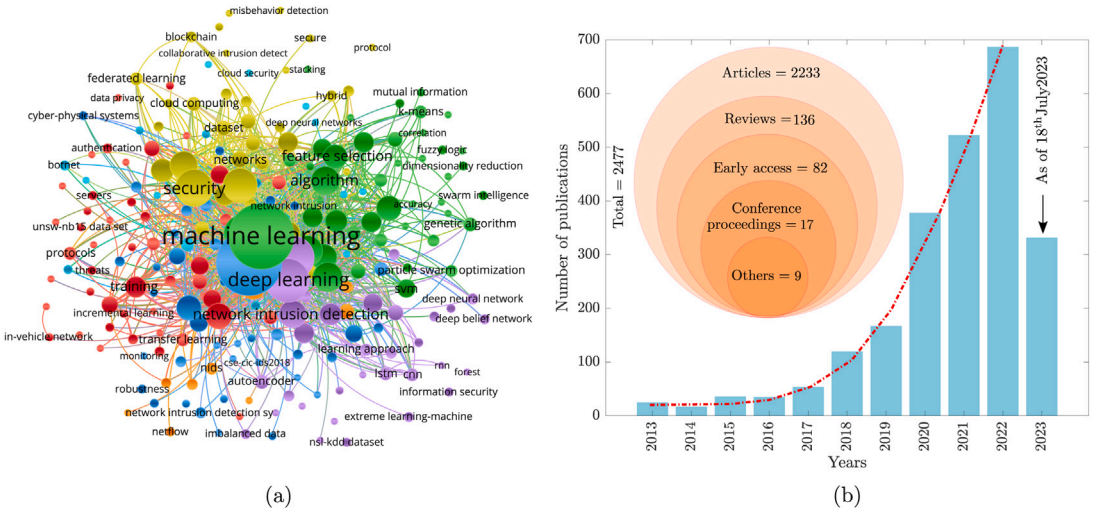


Fig. 1. Bibliometric analysis of the keywords 'Intrusion Detection' and 'Machine Learning.' Figure (a) illustrates the burst analysis of keywords used by authors in research papers published in the Web of Science from 2013 to 2023 (up until July 18th 2023). The circles, distinguished by various colours, represent different clusters of relevant keywords, with the circle's diameter indicating their frequency of appearance. Figure (b) demonstrates the publication trend over the past ten years regarding the application of machine learning in intrusion detection.

and functions autonomously in a decentralised manner (Nagar et al., 2020; Singh et al., 2021b; Kotiyal et al., 2021). Therefore, WSNs are highly in demand for monitoring, surveillance, intrusion detection, and reconnaissance purposes along international borders (Bhadwal et al., 2019; Arjun et al., 2019; Singh and Singh, 2021; Sood et al., 2022; Shukla et al., 2023). In addition, SNs are cheap, demand less power, are widely available, and quickly installable in emergency conditions where human intervention is almost negligible; therefore, WSNs also have many civilian applications such as industrial monitoring, precision agriculture, forest fire detection, health monitoring, remote landslides detection, structural health monitoring, and several others (Noel et al., 2017; Aponte-Luis et al., 2018; Nagar and Sharma, 2018; Ghosh et al., 2018; Singh et al., 2019; Kumar et al., 2020).

Surveillance, monitoring, and intrusion at international borders and unauthorised access to prohibited regions like no man's territories, crucial establishments, and military bases *etc.*, can be resolved by deploying an effective WSN. Researchers across the globe have proposed various algorithms and analytical frameworks to identify a potential intruder and alert the concerned authorities before it harms (Karthick et al., 2019; Arjun et al., 2019; Benahmed and Benahmed, 2019; Sharma and Nagar, 2020; Amutha et al., 2021; Karanja and Badru, 2021). One of the significant issues with these frameworks is that they need to be validated either through simulation runs or by actual deployment in a given Region of Interest (RoI). Validating analytical models via simulation runs is time-consuming, since it takes several hours to obtain a single result for a defined parameter set. Further, as the size of the network rises, the simulation time grows exponentially in terms of the size of the RoI, the number of SNs, and the transmission/sensing range of SNs. Further, practical deployment of WSNs is expensive, and money is an asset that is hard to get, if not impossible. One of the possible solutions to resolve the high simulation time issue is to use Machine Learning (ML) based approaches to predict the WSNs performance metrics such as the number of barriers, intrusion detection probability, coverage, connectivity, and so on. To gain insights into the current trends of machine learning for intrusion detection, we conducted a comprehensive bibliometric analysis of research papers published in Web of Science (WoS) over the past ten years (Singh et al., 2023a). Our analysis revealed a total of 2477 research papers, consisting of 2223 research articles, 136 review articles, 82 early access papers, 17 conference proceedings, and 9 other types of publications (Fig. 1). Remarkably, we observed an exponential surge in the number of publications dedicated to machine learning for intrusion detection, with a remarkable 687 publications occurring solely in 2022. These findings underscore the paramount importance of studying and exploring this subject matter. In Section 2, we discuss various ML-based approaches to predict several performance metrics, especially the $k-$barrier and $k-$barrier coverage probability rendered by a WSN for intrusion detection.

2. Related works

Surveillance and monitoring of crucial regions have become essential in today's scenario and can be addressed by deploying WSNs. The performance of the deployed WSNs can be measured in terms of $k-$barrier coverage probability, which is one of the crucial metrics for WSNs. A WSN is assumed to render $k-$barrier coverage if every possible path from the point of intrusion to the destination is covered by at least k distinct sensor nodes cumulatively, thus forming a $k-$barrier path (Keung et al., 2012). The researchers have proposed various ML algorithms to accurately map the $k-$barrier and $k-$barrier coverage probability that is used for intrusion detection and prevention (Fig. 2). Algorithms for accurately predicting the $k-$ barriers include LT-FS-ID algorithm (Singh et al., 2022c), Automated Machine Learning (AutoML) (Singh et al., 2022b) algorithm, FF-ANN-ID algorithm (Singh et al., 2022a), GPR-ID algorithm (Singh et al., 2021a), ANN model (Arora and Pal, 2022), and EFNNs algorithm (de Campos Souza et al., 2022).

Singh et al. (2022c) introduced an algorithm that relies on log transformation and scaling of the input features. They consider four features: area of the RoI, sensing range, transmission range, and the number of sensors for accurately mapping the k number of barriers in a rectangular RoI by considering uniform sensor deployment. In addition, they evaluated the relative importance score and feature sensitivity of all the features by employing the regression tree ensemble approach and Partial Dependency Plot (PDP) analysis, respectively. They reported that the LT-FS-SVR algorithm estimates the number of barriers with R = 0.98, RMSE = 6.47, and bias = 12.35. However, the major limitation associated with the LT-FS-ID algorithm is that it considers only positive real numbers to be used as input predictors. To overcome this limitation, recently, Singh et al. (2022b) have proposed an automated machine learning algorithm (*i.e.,* AutoML-ID) to precisely estimate the $k-$barriers in a rectangular RoI considering Gaussian node deployment. The proposed algorithm automatically determines the best ML model from the set of ML algorithms [Support Vector

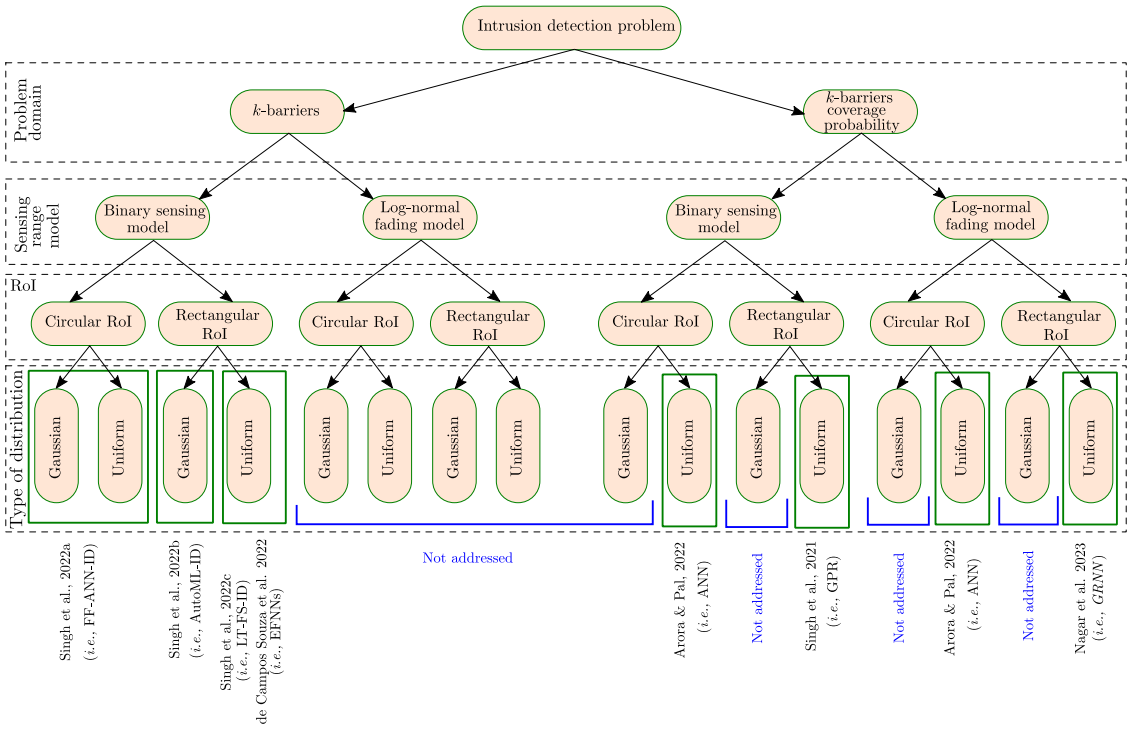


Fig. 2. Current state-of-the-art machine learning algorithms for solving intrusion detection problem domain.

Regression (SVR), Gaussian Process Regression (GPR), Binary Decision Tree (BDT), Random Forest (RF), Boosting Ensemble Learning (B-EL), and Linear Regression (LR)]. They employed a robust search strategy (Bayesian optimisation) which explored different combinations of algorithms and their hyperparameters. The performance of different algorithms is compared using the evaluation metrics (R, RMSE, and bias). Finally, the algorithms are ranked based on their performance, and the best-performing algorithms are selected as candidates for further optimisation. After the hyperparameter optimisation process, the algorithm that achieves the best performance on the evaluation metrics was selected. They reported that among all the algorithms involved in the AutoML-ID, GPR emerges as the best-performing algorithm with R = 1, RMSE = 0.007, and bias = −0.006. Although the AutoML approach yields the best results, the practical implementation of AutoML is very difficult. The recently proposed FF-ANN-ID model overcomes this limitation (Singh et al., 2022a). They trained and analysed the FF-ANN-ID model for a circular RoI by considering both uniform and Gaussian sensor distribution. They stated that the model estimates the $k-$barriers with R = 0.79 and RMSE = 48.36 for uniform distribution and with R = 0.78, RMSE = 41.1 for Gaussian distribution. The limitation associated with the FF-ANN-ID algorithm is that it fails to solve the problem of data stream regression issue. More recently, de Campos Souza et al. (2022) developed an Evolving Fuzzy Neural Networks (EFNNs) that solves the data stream regression issues, along with the prediction of the k-barriers in WSNs to detect unauthorised access. This system implements only if-then rules in the fuzzy system to estimate the number of barriers. To evaluate the proposed method's effectiveness, they compared it with existing evolving methods through empirical evaluations. The results highlight the superior performance of the EFNNs, as they demonstrate significantly lower RMSE values when tested on separate data sets. Furthermore, the evaluation includes a stream-based interleaved-predict-and-then-update procedure, further validating the proposed approach's efficacy. Recently, Muruganandam et al. (2023) tested the potential of a feed-forward neural network in accurately predicting the $k-$barriers on LT-FS-ID datasets (Singh et al., 2022c). They found that the feed-forward neural network accurately predicts the $k-$barriers with R = 0.95 and RMSE = 6.15. The achieved high correlation coefficient and relatively low RMSE indicate the effectiveness of the model in capturing the underlying patterns and providing accurate predictions. These results contribute to the growing body of literature on employing neural networks for barrier prediction tasks and emphasise their potential as a valuable tool in this domain.

Apart from developing ML models for predicting $k-$barriers, the researchers have also proposed several ML models for predicting $k-$barriers coverage probability. Recently, Singh et al. (2021a) developed three ML methods based on GPR algorithms [*i.e.,* scale-GPR (S-GPR), centre-mean-GPR (C-GPR), and Non-standardise GPR (NS-GPR)] to map the $k-$barrier coverage probability in a rectangular RoI considering Poisson point sensor distribution. They trained these models by using the squared exponential kernel and then evaluated their performance over the testing dataset. They reported that the NS-GPR outperforms all the other variants with R = 0.85 and RMSE = 0.095. More recently, Arora and Pal (2022) proposed an ANN-based architecture to predict the $k-$barriers coverage probability. In addition, they have also considered the Boundary Effects (BEs) into account to incorporate the shadowed environments and estimate the $k-$coverage probability with R = 0.98 and RMSE = 0.07 and outperform the result of Adaptive Neuro-Fuzzy Inference System (ANFIS) in terms of accuracy. Recently, Nagar et al. (2023) also proposed a Generalised Regression Neural Network (GRNN) based approach for predicting $k-$barriers coverage probability by considering BEs and Shadowing Effects (SEs). They considered six features: length, breadth, number of sensors, sensing range of sensors, required k, and standard deviation of SEs. They found that the proposed model accurately predicts the $k-$barriers coverage probability with R = 0.78 and RMSE = 0.14. The inclusion of BEs and SEs in the model allows for a more comprehensive understanding of the factors influencing barrier coverage probability. By considering these effects and incorporating relevant features, the GRNN-based approach proposed by Nagar et al. (2023) shows promise in accurately estimating the coverage probability of $k-$barriers. These findings contribute to advancing machine learning techniques in barrier prediction tasks and provide valuable insights into improving wireless sensor systems' performance.

The major issue associated with the work discussed above is the demand for new algorithms for each different scenario of WSNs. This paper proposes a novel ML-based approach for accurately predicting $k-$barriers to ensure fast intrusion detection and prevention by coupling probabilistic principal component analysis and generalised additive model. The proposed framework solves the problem of having different machine learning algorithms for different scenarios of WSNs based on RoI (circular or rectangular) and node deployment (uniform or Gaussian). In doing so, we trained the hybrid algorithms on three publicly available intrusion detection datasets. All these datasets use network and sensor properties as features for accurate mapping of $k-$barriers using regression-based machine learning. Finally, we evaluated its performance by considering the performance indicators, such as R, RMSE, and bias. By addressing the challenge of diverse WSN scenarios, our proposed approach demonstrates the potential to enhance intrusion detection and prevention. Through empirical evaluation on publicly available datasets, we aim to establish the effectiveness and robustness of our framework. The performance indicators provide quantitative insights into the accuracy and reliability of our model's predictions.

3. Datasets

We considered three intrusion datasets (LT-FS-ID, AutoML-ID, and FF-ANN-ID) to evaluate the performance of the proposed hybrid algorithm (Singh et al., 2022c,b,a). In addition to addressing the $k-$barriers intrusion detection problem, these datasets are employed to evaluate the performance of the newly proposed regression algorithms (de Campos Souza et al., 2022). We discussed these datasets in the upcoming subsections.

3.1. LT-FS-ID

The datasets for LT-FS-ID (Singh et al., 2022c) are obtained synthetically by simulations that utilise Network Simulator-2.35 (NS-2.35), which are intended for training and testing purposes. The main advantage of employing NS-2.35 is that it has gained prominence in networking due to its adaptability, scalability, and potential to simulate the algorithm's performance in wired or wireless networks. Hence, to extract the LT-FS-ID datasets, a finite number of sensor nodes, ranging from 100 to 400, are considered that are deployed randomly in the rectangular RoI. Each sensor node is considered homogeneous, meaning its sensing, transmission, and processing capabilities are the same. There exist several sensing and transmission range models in the literature, *viz.*, binary, log-normal, and Elfes model (Hossain et al., 2012; Nagar et al., 2022). A Binary Sensing Model (BSM) assumes that an event/object is sensed by a sensor node if and only if its Euclidean distance is less than or equal to the sensor's sensing range. In other words, a BSM assumes identical received power in all directions, which is not true for real scenarios. In real scenarios, various obstacles exist in the wireless signal propagation environment that cause variations in received signal power. Therefore, the characteristics of a wireless channel keep changing with the change in the signal propagation environment, making it crucial to incorporate the randomness in wireless channel characteristics while considering a sensing range model. The log-normal shadow-fading model incorporates the randomness in wireless channel characteristics denoted by its standard deviation (SD) of shadow-fading. A large value of SD represents large variations in received power and vice-versa. To analyse the performance of WSNs, the BSM is being used as it is very useful for initial mathematical formulations and analysis.

Singh et al. (2022c) used Monte Carlo simulation to extract relevant features, such as the area of RoI, the sensing range, the transmission range of sensor nodes, and the number of sensors from the network parameters. The RoI was varied in the range from 100×50 m^2 to 250×200 m^2. The primary reason for selecting a finite rectangular-shaped RoI was that most real estate in real life is rectangular. The sensing and transmission range of sensor nodes varied from 15 to 40 m and 30 to 80 m, respectively. This is because of the fact that the sensing and transmission range of sensor nodes should be kept less than or equal to half the width of the rectangular shaped RoI to avoid boundary effects. Moreover, a regression ensemble model was developed using boosting ensemble learning to measure the relative importance score of each feature. The sensitivity analysis of each feature was carried out using PDP, following feature scaling to the selected features. This dataset has gained significant recognition among researchers as a valuable resource for validating novel algorithms in k-barrier prediction for fast and accurate intrusion detection (de Campos Souza et al., 2022; Muruganandam et al., 2023).

3.2. AutoML-ID

The Automated Machine Learning (AutoML) model selects various ML models, such as binary decision tree, GPR, bagging ensemble learning, SVR, boosting ensemble learning, kernel regression, and LR model, to predict the number of $k-$barriers. Singh et al. (2022b) used Bayesian Optimisation (BO) to optimise the hyperparameters. They used a synthetic approach to extract the predictor datasets using Monte Carlo simulations, and the entire dataset was generated with the NS-2.35 simulator. Depending upon the application needs, sensor nodes in a given RoI can be distributed either following a uniform or Gaussian distribution model. Therefore, this dataset was obtained by deploying a finite number of homogeneous sensor nodes in a finite rectangular RoI using a Gaussian distribution.

The Gaussian distribution is suited for practical applications and offers distinct capabilities to sensor nodes positioned at various locations. For better-detecting capabilities, more sensor nodes should be deployed. In a Gaussian distributed WSN, more sensor nodes are deployed closer to the central of the RoI denoted by P. In locations far away from P, fewer sensors are deployed, lowering the cost of network deployment (Wang et al., 2012). The Probability Density Function (PDF) is represented as:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{(x-x_i)^2}{2\sigma_x^2} + \frac{(y-y_i)^2}{2\sigma_y^2}\right)} \tag{1}$$

where $(x_i,\ y_i)$ represents the deployment point, σ_x and σ_y are the standard deviations for x and y dimensions, respectively. The objective function *(f)* in BO is determined using the Gaussian Process (GP) as:

$$f(x) \sim GP(\rho(x), \tau(x_i, x_j)) \tag{2}$$

where ρ and τ are calculated from the observations of x. BSM was used to evaluate the WSN's performance. In BSM, if the target is present inside the sensor's sensing range, it can be detected by any sensor at random. Otherwise, the probability of detecting the target will be zero. Furthermore, the relevancy of the selected predictors, namely, the area of the RoI, sensing range, transmission range, and the number of sensors in determining the $k-$barriers, was computed by the regression tree ensemble technique by estimating the relative score of each predictor in the AutoML-ID approach.

3.3. FF-ANN-ID

The dataset for Fully Connected Feed Forward Intrusion Detection (FC-FF-ID) is obtained using NS-2.35 (Singh et al., 2022a). Here, a finite number of sensor nodes, ranging from 100 to 400, were assumed to be deployed randomly and uniformly in a finite circular RoI having a radius of R meters. A circular region was another real-life shape we might have seen almost every day in the form of parks, buildings, and playing grounds like cricket stadiums, etc. The circular region's radius was considered to vary from 40 m to 127 m. In this scenario as well,

the sensing range of sensor nodes was assumed to be less than or equal to half the radius of the assumed circular region to avoid the boundary effects. This model considers two sensor nodes distribution models, *viz.*, Gaussian and uniform (Wang et al., 2008). For the Gaussian model, the PDF for a point *(x, y)* deployed with a sensor node in a circular RoI is determined by:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\left(\frac{(x-x_c)^2}{2\sigma_x^2} + \frac{(y-y_c)^2}{2\sigma_y^2}\right)} \tag{3}$$

where (x_c, y_c) represents the centre of the circular RoI, σ_x and σ_y are the standard deviations for x and y dimensions, respectively. Moreover, the location of a sensor node within the circular RoI denoted by (α, ϕ) can also be represented using position coordinates *(x, y)* if

$$f(x,y) : \sqrt{(x-x_c)^2 + (y-y_c)^2} \le R = f(\alpha,\phi) : \alpha \le R \tag{4}$$

For the uniform sensor node distribution model, the PDF for the position of a random sensor node is given by:

$$f_p(\Re) = \begin{cases} 1, if\, P_c \in \Re \\ 0, otherwise \end{cases} \tag{5}$$

where $P_c = (\alpha, \phi)$ represents the position of the node which is randomly deployed in the circular RoI, $\alpha \in [0, R]$, represents the node's distance from the centre of the circular RoI, and $\phi \in [0, 2\pi]$, represents the angle formed by the x-axis and the line passing through the position of the sensor. Moreover, the probability that a node positioned at an arbitrary location $P_c = (\alpha, \phi)$ is determined by:

$$f(P_c) = \frac{1}{\pi R^2} \tag{6}$$

4. Methodology

In this section, we provided an overview of the theoretical background of two key methods: probabilistic principal component analysis and generalised additive model. We explained their individual concepts and characteristics. Next, we explored the coupling of probabilistic principal component analysis and generalised additive model to leverage their combined strengths in capturing complex patterns and relationships in the data. We highlighted how this coupling enhances the modelling capabilities and leads to improved performance. Finally, we delved into the optimisation process of the proposed model, describing the steps involved in fine-tuning the model's parameters and achieving the best possible results. To provide a visual representation of the methodology, please refer to Fig. 3 for a detailed flowchart illustrating the different stages and interactions of the proposed approach.

4.1. Probabilistic Principal Component Analysis (P^2CA)

Tipping and Bishop introduced the P^2CA (Tipping and Bishop, 1999), which is a probabilistic formulation of Principal Component Analysis (PCA) relying on the Gaussian latent variable model. The probability model can quantify the noise level in the observed data. Moreover, it can predict the Principal Components (PCs) in scenarios with missing data (Singh et al., 2023b). P^2CA is rapidly being used in fault detection and data recovery (Ma et al., 2021).

Consider $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ as the feature vectors which are extracted from the variables of interest for the i^{th} subject, $i = 1,2, \ldots, n$. Mathematically, the probabilistic representation of PCA (Suresha and Parthasarathy, 2021; Geraci and Farcomeni, 2016) is given by:

$$y_i = \mu + W u_i + \epsilon_i,\ i = 1, \ldots, n \tag{7}$$

where u_i is a vector of principal components and W is stated as $p \times q$ matrix with elements W_{jh}, $j = 1, \ldots, p$, $h = 1, \ldots, q$. Moreover, u is stochastically independent from ϵ, and is represented as:

$$u_i \sim Y(0, I_q) \tag{8}$$

where I represents q-dimensional identity matrix. If the error is assumed to be zero-centred Gaussian with a covariance matrix Γ, $\epsilon_i \sim Y(0, \Gamma)$, the multivariate distribution is then determined as:

$$y_i \sim Y(\mu, C), C = WW^T + \Gamma \tag{9}$$

4.2. Generalised Additive Model (GAM)

The GAM is a non-parametric regression model based on the combination of the generalised linear model and the additive model. The fundamental principle of this model is to use a connection function that links the dependent variable to the sum of smooth functions that relate to every independent variable. In addition, it can prevent the issue of dimensional disaster and more accurately depict the complex nonlinear relationship among variables (Wang et al., 2021a). Mathematically, the GAM is represented as:

$$g(\mu) = \sum_{j=1}^{p} S_j(X_j) + \alpha \tag{10}$$

where $\mu = E(Y|X_1, X_2, \ldots, X_p)$ is the expectation of Y, $g(.)$ represents the connection function, $S_j(.)$, $j = 1, 2, \ldots, p$ represents the nonparametric smooth function, α is a constant.

To attain a good fit, penalty regression is incorporated into the GAM model. The smooth function's high variability and subsequent model overfit are avoided by adding a penalty term to the basis function's coefficient.

$$l_p(\gamma) = l(\gamma) - \lambda B' S B \tag{11}$$

where γ is the regression coefficient, and S is the penalty matrix. The Eq. (11) obtains the following form (Eq. (12)) if the loss function is set to be the least squares (Tong et al., 2021).

$$l_p(\gamma) = \sum (y - X\gamma)^2 - \lambda B' S B \tag{12}$$

GAMs leverage the Explainable Boosting Machine for interpretation, in which shape functions are ensembles of bagged trees that have been gradient-boosted and act on a single variable. Gradient Boosting Machines (GBMs) construct an additive ensemble model of M size by introducing base learners that perform better than the earlier ones, iteratively enhancing the predictions of y from x in relation to loss function:

$$g_m(x) = g_{m-1}(x) + \rho_m h_m(x) \tag{13}$$

where ρ_m represents the weight of the m^{th} function $h_m(x)$, which serves as the ensemble models (Bentéjac et al., 2021). The gradient boosting algorithm uses decision trees as its basis model to reduce the expected loss function. Some of its metrics include the maximum number of splits per predictor (maxNumSplits) and the number of trees per predictor (numTrees). They are chosen according to the specified task and to deliver a high level of generalisation and accuracy (Konstantinov and Utkin, 2021). The generalisation and accuracy of regularisation techniques introduced for GBMs can be significantly improved through the process of subsampling. Subsampling introduces randomness to the fitting process by using only a random subset of the training data to fit a consecutive base-learner at each learning iteration. The "bag fraction" is a crucial parameter in the subsampling process, indicating the ratio of data utilised at each iteration. It is a positive number that should not exceed one. One of the key advantages of subsampling is that it reduces computational efforts, particularly for large data sets. By using a smaller fraction of the training data at each iteration, the algorithm can adapt to large-scale datasets more efficiently. Additionally, subsampling allows for achieving the desired accuracy with a lower bag fraction, meaning a smaller proportion of the training data is used while increasing the number of base-learners. To support these claims, Natekin and Knoll (2013) and Sutton (2005) have provided further evidence of the effectiveness of subsampling in GBMs. Moreover, the gradient boosting algorithm solves regression issues and can handle complex non-linear function dependencies.

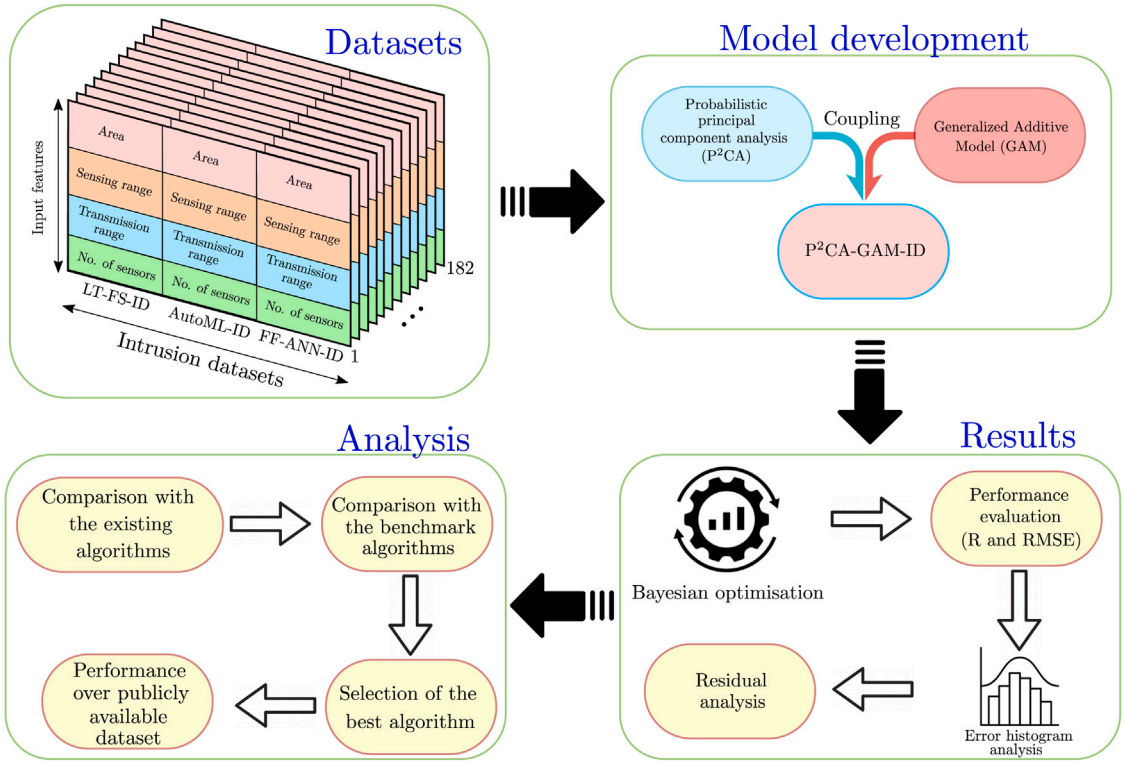


Fig. 3. Flowchart of the methodology.

4.3. Coupling P^2CA and GAM

After standardising the input feature set using the z-score scaling, we feed the transformed features as input to the P^2CA. It performs dimension reduction by leveraging a probabilistic mechanism. We reconstructed the data by considering 90% of the variance. Finally, we fed the P^2CA reconstructed data as an input to the GAM for creating a mapping function between the input and the response variable (Fig. 3).

The coupling of P^2CA and GAM is essential for several reasons. Firstly, P^2CA allows us to capture the underlying probabilistic structure of the data, enabling a more comprehensive understanding of the intrusion patterns and characteristics within the network. This helps in identifying relevant features and reducing the dimensionality of the data, improving the accuracy of prediction models. Secondly, GAM is a powerful statistical modelling technique that allows for flexible modelling of complex relationships between predictors and the response variable. By incorporating GAM, we can capture non-linear relationships, interactions, and complex dependencies within the data. This is particularly beneficial for accurately predicting the $k-$barriers for intrusion detection and prevention, as it enables us to capture the intricate nature of intrusion patterns and their associated factors.

Furthermore, the coupling of P^2CA and GAM leverages the complementary strengths of both methods. P^2CA aids in feature extraction and dimensionality reduction, while GAM provides a flexible modelling framework for accurately capturing complex relationships. This combination enhances the predictive capabilities of the model and improves the accuracy of predicting the $k-$barriers.

4.4. Bayesian Optimisation (BO)

It is a very efficient optimisation approach that integrates prior knowledge of the unknown function (η) with sampling points (p) to acquire posterior information about the function distribution by using the Bayesian algorithm. The global optimal solution is determined by this posterior information (Shi et al., 2021). Mathematically, the hyperparameter optimisation using BO is denoted as:

$$p^{+} = arg \max_{p \in \vartheta} \eta(p) \tag{14}$$

where ϑ indicates the search space of p and p^{+} indicates the position at which the unknown function η is maximised.

BO consists of (a) an objective function that is modelled by a Bayesian statistical model and (b) an acquisition function, which determines where to sample next. The acquisition function accomplishes sampling points within the search space. The time complexity of BO is $O(n^3)$, where n indicates the number of observations (Wang et al., 2021b). Moreover, BO finds better hyperparameters than random search in fewer iterations in various applications such as recommender systems, robotics and reinforcement learning, environmental monitoring, and sensor networks, preference learning and interactive interfaces, automatic machine learning and hyperparameter tuning, combinatorial optimisation, and natural language processing due to its ability to exploit regions likely to contain optimal solutions (Shahriari et al., 2015; ALGorain and Clark, 2022). BO outperforms random search by intelligently exploring the search space, adapting the sampling strategy based on previous evaluations, leveraging acquired information, and effectively handling constraints. Furthermore, BO is effective in optimising complex, high-dimensional search spaces. The random search may struggle in such cases due to its lack of guidance, often requiring a large number of random samples to cover the search space adequately. BO, on the other hand, leverages the probabilistic surrogate model to make informed decisions about which hyperparameter configurations to evaluate next. This enables BO to efficiently explore and exploit the search space, leading to faster convergence towards better hyperparameters. These factors contribute to its superior performance and make it a popular choice for optimising complex functions such as Deep Belief Networks (DBNs) when limited evaluations are available (Bergstra et al., 2011).

Hence, in this study, we have optimised two hyperparameters of GAM using BO. The two hyperparameters are MaxNumSplits, which

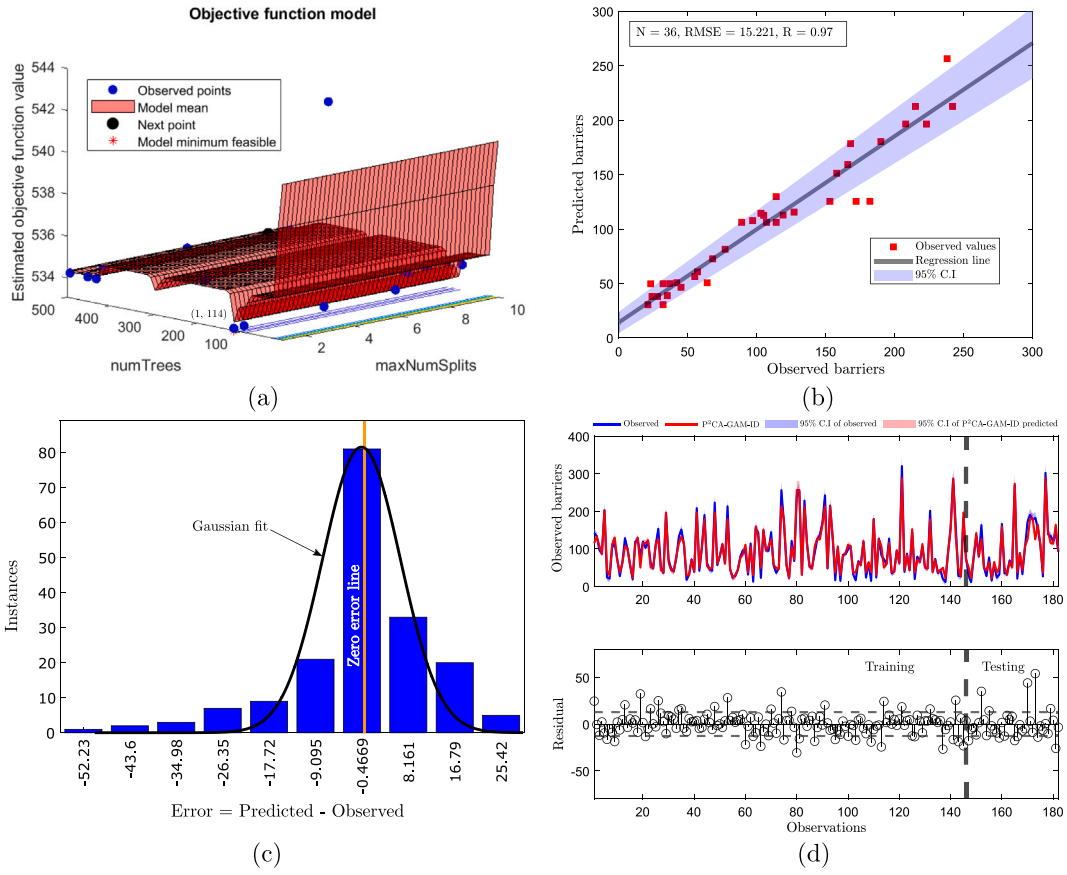


Fig. 4. (a) Illustration of the BO process for the optimisation of the hyperparameters, (b) linear regression plot between the observed and model predicted values on the LT-FS-ID dataset. The shade of blue represents the 95% confidence interval, (c) error histogram analysis using 10 bins, (d) residual plot. The dashed line illustrates the ± testing RMSE value.

represents the maximum number of branch node splits, and NumTrees, which represents the total number of trees in a forest.

5. Results

5.1. Model performance over LT-FS-ID dataset

We optimised the hyperparameters (*i.e.,* maxNumSplits and numTrees) using BO and illustrated the estimated objective function value in Fig. 4a. In doing so, we found that for maxNumSplits = 1 and numTreesWe = 114, the model records the minimum cross-validation loss between the observed and predicted values. Afterwards, we used 80% of the dataset to train the optimised model. We used the remaining dataset to test the model performance using R and RMSE as the performance metrics. To compute the performance metrics, we fitted a linear curve between the observed and predicted values (Fig. 4b). We found that the predicted barriers suit the observed values with R = 0.97 and RMSE = 15.22. Subsequently, to analyse the distribution of the errors, an error histogram (using ten bins) analysis is performed (Fig. 4c). We found that the error ranges between −52.23 (leftmost bin) to 25.42 (rightmost bin). The vertical orange line represents the zero error line. The region left to the vertical line represents the underestimation region (*i.e.,* negative errors), and the region right to the vertical line represents the overestimation region (*i.e.,* positive errors). It is observed that the error follows a Gaussian distribution with its peak exactly superimposed with the zero error line, indicating a good-fit model. Finally, we plotted the residual plot to check the appropriateness of the model (Fig. 4d). We found that the residuals are stochastic, indicating a good fit. Also, most of the residuals lie within the ± testing RMSE values with occasional peaks.

5.2. Model performance over AutoML-ID dataset

For the AutoML-ID dataset, we first optimised the hyperparameters by leveraging BO. In this case, we found that for maxNumSplits = 1, and numTrees = 329 the model attained the minimum cross-validation loss (Fig. 5a). We then trained the optimised model by using 80% of the dataset and used the remaining dataset (*i.e.,* 20%) for the testing of the trained model. We found that the model-predicted barriers align well with the observed values (with R = 0.99 and RMSE = 10.88). The regression line coincides with the 1:1 line, and all the data points cluster along the line, indicating the best-fit model (Fig. 5b). Then, we performed the error histogram analysis and found that the error ranges from −27.24 to 41.16. The error follows a Gaussian distribution with the peak lying near the zero error line, indicating a good-fit model (Fig. 5c). Finally, we performed the residual analysis and plotted the residuals at each instance (Fig. 5d). We found that most residuals are well inside the ± testing RMSE value with occasional peaks.

5.3. Model performance over FF-ANN-ID dataset

Using the FF-ANN-ID dataset, we have optimised the hyperparameters of the models by using BO and found that for maxNumSplits = 1 and numTrees = 150, the model obtained minimum cross-validation loss (Fig. 6a). We then divided the entire dataset into 80:20 ratios for training and testing, respectively. We used the training dataset to train the model using optimised hyperparameters. Afterwards, we analysed the trained model's performance by using the testing dataset. In this way, the testing data are fed into the model input, and its performance is recorded. We found that the predicted barriers accord well with the observed values with R = 0.98 and RMSE = 13.74 (Fig. 6b).

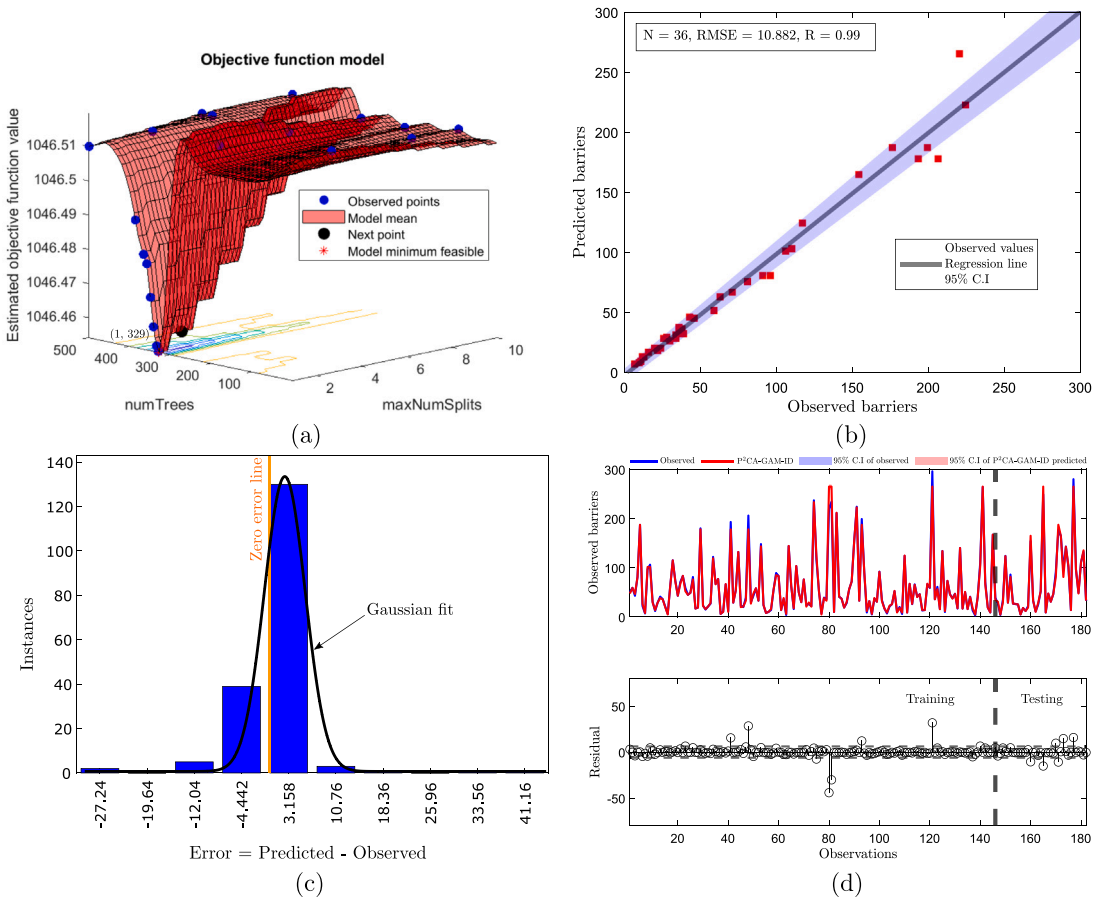


Fig. 5. (a) Illustration of the BO process for the optimisation of the hyperparameters, (b) linear regression plot between the observed and model predicted values on the AutoML-ID dataset. The shade of blue represents the 95% confidence interval, (c) error histogram analysis using 10 bins, and (d) residual plot. The dashed line illustrates the ± testing RMSE value.

We then performed the error histogram analysis and found that the error ranges from −30.49 to 45.11 (Fig. 6c). Also, the error follows a Gaussian distribution, indicating a good fit. Finally, we performed residual analysis and plotted the residuals in Fig. 6d. Every residual is random and lacks any pattern determining a good-fit model. Moreover, most of the residuals lie within the ± testing RMSE value.

After critically comparing the results of the proposed approach on all three datasets, we discovered the following profound insights that shed light on key aspects of the analysis:

- Consistency: The proposed model consistently performs well across all three datasets, suggesting its robustness and generalisability. This consistency indicates that the model is effective in capturing the underlying patterns and relationships present in different intrusion detection datasets.
- Dataset Variability: The performance metrics (R and RMSE) vary slightly across the datasets. This variability can be attributed to differences in the characteristics and complexities of the datasets themselves. It highlights the need to consider dataset-specific factors when evaluating and comparing model performance.
- AutoML-ID Superiority: The AutoML-ID dataset stands out with the highest correlation coefficient (R = 0.99) and the lowest root mean square error (RMSE = 10.88) among the three datasets. This suggests that the proposed model is particularly well-suited for this dataset, as it captures the patterns and relationships more accurately.
- General Performance: Overall, the proposed model demonstrates strong performance on all three datasets. The obtained correlation coefficients (R) are relatively high, indicating a good fit between predicted and actual values. The RMSE values are also within acceptable ranges, suggesting that the model's predictions are reasonably close to the true values.

In summary, the proposed machine learning model shows good performance on all three intrusion detection datasets for different scenarios of WSNs. While there are slight variations in the results, the model consistently demonstrates its efficacy in capturing the underlying patterns and relationships within the data. The AutoML-ID dataset particularly showcases the model's superior performance. These findings emphasise the model's potential as an effective tool for intrusion detection in WSNs and warrant further investigation and application in real-world scenarios.

6. Discussion

6.1. Ablation experiments and comparison

We perform an ablation experiment on the feature set to assess the individual contribution and impact of each input feature. We conducted a series of ablation experiments that systematically removed specific features from the input data while keeping other factors constant. We considered all the possible combinations of two and three input features. We then evaluated the performance of the P^{2}CA-GAM-ID model with each ablated feature combination, measuring various performance metrics such as R, RMSE, and bias on all three datasets (Table 1). By comparing the results of these ablation experiments, we were able to determine the optimal feature combination that yielded the highest performance for the P^{2}CA-GAM-ID model. The ablation experiments

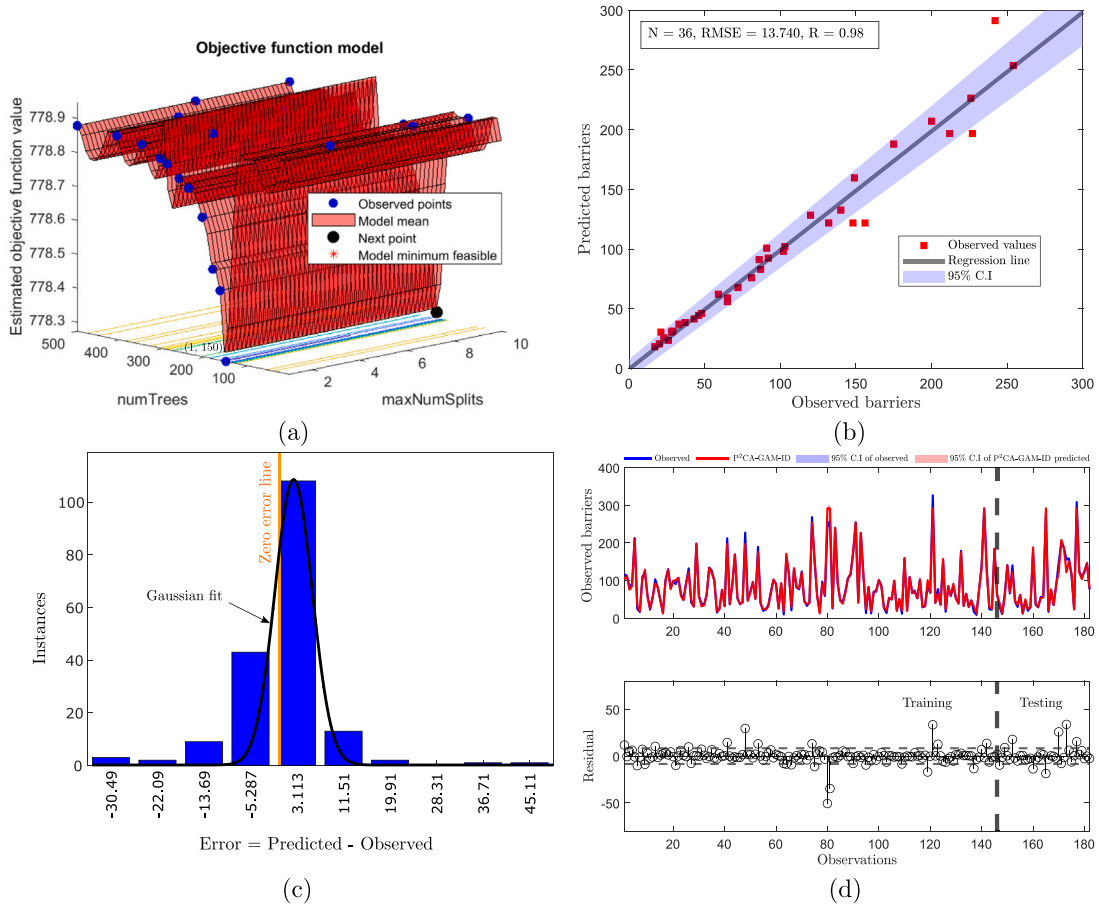


Fig. 6. (a) Illustration of the BO process for the optimisation of the hyperparameters, (b) linear regression plot between the observed and model predicted values on the FF-ANN-ID dataset. The shade of blue represents the 95% confidence interval, (c) error histogram analysis using 10 bins, and (d) residual plot. The dashed line illustrates the ± testing RMSE value.

Table 1
Ablation study on input features: A stands for Area, SR stands for Sensing range, TR stands for Transmission range, and NS stands for Number of sensors.

Ablated feature combination	Datasets								
	LT-FS-ID			AutoML-ID			FF-ANN-ID		
	R	RMSE	Bias	R	RMSE	Bias	R	RMSE	Bias
A+SR	0.36	26.93	−16.73	0.70	29.01	−12.77	0.52	29.43	−16.46
A+TR	0.36	26.93	−16.73	0.70	29.01	−12.77	0.52	29.43	−16.46
A+NS	0.36	26.93	−16.73	0.70	29.01	−12.77	0.52	29.43	−16.46
SR+TR	0.36	26.93	−16.73	0.40	36.67	−8.12	0.66	42.07	−4.28
SR+NS	0.36	26.93	−16.73	0.40	36.67	−8.12	0.66	42.07	−4.28
TR+NS	0.74	39.02	−5.81	0.40	36.67	−8.12	0.66	42.07	−4.28
A+SR+TR	0.94	23.07	−2.83	0.92	24.11	−2.26	0.96	24.56	−1.86
SR+TR+NS	0.86	28.34	−3.81	0.92	23.78	−3.57	0.97	16.71	−1.77
TR+NS+A	0.86	29.64	−2.82	0.93	23.28	−4.42	0.97	15.83	−1.61
NS+A+SR	0.94	21.14	0.55	0.92	26.60	1.28	0.96	21.49	0.49
All features (A+SR+TR+NS)	0.97	15.22	−1.49	0.99	10.88	−1.17	0.98	13.74	−0.93

revealed that the P²CA-GAM-ID model achieved the highest accuracy when all the input features were considered, indicating the collective importance of the feature set in capturing the underlying patterns and achieving optimal performance in all three datasets. In addition, we observed minimal variation in the model accuracy when considering a combination of two input features, whereas significant changes were observed when incorporating a combination of three input features. Our findings provided insights into the feature's importance and enabled us to make informed decisions regarding the inclusion or exclusion of specific features in the final model.

6.2. Comparison with the existing algorithms

We compared the results obtained from P²CA-GAM-ID with the primary algorithms for each dataset. We have compiled and tabulated the results in Table 2. For the LT-FS-ID dataset, we found alike performance with a comparable correlation coefficient. However, the RMSE of the P²CA-GAM-ID is higher than the primary algorithm. We found a similar observation in the case of the AutoML-ID dataset. Nevertheless, for the FF-ANN-ID dataset, we found that the results obtained through P²CA-GAM-ID outperform the primary algorithm with a high

Table 2
Comparison of the P²CA-GAM-ID results with the primary algorithms.

Performance metrics	LT-FS-ID (Singh et al., 2022c)		AutoML-ID (Singh et al., 2022b)		FF-ANN-ID (Gaussian) (Singh et al., 2022a)	
	Primary	P²CA-GAM-ID	Primary	P²CA-GAM-ID	Primary	P²CA-GAM-ID
R	0.98	0.97	1	0.99	0.76	0.98
RMSE	6.47	15.22	0.007	10.88	29.86	13.74

Table 3
Comparison with the benchmark algorithms.

Datasets	LT-FS-ID						AutoML-ID						FF-ANN-ID					
Algorithms	ANN	GRNN	RF	RBN	ERBN	This study	ANN	GRNN	RF	RBN	ERBN	This study	ANN	GRNN	RF	RBN	ERBN	This study
R	0.38	0.96	0.99	0.01	0.03	0.97	0.47	0.97	0.97	0.41	0.30	0.99	0.78	0.89	0.99	0.68	0.69	0.98
RMSE	46.37	57.56	32.15	40.45	65.03	15.22	36.96	64.61	75.72	161.11	107.95	10.882	41.15	39.66	11.37	65.67	75.12	13.74
Bias	−36.12	49.62	28.62	−13.15	52.57	−1.49	21.47	60.18	66.78	139.23	86.21	−1.17	3.02	60.70	53.90	39.19	60.11	−0.93

correlation coefficient (R = 0.98) and low RMSE (RMSE = 13.74).

6.3. Comparison with the benchmark algorithms

For a fair and unbiased evaluation of the proposed framework, we compared the P²CA-GAM-ID results with the benchmark algorithms that are frequently used to solve intrusion detection problems. We selected Artificial Neural Network (ANN), General Regression Neural Network (GRNN), Random Forest (RF), Radial Basis Neural Network (RBN), and Exact Radial Basis Neural Network (ERBN) algorithm for comparison and evaluated their performance on all the three datasets (*i.e.,* LT-FS-ID, AutoML-ID, and FF-ANN-ID). To ensure a fair comparison, we employed a common optimisation technique, namely BO, to optimise the hyperparameters of each algorithm. By employing this approach, we aimed to find the most suitable parameter settings for each method, effectively minimising any potential bias arising from inconsistent or suboptimal parameter choices. We considered R, RMSE, and bias as the performance metrics (Table 3). We found that P²CA-GAM-ID outperforms all the benchmark algorithms when we consider all the performance metrics for all three datasets. Also, among all the benchmark algorithms, RF emerged as the best-performing algorithm for all three datasets.

6.4. Performance over publicly available dataset

For generalisation of the P²CA-GAM-ID, we have tested its performance on dataset other than intrusion detection. In doing so, we have selected the widely used (Average Localisation Error) ALE regression dataset (Singh et al., 2020; Chen et al., 2022). We downloaded the dataset from UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/844/average+localization+error+(ale)+in+sensor+node+localization+process+in+wsns). It consists of four input features; anchor ratio, transmission range, node density, and iteration with ALE as the target variable. It has a dimension of 107 × 6. We applied the same methodology and optimised the model by using BO. In doing so, we found that for maxNumSplits = 6 and numTrees = 2 the model obtained minimum cross-validation loss (Fig. 7a). We used these tuning parameters to train the model by using 70% of the dataset and evaluated its performance by using the remaining 30% dataset. It is identified that the predicted ALE suits the observed values with R = 0.82 and RMSE = 0.202 (Fig. 7b). Afterwards, we performed error analysis and found that the error follows the Gaussian distribution, indicating a good-fit model with its peak lying near the zero error line (Fig. 7c). In addition, we performed the residual analysis and observed that the residuals are random in nature, and most of the residuals lie within the ± testing RMSE values (Fig. 7d).

Finally, we compared the performance of P²CA-GAM-ID with the primary ALE algorithms (such as S-SVR, Z-SVR, R-SVR, S-GPR, Z-GPR, and R-GPR) and other algorithms that uses ALE dataset (such as ResTT) (Table 4). We found that the P²CA-GAM-ID outperforms most of the primary algorithms. However, ResTT emerges as the best algorithm over the ALE dataset.

6.5. Limitations and future work

In this study, we proposed a novel approach by coupling P²CA and GAM to address a regression problem. P²CA was utilised to handle missing values in input features, effectively managing their impact on the model. However, it is important to note that excessive missing values within a single feature can lead to non-convergence and instability during the training of GAM. Further investigation is needed to explore strategies for handling such scenarios to ensure robust and stable model performance. Additionally, the datasets used for evaluating the model's performance assume time-invariance. However, in real-life scenarios, sensor performance may degrade over time due to the aging effect. This introduces a challenge in maintaining accurate results. It is crucial to consider periodic sensor maintenance or routine updating of the training datasets to account for these changes and ensure the continued accuracy and reliability of the model.

The successful application of coupling P²CA and GAM for accurately predicting the $k-$barriers in intrusion detection and prevention opens up several avenues for future research. Presented below are a range of thought-provoking and strategic recommendations for future work:

- To enhance the practical applicability of our approach, future research should focus on developing methodologies to effectively handle excessive missing values in individual features during GAM training.
- Furthermore, investigating techniques to incorporate the temporal dynamics of sensor performance and integrating sensor maintenance strategies into the modelling process will contribute to more accurate and reliable results in real-world settings.
- One promising direction is to explore the potential of nature-inspired algorithms for optimising hyperparameters in this context. Although our study utilised BO, there is a growing body of research on advanced optimisation algorithms that could offer additional benefits in optimising the hyperparameters of the proposed model. The effectiveness of new types of hybrid heuristics, metaheuristics, adaptive algorithms, self-adaptive algorithms, island algorithms, goat search algorithm, and polyploid algorithms has been demonstrated in various fields, including online learning, scheduling, multi-objective optimisation, transportation, medicine, data classification, and more (De, 2022; Zhao and Zhang, 2020; Dulebenets, 2021; Pasha et al., 2022; Gholizadeh et al., 2021; Dulebenets et al., 2018; Rabbani et al., 2022). Specifically, these algorithms could be explored to find optimal hyperparameter settings that improve the accuracy and robustness of the intrusion detection and prevention system.

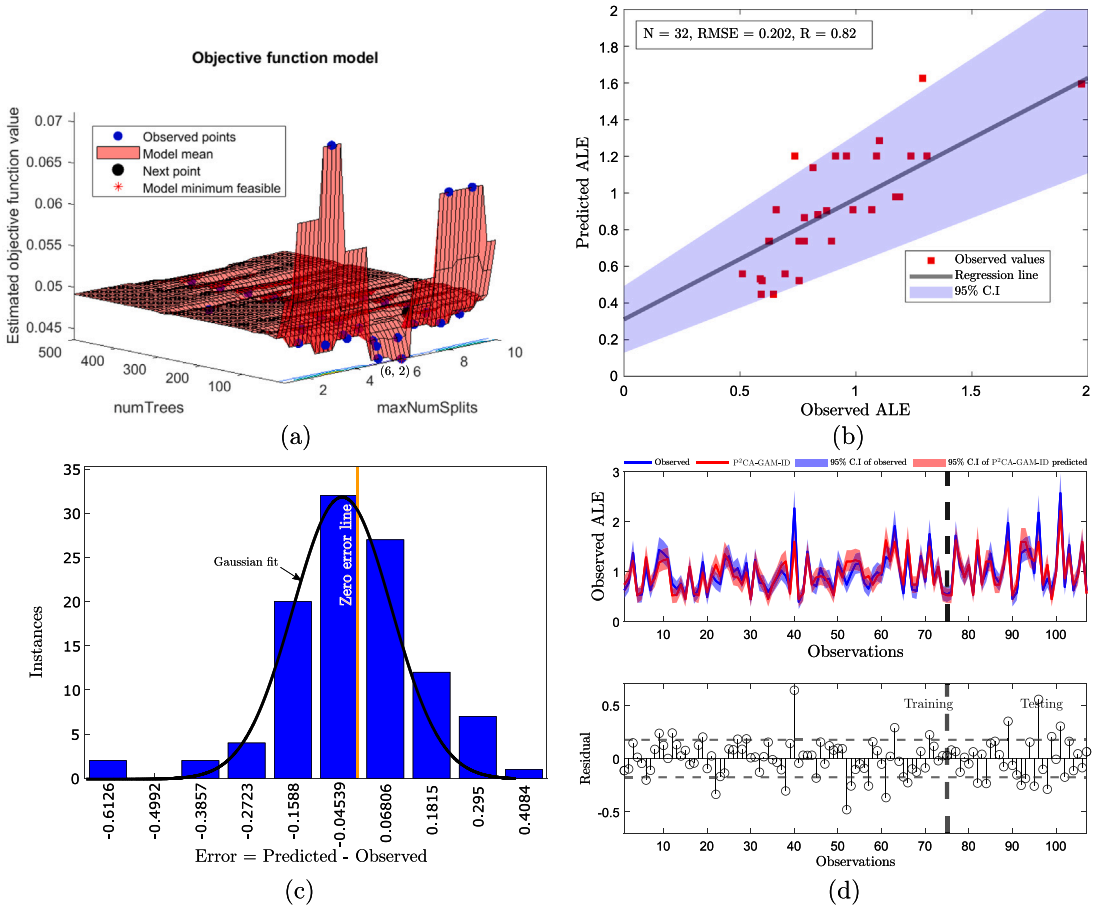


Fig. 7. (a) Illustration of the BO process for the optimisation of the hyperparameters, (b) linear regression plot between the observed and model predicted values. The shade of blue represents the 95% confidence interval, (c) error histogram analysis using 10 bins, and (d) residual plot. The dashed line illustrates the ± testing RMSE value.

Table 4
Comparison of the performance of P²CA-GAM-ID on publicly available datasets other than intrusion detection.

Performance metrics	ALE algorithms (Singh et al., 2020, 2021a)						ResTT (Chen et al., 2022)	P²CA-GAM-ID (This study)
	S-SVR	Z-SVR	R-SVR	S-GPR	Z-GPR	R-GPR		
R	0.80	0.81	0.82	0.74	0.72	0.71	0.86	0.82
RMSE [m]	0.23	0.20	0.15	0.22	0.23	0.23	0.18	0.20

By addressing these challenges, we can improve the robustness and long-term viability of the P²CA-GAM coupling approach for regression problems and enhance its real-world applicability in domains where data quality and temporal dynamics are significant considerations.

7. Conclusion

In this study, we proposed a novel regression-based algorithm for fast intrusion detection and prevention by coupling P²CA and GAM. We utilise the publicly available synthetic intrusion detection datasets (*i.e.,* LT-FS-ID, AutoML-ID, and FF-ANN-ID) to test the model's performance. All these datasets are widely used for the prediction of $k-$barriers using WSNs. We found that the proposed approach gives excellent results for all three datasets in terms of accuracy, hence eradicating the need for scenario-specific algorithms for the accurate prediction of $k-$barriers. Further, it also outperforms various benchmark algorithms such as ANN, RF, GRNN, RBN, and ERBN.

For a more robust conclusion, we also tested the model performance over other problem domains of WSNs. We used the publicly available sensor node localisation ALE dataset for testing the performance of the proposed approach. We found that the proposed algorithms outperform the primary results over ALE datasets. This study is a step towards a single algorithm solution for a multi-scenario-based intrusion detection problem. The proposed technique can be used for near-real-time border surveillance.

Code availability

The code is available for download at https://abhilashsingh.net/codes.html.

CRediT authorship contribution statement

Abhilash Singh: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization, Investigation, Writing – review & editing. **Jaiprakash Nagar:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. **J. Amutha:** Data curation, Visualization, Investigation, Writing – original draft. **Sandeep Sharma:** Methodology, Visualization, Investigation, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used in this study are publicly available.

Acknowledgements

The authors would like to acknowledge IISER Bhopal, IIT Kharagpur, and the Department of Higher and Technical Education, Ranchi, Jharkhand for providing institutional support. They would like to thank to the editor and all the six anonymous reviewers for providing helpful comments and suggestions.

Appendix

See Table A.1

Table A.1
List of abbreviations.

ANFIS	Adaptive Neuro-FIS
AutoML	Automated Machine Learning
ALE	Average Localisation Error
BO	Bayesian Optimisation
BDT	Binary Decision Tree
BSM	Binary Sensing Model
B-EL	Boosting Ensemble Learning
BEs	Boundary Effects
C.I	Confidence Interval
C-GPR	Centre-mean-GPR
DBNs	Deep Belief Networks
EFNNs	Evolving Fuzzy Neural Networks
ERBN	Exact Radial Basis Neural Network
FC-FF-ID	Fully Connected Feed Forward Intrusion Detection
GP	Gaussian Process
GPR	Gaussian Process Regression
GAM	Generalised Additive Model
GRNN	Generalised Regression Neural Network
GBMs	Gradient Boosting Machines
LR	Linear Regression
ML	Machine Learning
maxNumSplits	Maximum Number of Splits
NS-2.35	Network Simulator-2.35
NS-GPR	Non-standardise GPR
numTrees	Number of Trees
PDP	Partial Dependency Plot
PCA	Principal Component Analysis
PCs	Principal Components
P2CA	Probabilistic PCA
PDF	Probability Density Function
RBN	Radial Basis Neural Network
RF	Random Forest
RoI	Region of Interest (RoI)
S-GPR	Scale-GPR
SNs	Sensor Nodes
SEs	Shadowing Effects
SD	Standard Deviation
SVR	Support Vector Regression
WoS	Web of Science
WSNs	Wireless Sensor Networks

(continued on next page)

Table A.1 (*continued*).

References

ALGorain, F.T., Clark, J.A., 2022. Bayesian hyper-parameter optimisation for malware detection. Electronics 11 (10), 1640.

Amutha, J., Nagar, J., Sharma, S., 2021. A distributed border surveillance (DBS) system for rectangular and circular region of interest with wireless sensor networks in shadowed environments. Wirel. Pers. Commun. 117 (3), 2135–2155.

Aponte-Luis, J., Gómez-Galán, J.A., Gómez-Bravo, F., Sánchez-Raya, M., Alcina-Espigado, J., Teixido-Rovira, P.M., 2018. An efficient wireless sensor network for industrial monitoring and control. Sensors 18 (1), 182.

Arjun, D., Indukala, P., Menon, K.U., 2019. PANCHENDRIYA: A multi-sensing framework through wireless sensor networks for advanced border surveillance and human intruder detection. In: 2019 International Conference on Communication and Electronics Systems. ICCES, IEEE, pp. 295–298.

Arora, M., Pal, A., 2022. A deep learning approach to accurately predict the κ-coverage probability in wireless sensor networks. Wirel. Pers. Commun. 1–18.

Benahmed, T., Benahmed, K., 2019. Optimal barrier coverage for critical area surveillance using wireless sensor networks. Int. J. Commun. Syst. 32 (10), e3955.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54, 1937–1967.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. Adv. Neural Inf. Process. Syst. 24.

Bhadwal, N., Madaan, V., Agrawal, P., Shukla, A., Kakran, A., 2019. Smart border surveillance system using wireless sensor network and computer vision. In: 2019 International Conference on Automation, Computational and Technology Management. ICACTM, IEEE, pp. 183–190.

Chen, Y., Pan, Y., Dong, D., 2022. Residual tensor train: A quantum-inspired approach for learning multiple multilinear correlations. IEEE Trans. Artif. Intell..

De, S.K., 2022. The goat search algorithms. Artif. Intell. Rev. 1–37.

de Campos Souza, P.V., Lughofer, E., Rodrigues Batista, H., 2022. An explainable evolving fuzzy neural network to predict the k barriers for intrusion detection using a wireless sensor network. Sensors 22 (14), 5446.

Dulebenets, M.A., 2021. An adaptive polyploid memetic algorithm for scheduling trucks at a cross-docking terminal. Inform. Sci. 565, 390–421.

Dulebenets, M.A., Kavoosi, M., Abioye, O., Pasha, J., 2018. A self-adaptive evolutionary algorithm for the berth scheduling problem: Towards efficient parameter control. Algorithms 11 (7), 100.

Geraci, M., Farcomeni, A., 2016. Probabilistic principal component analysis to identify profiles of physical activity behaviours in the presence of non-ignorable missing data. J. R. Stat. Soc. Ser. C. Appl. Stat. 65 (1), 51–75.

Gholizadeh, H., Fazlollahtabar, H., Fathollahi-Fard, A.M., Dulebenets, M.A., 2021. Preventive maintenance for the flexible flowshop scheduling under uncertainty: A waste-to-energy system. Environ. Sci. Pollut. Res. 1–20.

Ghosh, K., Neogy, S., Das, P.K., Mehta, M., 2018. Intrusion detection at international borders and large military barracks with multi-sink wireless sensor networks: An energy efficient solution. Wirel. Pers. Commun. 98 (1), 1083–1101.

Hossain, A., Chakrabarti, S., Biswas, P.K., 2012. Impact of sensing model on wireless sensor network coverage. IET Wirel. Sensor Syst. 2 (3), 272–281.

Karanja, S., Badru, R., 2021. Development of a low cost wireless sensor network for surveillance along Kenya-Somalia border. Preprint.

Karthick, R., Prabaharan, A.M., Selvaprasanth, P., 2019. Internet of things based high security border surveillance strategy. Asian J. Appl. Sci. Technol. (AJAST) Vol. 3, 94–100.

Keung, G.Y., Li, B., Zhang, Q., 2012. The intrusion detection in mobile sensor network. IEEE/ACM Trans. Netw. 20 (4), 1152–1161. http://dx.doi.org/10.1109/TNET.2012.2186151.

Konstantinov, A.V., Utkin, L.V., 2021. Interpretable machine learning with an ensemble of gradient boosting machines. Knowl.-Based Syst. 222, 106993.

Kotiyal, V., Singh, A., Sharma, S., Nagar, J., Lee, C.-C., 2021. ECS-NL: An enhanced cuckoo search algorithm for node localisation in wireless sensor networks. Sensors 21 (11), 3576.

Kumar, S., Duttagupta, S., Rangan, V.P., Ramesh, M.V., 2020. Reliable network connectivity in wireless sensor networks for remote monitoring of landslides. Wirel. Netw. 26 (3), 2137–2152.

Ma, Z., Yun, C.-B., Wan, H.-P., Shen, Y., Yu, F., Luo, Y., 2021. Probabilistic principal component analysis-based anomaly detection for structures with missing data. Struct. Control Health Monit. 28 (5), e2698.

Muruganandam, S., Joshi, R., Suresh, P., Balakrishna, N., Kishore, K.H., Manikanthan, S., 2023. A deep learning based feed forward artificial neural network to predict the K-barriers for intrusion detection using a wireless sensor network. Measurement: Sensors 25, 100613.

Nagar, J., Chaturvedi, S.K., Soh, S., 2020. An analytical model to estimate the performance metrics of a finite multihop network deployed in a rectangular region. J. Netw. Comput. Appl. 149, 102466.

Nagar, J., Chaturvedi, S.K., Soh, S., 2022. Wireless multihop network coverage incorporating boundary and shadowing effects. IETE Tech. Rev. 39 (5), 1124–1139.

Nagar, J., Chaturvedi, S.K., Soh, S., Singh, A., 2023. A machine learning approach to predict the k-coverage probability of wireless multihop networks considering boundary and shadowing effects. Expert Syst. Appl. 226, 120160.

Nagar, J., Sharma, S., 2018. K-barrier coverage-based intrusion detection for wireless sensor networks. In: Cyber Security. Springer, pp. 373–385.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobot. 7, 21.

Noel, A.B., Abdaoui, A., Elfouly, T., Ahmed, M.H., Badawy, A., Shehata, M.S., 2017. Structural health monitoring using wireless sensor networks: A comprehensive survey. IEEE Commun. Surv. Tutor. 19 (3), 1403–1423.

Pasha, J., Nwodu, A.L., Fathollahi-Fard, A.M., Tian, G., Li, Z., Wang, H., Dulebenets, M.A., 2022. Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings. Adv. Eng. Inform. 52, 101623.

Rabbani, M., Oladzad-Abbasabady, N., Akbarian-Saravi, N., 2022. Ambulance routing in disaster response considering variable patient condition: NSGA-II and MOPSO algorithms. J. Ind. Manag. Optim. 18 (2), 1035–1062.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2015. Taking the human out of the loop: A review of Bayesian optimization. Proc. IEEE 104 (1), 148–175.

Sharma, S., Nagar, J., 2020. Intrusion detection in mobile sensor networks: A case study for different intrusion paths. Wirel. Pers. Commun. 115 (3), 2569–2589.

Shi, R., Xu, X., Li, J., Li, Y., 2021. Prediction and analysis of train arrival delay based on xgboost and Bayesian optimization. Appl. Soft Comput. 109, 107538.

Shukla, A.K., Srivastav, S., Kumar, S., Muhuri, P.K., 2023. UInDeSI4. 0: An efficient unsupervised intrusion detection system for network traffic flow in Industry 4.0 ecosystem. Eng. Appl. Artif. Intell. 120, 105848.

Singh, A., Amutha, J., Nagar, J., Sharma, S., 2022a. A deep learning approach to predict the number of k-barriers for intrusion detection over a circular region using wireless sensor networks. Expert Syst. Appl. 118588.

Singh, A., Amutha, J., Nagar, J., Sharma, S., Lee, C.-C., 2022b. AutoML-ID: Automated machine learning model for intrusion detection using wireless sensor network. Sci. Rep. 12 (1), 1–14.

Singh, A., Amutha, J., Nagar, J., Sharma, S., Lee, C.-C., 2022c. LT-FS-ID: Log-transformed feature learning and feature-scaling-based machine learning algorithms to predict the k-barriers for intrusion detection using wireless sensor network. Sensors 22 (3), 1070.

Singh, A., Gaurav, K., Sonkar, G.K., Lee, C.-C., 2023a. Strategies to measure soil moisture using traditional methods, automated sensors, remote sensing, and machine learning techniques: Review, bibliometric analysis, applications, research findings, and future directions. IEEE Access.

Singh, A., Kotiyal, V., Sharma, S., Nagar, J., Lee, C.-C., 2020. A machine learning approach to predict the average localization error with applications to wireless sensor networks. IEEE Access 8, 208253–208263.

Singh, A., Mehra, M., Kumar, A., Niranjannaik, M., Priya, D., Gaurav, K., 2023b. Leveraging hybrid machine learning and data fusion for accurate mapping of malaria cases using meteorological variables in western India. Intell. Syst. Appl. 17, 200164.

Singh, A., Nagar, J., Sharma, S., Kotiyal, V., 2021a. A Gaussian process regression approach to predict the k-barrier coverage probability for intrusion detection in wireless sensor networks. Expert Syst. Appl. 172, 114603.

Singh, A., Sharma, S., Singh, J., 2021b. Nature-inspired algorithms for wireless sensor networks: A comprehensive survey. Comp. Sci. Rev. 39, 100342.

Singh, A., Sharma, S., Singh, J., Kumar, R., 2019. Mathematical modelling for reducing the sensing of redundant information in WSNs based on biologically inspired techniques. J. Intell. Fuzzy Systems 37 (5), 6829–6839.

Singh, R., Singh, S., 2021. Smart border surveillance system using wireless sensor networks. Int. J. Syst. Assur. Eng. Manag. 1–15.

Sood, T., Prakash, S., Sharma, S., Singh, A., Choubey, H., 2022. Intrusion detection system in wireless sensor network using conditional generative adversarial network. Wirel. Pers. Commun. 1–21.

Suresha, H.S., Parthasarathy, S.S., 2021. Probabilistic principal component analysis and long short-term memory classifier for automatic detection of Alzheimer's disease using MRI brain images. J. Inst. Eng. (India): Series B 102 (4), 807–818.

Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. In: Handbook of Statistics, Vol. 24. Elsevier, pp. 303–329.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (3), 611–622.

Tong, C., Zhang, C., Liu, C., 2021. Investigation on the relationship between satellite air quality measurements and industrial production by generalized additive modeling. Remote Sens. 13 (16), 3137.

Wang, Y., Fu, W., Agrawal, D.P., 2012. Gaussian versus uniform distribution for intrusion detection in wireless sensor networks. IEEE Trans. Parallel Distrib. Syst. 24 (2), 342–355.

Wang, W., Jia, Y., Yu, W., Pang, H., Cai, K., 2021a. On-line ammonia nitrogen measurement using generalized additive model and stochastic configuration networks. Measurement 170, 108743.

Wang, Y., Wang, H., Peng, Z., 2021b. Rice diseases detection and classification using attention based neural network and Bayesian optimization. Expert Syst. Appl. 178, 114770.

Wang, D., Xie, B., Agrawal, D.P., 2008. Coverage and lifetime optimization of wireless sensor networks with gaussian distribution. IEEE Trans. Mob. Comput. 7 (12), 1444–1458.

Zhao, H., Zhang, C., 2020. An online-learning-based evolutionary many-objective algorithm. Inform. Sci. 509, 1–21.