Coverage Prediction for Target Coverage in WSN Using Machine Learning Approaches

Pooja Chaturvedi (chaturvedi.pooja03@gmail.com)
Nirma University Institute of Technology https://orcid.org/0000-0001-5207-2696
Ajai Kumar Daniel
Madan Mohan Malaviya Engineering College: Madan Mohan Malaviya University of Technology
Vipul Narayan
Madan Mohan Malaviya Engineering College: Madan Mohan Malaviya University of Technology

Research Article

Keywords: Wireless Sensor Network, Target Coverage, Probabilistic Coverage Model, Trust, Network Lifetime, Machine Learning, Prediction.

Posted Date: December 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1163536/v1

Coverage Prediction for Target Coverage in WSN Using Machine Learning Approaches

Pooja Chaturvedi[1], A K Daniel[2], Vipul Narayan[3]

Institute of Technology, Nirma University Ahmedabad Gujrat[1]

Madan Mohan Malaviya University of Technology Gorakhpur[2, 3]

chaturvedi.pooja03@gmail.com[1], danielak@rediffmail.com[2],vipul.upsainian2470@gmail.com[3]

Abstract

Mathematical programming techniques are widely used in the determination of optimal functional configuration of a wireless sensor network (WSN). But these techniques have usually high computational complexity and are often considered as Non Polynomial (NP) complete problems. Therefore, machine learning (ML) techniques can be utilized for the prediction of the WSN parameters with high accuracy and lesser computational complexity than the mathematical programming techniques. This paper focuses on developing the prediction model for determination of the node status to be included in the set cover based on the coverage probability and trust values of the nodes. The set covers are defined as the subset of nodes which are scheduled to monitor the region of interest with the desired coverage level. Several machine learning techniques have been used to determine the node activation status based on which the set covers are obtained. The results show that the random forest based prediction model yields the highest accuracy for the considered network setting.

Keywords: Wireless Sensor Network, Target Coverage, Probabilistic Coverage Model, Trust, Network Lifetime, Machine Learning, Prediction.

1. Introduction

WSN find their applicability in diverse real time scenarios due to their inherent characteristic such as small size, low cost and convenient deployment. The main functionality of the sensor network is to observe and record the environmental parameters and transfer this information to the central node known as base station for the user analysis and inference. The sensor network usually comprise of a large number of nodes which perform the environmental monitoring in collaboration. However, the management of such a large network is a complex and tedious task, so distributed and scalable mechanisms are required. In addition to this, the topology of the sensor network is categorized of being dynamic due to the different factors such as node mobility, link failure or node failure due to hardware or software characteristics. The dynamic nature of the sensor network affects the network functionalities such as routing, localization, coverage and other Quality of Service (QoS) parameters, which may lead to redesign or redeployment of the network components [1] [2] [3].

Coverage is defined as a service quality parameter which aims to determine the duration for which the node can provide the monitoring of the region of interest with the required reliability. The coverage problem in the sensor network can be of three types as: area coverage-which emphasize on the observation of all the points in the region of observation, target/point coverage- which emphasize on the monitoring of a predefined set of points and the barrier coverage- which emphasize on tracking of the movement of intruder inside the region of interest. The coverage functionality of the sensor node is usually dependent on the sensing range and communication range. The sensing model of the sensor node is usually defined as a circular region, in which the node is situated at the center of the circle. The communication model is also modeled as a circular region, within which the node can transmit and receive the information from nearby points. A point in the region of interest is said to be covered by a sensor node if it lies within the sensing range of the sensor node. The coverage provided by a sensor node is a function of the distance of the sensor node from the target point. The sensing model of a sensor node may be either Boolean or Probabilistic. In the Boolean sensing model, the point in the region of interest is said to be monitored with

probability 1 if its distance is lesser than the sensing range of the node, otherwise the monitoring probability is zero. The probabilistic sensing model is considered to be more realistic as it states that the observation probability of the point under observation decreases exponentially with the increase in the distance between the sensor node and the target point.

The node scheduling approaches are considered as one of most efficient mechanism to address the energy conservation and coverage problem which emphasizes on the prolonging the network lifetime by scheduling the nodes to be in active state based on certain predefined parameters. The subsets of nodes which can monitor all the target points in a region are considered to constitute a set cover. In such cases, the coverage problem emphasizes on determining as many number of set covers as possible, such that network lifetime is maximized. The related problem with coverage is the connectivity problem, which ensures the existence of path between every two nodes [4] [5].

The organization of the paper is as follows: motivation and contribution in section 2, related work in section 3, machine learning models in section 4, network model in section 5, proposed approach and methodology in section 6, results and discussion in section 7 and section 8 concludes the paper with the future scope.

2. Motivation and Contribution

The traditional approaches for the sensor network deployment and organization are dependent on the predefined programming, which is not applicable for the dynamic network characteristics. The incorporation of machine learning based techniques may significantly improve the network performance and lowers the human intervention in explicitly programming the network functionality. Machine learning may adapt to the dynamic nature of the network by automatically improving through the automated learning from past experience or training. Thus the machine learning approaches make the computing processes more efficient, cost effective and robust. Machine learning approaches develop the models after analyzing even more complex data more easily and accurately. The learning approaches in the machine learning based model are usually categorized as supervised, unsupervised and reinforcement learning. In the supervised learning, the developed model predicts the class of the new dataset based on the analysis of the some predefined training dataset. The examples of supervised learning can be seen in the regression analysis and classification approaches. The unsupervised learning is somewhat autonomous in nature, as it tries to evaluate the environmental parameters on its own and adapt into it based on its own learning and observation. The most common example of unsupervised learning is found in the clustering applications, in which a number of data sets are divided into a number of clusters based on certain similarity metrics. The reinforcement learning approaches are based on the concepts of reward and penalty. The agent is provided with the feedback regarding the action performed by it. The reinforcement learning approaches are widely used in problem solving and game playing. The machine learning based models effectiveness lies in the fact that they can provide generalized solutions for the complex problems based on their learning and adaptability. Due to the interdisciplinary nature of machine learning models have been utilized in addressing the different issues related to the sensor network. Some of the problems in which machine learning models have been applied are as follows [6]-[12]:

1. The node energy consumption in different time slots can be predicted which can enhance the network performance.
2. The machine learning based models may be used in identification and removal of malicious nodes from the network.
3. Machine learning approaches may be utilized in the accurate prediction of the location of nodes within the target region.
4. The dynamic routing approaches may be designed using the machine learning based approaches.

The main motivation behind the proposed model is the need to incorporate the machine learning approach to address the target coverage problem which can provide the coverage to the predefined set of targets with the desired

coverage level. To the best of our knowledge, this work is the first attempt to utilize the machine learning based model to predict the node status for the target coverage problem such that all the targets are monitored with the desired coverage degree with activating the optimal number of nodes. The key contribution of the proposed model is as:

1. The prominent attributes of the network affecting the node activation status are identified.
2. Several prediction models have been designed to determine the node status based on historical data so that unnecessary and redundant data transmissions may be reduced. The performance of the different prediction model is analyzed in terms of different performance metrics.
3. Cross validation techniques are employed to identify the best model attributes.

3. Related Work

There are various approaches proposed in the literature to prolong the network lifetime while addressing the target coverage problem. The target coverage approaches can be broadly categorized into three types: adjusting the sensing range to enhance the coverage, optimal deployment techniques based on geometrical constructs and virtual force based approach and the node scheduling based approaches. The sensing range of the sensor node may be adjusted according to the energy level to maximize the coverage of the monitored region. The geometrical construct based approaches place the nodes based on designing the Voronoi diagram and Delaunay Triangulation. The virtual force based approaches address the coverage problem by pushing the nodes away from each other or pulling the nodes towards each other to achieve the desired coverage [13]-[17].

The node scheduling based approaches are based on the concept of energy conservation by activating only a subset of nodes instead of activating all the nodes at once. The node scheduling based approach is based on the fact that the number of set covers a node can participate is dependent on the initial energy of the nodes. There are various node scheduling based approaches proposed in the literature. In [17], authors have addressed the target coverage problem by determining the non-disjoint set covers which are activated according to the schedule determined by the base station. In this approach, a node can participate only in one set cover. In [18], the number of set covers is enhanced by employing the disjoint set cover approach, in which a node can participate in multiple set covers depending on the residual energy of the nodes. In [19], authors have addressed the connected target coverage problem, which ensures that there is a path from the active node towards the base station.

Most of these approaches consider the Boolean coverage model, which is not feasible in real world applications. In [20], authors have proposed a node scheduling based approach which determines the number of set covers on the basis of coverage probability, trust values of the nodes and the node contribution. This approach also aims to obtain the set covers consisting of minimal number of nodes. The probabilistic coverage model considers that the detection probability of a node decreases exponentially with the increasing distance between the sensor node and the target. The trust value of a node is determined as a commutative function of direct trust, indirect trust and recommendation trust. The direct trust is determined as a function of data trust, communication trust and energy trust. The node contribution refers to the number of targets, a node can monitor. In this approach, the network lifetime is defined as the duration until all the target points in the region are monitored with desired confidence level. The network lifetime is proportional to the number of set covers obtained. It has been established that the determination of the number of set covers is NP complete problem [21]-[27]. Hence the traditional computation models are more complex to apply. In the proposed work, we aim to incorporate the machine learning based prediction model which can predict the node status as either active or sleep based on the historical data. The machine learning based approach may reduce the number of data transmission and hence can improve the network performance.

4. Machine Learning Models

The machine learning models can be categorized into three classes as supervised learning, unsupervised learning and reinforcement learning. The supervised learning techniques are based on the labeled data sets based on which the new data values are classified to a particular class. The supervised learning approaches are linear regression, multiple regression, logistic regression, naïve bayes, random forest based classification and support vector machine. The unsupervised learning has the capability to evolve based on the unlabeled data set. The different clustering approaches such as K-means, hierarchical clustering is considered as unsupervised learning mechanism. The reinforcement learning approaches is similar to the unsupervised in the case that it is also based on the unlabeled data set but it also has a concept of reward/penalty. It is commonly used in game playing in which the agent learns about the environment on its own exploration.

The supervised machine learning based approaches can be used to solve several issues in wireless sensor networks such as localization, fault detection and identification, coverage and connectivity, routing approaches and anomaly detection. The ML based localization approaches aim to determine the location of anchor nodes and sensor nodes. The node failure probability can be predicted using machine learning based approaches. The coverage and connectivity problems may utilize the machine learning algorithms to determine the optimal number of nodes to keep in active or sleep state. The supervised learning algorithms may be used for determining the optimal routing approaches which can enhance the network lifetime. The anomaly detection approaches utilize machine learning based approach to determine the malicious or faulty nodes. The unsupervised learning based algorithms are mainly used for solving the clustering and dimensionality reduction problems. The reinforcement learning based approaches are based on the application of Q-learning to solve the different issues in WSN such as coverage, routing and QoS parameters. The machine learning based approaches provide these advantages in the solution of coverage and connectivity problems [29]:

1. The minimum number of nodes to monitor a given region of interest can be determined in quick time even in a dynamic environment.
2. The connected or disconnected nodes can be identified easily and the network routes can be updated dynamically.

In [30] and [31], regression based connectivity improvement approaches have been proposed. In [30], the network quality and reliability in deployment of sensor nodes is optimized using the mapping function on the basis of received signal strength, noise and packet reception rate. The approach in [31] improves the connectivity of the network by reducing the overhead involved in the in-network aggregation. In [32], Support Vector Machine (SVM) based connectivity improvement has been proposed which reduces the communication complexity and shorter message exchanges. In [33], SVM based decision classification method has been proposed to estimate the link quality on the basis of received signal strength and link quality indicator. In [34], Random Forest (RF) based target coverage improvement approach has been proposed. In [35], Naïve bayes based tracking approach has been proposed to track the human in the sensor network. This approach has been found to be computationally simple. In [36] fuzzy c-means algorithm has been proposed to classify the work load of the sensor nodes. In [37] Q, learning based two step scheduling approach has been proposed which aim to select the nodes to cover the desired set of points. In [38], a reinforcement learning based probe algorithm is proposed to estimate the quality of the link through minimum energy consumption and overhead. In [39], several evolutionary optimization approaches has been proposed to enhance the coverage and link quality. The different machine learning based approaches for coverage and connectivity problems have been summarized in the Table 1 as shown:

Table 1. Machine Learning Approaches for Coverage and Connectivity Issues

S. No.	Reference	ML Technique Used	Objective
1	[30][31]	Regression	Connectivity
2	[32]	SVM	Connectivity
3	[33]	SVM along with Decision Tree	Connectivity
4	[34]	Random Forest	Coverage
5	[35]	Bayesian	Coverage
6	[36] [37]	K-means with Fuzzy C-means	Connectivity
7	[38] [39]	Reinforcement	Coverage Connectivity

It is clear from the above table the machine learning approaches have been mostly applied in the connectivity problems with the objective to enhance the network lifetime. There is still a lot of research required in the application of the machine learning based approaches to determine the minimum number of nodes to monitor all the targets. The proposed approach aims to predict the node status with respect to the observation of target points.

5. Network Model

The network consists of *m* sensor nodes and *n* target nodes deployed randomly in a target region of dimension 50*50 square meters. The nodes and targets remain static after the deployment. All the nodes in the sensor network are homogeneous in terms of the storage, communication and processing capabilities. The sensing model and communication model of the nodes are considered as a circular disk. The objective of the node scheduling based approach is to determine the number of set covers which can monitor all the targets

5.1 Coverage probability

The proposed approach considers the probabilistic coverage model, in which the detection probability decreases exponentially with respect to the increase in the distance between the sensor node and the target. The mathematical representation of the coverage probability is as shown in the equation 1:

$$\operatorname{cov}(m,n)=\begin{cases}0 & if\ r_s+r_c \le D(S_m,T_n)\\ e^{-\lambda x^y} & if\ r_s-r_c < D(S_m,T_n) < r_s+r_c\\ 1 & if\ r_s-r_c \ge D(S_m,T_n)\end{cases} \quad (1)$$

where r_s is the sensing range of the sensor node, r_c is the detection error range which represents the range within which error is permissible and D represents the distance between the sensor node m and the target node n. λ, x and y are the sensing parameters which depends on the hardware charactersistics. The

observation probability of a target with respect to a set of sensor nodes is defined as a function of coverage probability and trust value of the nodes as shown in the equation 2 below:

$$P_{mn}^{obs} = \mathrm{cov}(m,n) \times T \tag{2}$$

5.2 Trust calculation

The reliability of the data sensed and communicated between the nodes depends on the trust value of the nodes. The trust value of a node is defined as the subjective opinion of the nodes which have communicated with the nodes earlier. In the proposed work we have considered the trust model presented in [26] [27]. In this the trust of a node is determined as the weighted average of the direct trust, indirect trust and recommendation trust. The node who directly communicates with a node would contribute more towards the direct trust. The direct trust is in turn defined as the weighted sum of data trust, communication trust and energy trust. In the case where the nodes cannot directly communicate with each other but can communicate through common neighbors can utilize the recommendation trust. Recommendation trust is further defined as the weighted average of the recommendation reliability and familiarity. The nodes which are not directly reachable and those who are not connected through direct neighbors utilize the concept of indirect trust. The trust values are updated periodically to address the dynamics of the network.

5.3 Machine learning models

The several machine learning models considered in this paper are discussed as follows [12]:

5.3.1 Regression

Regression is an example of supervised learning algorithm, which is used to interpolate the value of a dependent variable based on the values of the independent variable. The mathematical representation of the regression model is as shown in the equation 3:

$$\beta = f(\alpha) + \delta \tag{3}$$

Where β represents a dependent variable and α is an independent variable. δ represents the error induced during the learning mechanism. *f* is a function which establishes the relationship between the independent and dependent variables. The regression based machine learning model assumes the variables to contain continuous values. In the proposed approach we are trying to classify the nodes according to their status in either active or sleep state, so linear regression is not feasible to apply. The logistic regression based machine learning model is more suitable for such kind of classification problem. The logistic regression aims to determine the probability of assignment of binary values to the dependent variable using the sigmoid activation function as shown in the equation 4:

$$P = \frac{1}{1 + e^{-(X + \alpha Y)}} \tag{4}$$

Where P is the probability of assigning a particular value to the dependent variable, X and Y are the model characteristics and α is the independent variable. The application of regression based approaches for solving different WSN issues are localization, data aggregation, coverage and connectivity and energy conservation.

5.3.2 Naive Bayes Classification

Naïve Bayes classification is a supervised machine learning based approach which is based on Bayes theorem. Bayes theorem is applied to determine the posterior probability of occurrence of an event. The Naïve Bayes classification approach can be categorized into three types as Gaussian Naïve Bayes Classification, Bernoulli Naïve Bayes Classification and Multinomial Naïve Bayes Classification model. Gaussian Naïve Bayes classification approach works on the input data which follows Gaussian distribution and are continuous in nature. Similarly Multinomial Naïve Bayes classification approach also works for the continuous datasets. The Bernoulli Naïve Bayes approach works for the data set which follows the Bernoulli distribution and produces the discrete output. The Naïve Bayes classification based approach has been used in solution of different WSN issues such as anomaly detection, localization, routing, coverage and tracking of the targets or sink nodes.

5.3.3 Random Forest based Classifier

Random Forest based classification is a supervised machine learning approach. In this approach several decision trees are constructed where each tree represents a classification. The random forest based classification approach consists of two steps: in the first step, the random forests are created and in the second step, the classifier is used to predict the class labels for the new test data. The random forest based classification approach has been used in solving the coverage problem and designing the Medium Access Control (MAC) protocols.

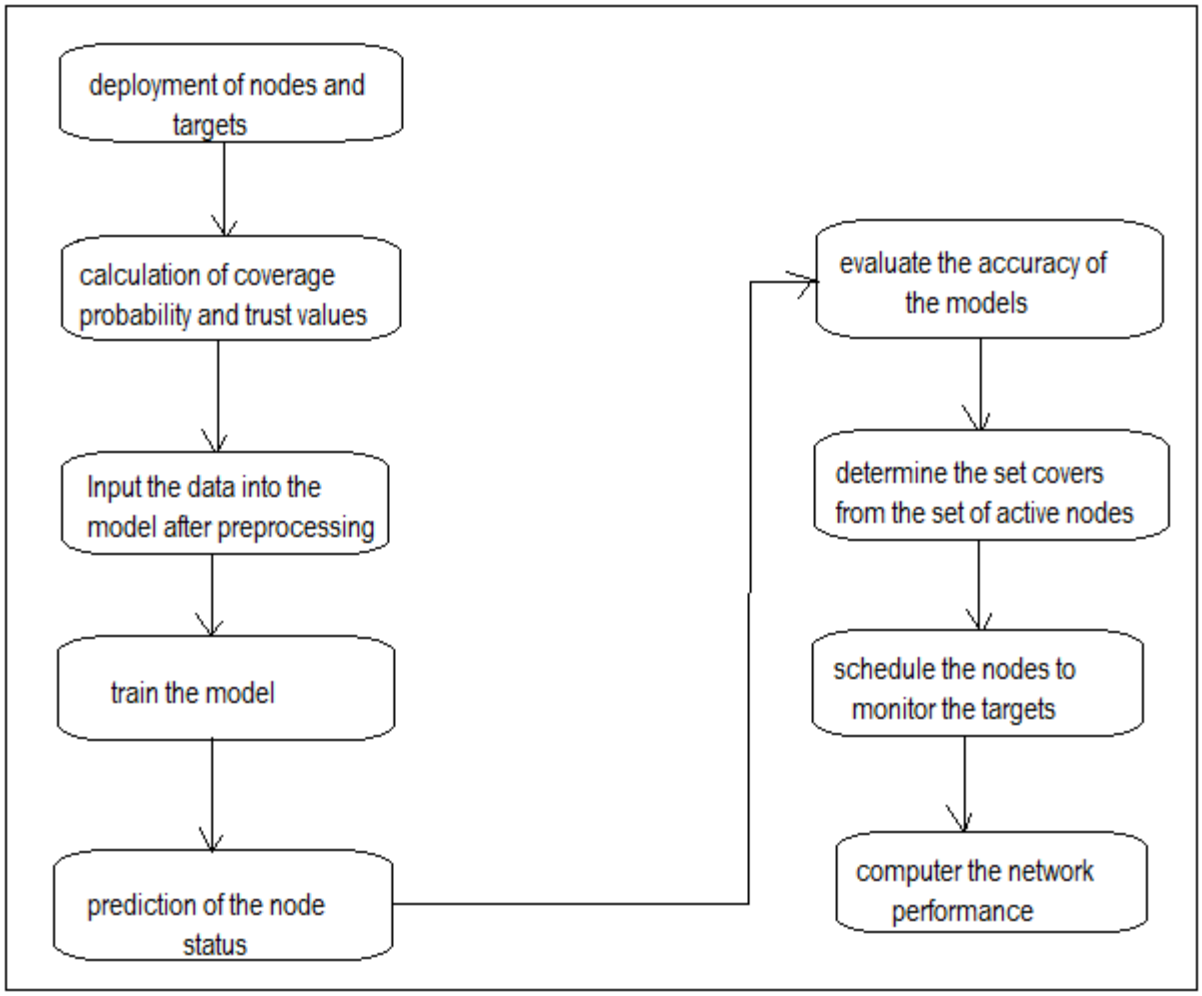
5.3.4 Support Vector Machine Based Classifier

The support vector machine is a supervised machine learning which can be applied to various classification problems. The support vector based approach aims to maximize the separation between the datasets of different classes. The data points which aid in maximizing this separation are known as support vectors. The SVM based approach complexity is independent of the number of attributes considered in the data set. SVM based approaches have been applied in the literature to solve the different problems in WSN such as connectivity, routing, fault detection and congestion control.

6. Proposed Approach and Methodology

The different parameter values of the network are collected after the random deployment of the network. The proposed approach consists of following steps as shown in the Fig.1:

1. Deploy the nodes and targets in the region of interest randomly.
2. Determine the coverage probability and trust values of the nodes.
3. Input the parameters in the prediction model.
4. Train the model.
5. Predict the node status.
6. Evaluate the accuracy of the models.
7. Determine the set covers from the set of active nodes.
8. Schedule the nodes to monitor the targets.
9. Compute the network lifetime.



Fig, 1. Steps of the proposed model

6.1 Performance Metrics

There are four key parameters required for the evaluation of the machine learning based prediction model as:

i. *True Positive (TP)*- This parameter represents the number of correctly labeled tuples whose outcome was positive in nature.

ii. *False Positive (FP)* – This parameter represents the number of incorrectly labeled tuples whose outcome was positive in nature.

iii. *True Negative (TN)*- This parameter represents the number of correctly labeled tuples whose outcome was negative in nature.

iv. *False Negative (FN)*- This parameter represents the number of incorrectly labeled tuples whose outcome was negative in nature.

Based on these parameters the performance metrics used for the evaluation of a machine learning based model are discussed in this section. Let *P* represents the number of positive outcomes and *N* represents the number of negative outcomes.

a. *Precision*- This metric represents the number of positive tuples which are actually positively labeled. In other words, it represents the correctness level of the model. The precision parameter is calculated as shown in the equation 5:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{5}$$

b. *Recall*- This performance metric represents the number of positive tuples correctly identified by the model. It is calculated as shown in the equation 6:

$$\mathrm{Re}\, call = \frac{TP}{P} \tag{6}$$

c. *F1-Score*- This parameter represents the weighted average of precision and recall as shown in the equation 7:

$$F1\ Score = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{7}$$

d. *Accuracy*- This metric represents the accuracy of the model and is mathematically represented as shown in the equation 8:

$$Accuracy = \frac{TP + TN}{P + N} \tag{8}$$

e. *Confusion Matrix*- Confusion matrix is considered as an important representation of the machine learning model. The different parameters of the machine learning model are represented in the square matrix as shown in the Fig.2.

Fig. 2 Confusion Matrix

TP	*FP*
TN	*FN*

6.2 Experimental set up and simulation parameters

We have considered a network of 10 nodes and 5 targets which are randomly deployed in the region of dimension 50*50 square meters. The other simulation parameters considered in the proposed approach are as shown in the Table 2:

Table 2. Experimental parameters

Parameter	Value
Number of nodes	10-50
Number of targets	5-10
Target area	50*50 square meter
Initial energy of the node	1 J
Sensing Range	10 m
Detection error range	5 m
Energy consumed in the electronics circuit to transmit or receive the signal (E_{elec})	50nJ/bit
Energy consumed by the amplifier to transmit at a short distance (E_{fs})	10pJ/bit/m^2
Energy consumed by the amplifier to transmit at a longer distance (E_{amp})	0.0013pJ/bit/ m^4
Data Aggregation Energy (E_{DA})	5nJ/bit/report
Packet size (L)	500 bytes
Regulatory factor (R)	0.5
Required coverage level for each target	0.5

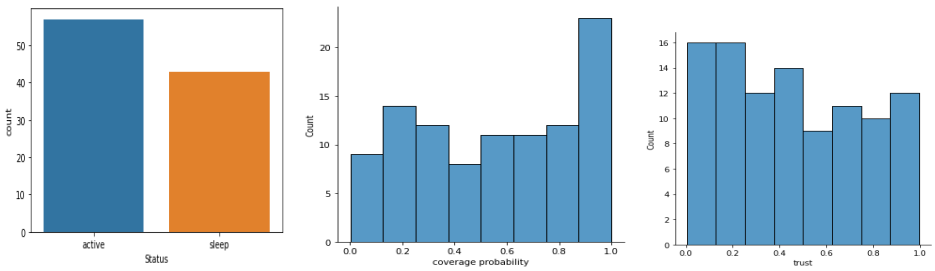
7. Results and Discussion

The pre requisite of designing a prediction model is data preprocessing step, which implies preparing the data set for prediction. The data pre processing approach basically aims to visualize the data values with an objective of finding the correlation between the different features on the target attribute. Based on the information gain parameter the most relevant features are identified as shown in the Table 3.

Table 3. Most relevant attributes

S. No.	Feature Name
1	Node Id
2	Coverage Probability
3	Trust
4	No of Communications
5	No of Packets
6	CH or not
7	Degree
8	Energy Consumed
9	Status

The node status parameter is considered as the target attribute for the prediction model, where as the remaining attributes are considered as the input parameters. To determine the number of active and sleep nodes, the count value is plotted as shown in the Fig. 3. To analyze the values of the coverage probability, the histogram is plotted for all the values as shown in the Fig. 4. Similarly the histogram for the trust values of the nodes is plotted as shown in the Fig.5.



(3) **(4)** **(5)**

Fig.3. Number of nodes vs status 4. Histogram of coverage probability 5. Histogram of trust

To analyze the relationship of all the other parameters on the node status, the correlation heatmap is plotted as shown in the Fig. 6.

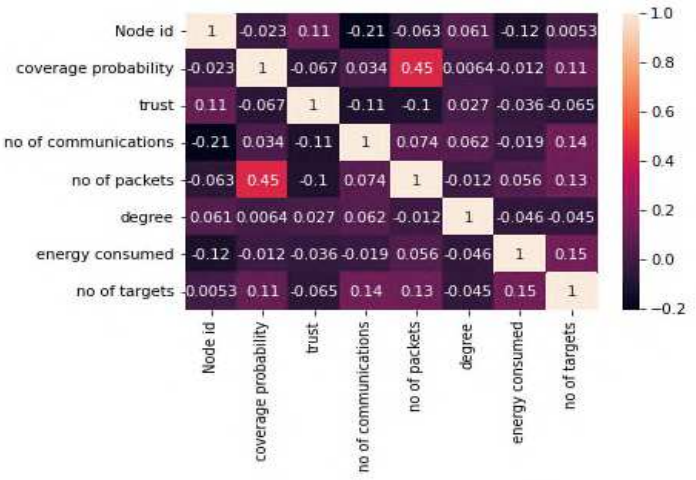


Fig. 6. Heatmap plot of different parameters

7.1 Comparison of different models

We have designed several prediction models utilizing different machine learning techniques such as Logistic regression, Gaussian naïve Bayes, Bernoullie Naïve Bayes, Random Forest and Support Vector based classification. The dataset is divided into two parts as training and testing set. The training set consists of 70% data values and the test size is considered as 30%. The performance metrics for the different models is as shown in the Table 4.

Table 4. Comparison of different prediction models

Model		Precision	Recall	F-1 Score	Support	Accuracy
Logistic Regression	Active	1	0.96	.98	26	0.97
	Sleep	.93	1	.97	14	
Gaussian Naïve Bayes	Active	1	.54	.7	26	0.7
	Sleep	.54	1	.7	14	
Bernoullie Naïve Bayes	Active	1	0.96	.98	26	0.975
	Sleep	.93	1	.97	14	

Random Forest Classifier	Active	1	1	1	26	1.00
	Sleep	1	1	1	14	
Support Vector Machine (Linear Kernel)	Active	.96	.92	.94	26	0.925
	Sleep	.87	.93	.90	14	
Support Vector Machine (Polynomial Kernel)	Active	.96	.92	.94	26	0.925
	Sleep	.87	.93	.90	14	
Support Vector Machine (Radial Basis Kernel)	Active	.96	.92	.94	26	0.93
	Sleep	.87	.93	.90	14	
Support Vector Machine (Sigmoid Kernel)	Active	0.96	0.96	0.96	26	0.95
	Sleep	0.93	0.93	0.93	14	

The result shows that the accuracy of the Random forest classifier is maximum for the considered dataset. The different variant of support vector machine based classification is used for the different kernel functions such as Linear, Polynomial, Radial Basis and Sigmoid kernel function as clf, clf2, clf3 and cl4. The results show that the sigmoid kernel function based prediction model has highest accuracy as compared to other SVM based models. The accuracy of Logistic Regression model is 97% initially but it can be further improved by incorporating the hyper parameter tuning mechanism. The results show that the best parameters with accuracy 100% and score 98% for the parameter value as 1 and laso regularization technique for the penalty.

Tuned hpyerparameters ☹est parameters) {'C': 1.0, 'penalty': 'l1'}
accuracy : 1.0
score 0.975

The accuracy score for the different prediction model is as shown in the Fig. 7.

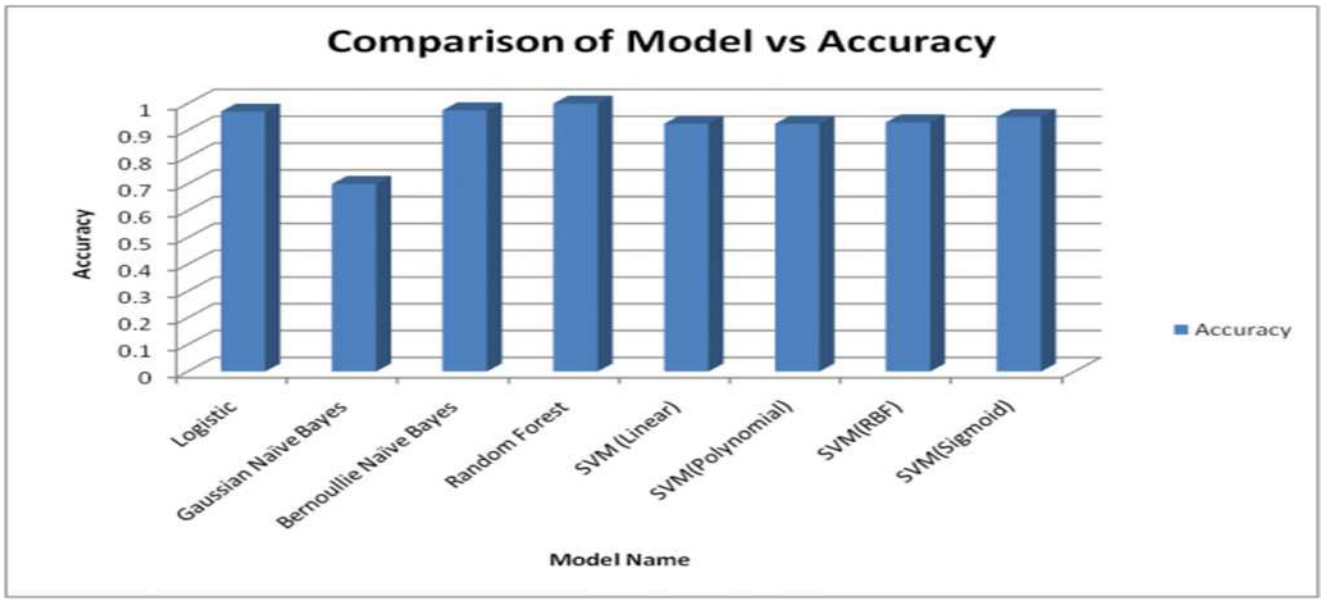


Fig. 7. Comparison of Accuracy of different models

We have also analyzed the accuracy of the polynomial kernel function for different degree values ranging from 1 to 40. The accuracy of the different polynomial kernel function is as shown in the Fig. 8. It can be seen from the figure that there is no improvement in the accuracy beyond the degree 20, so we have only considered the degree of polynomial kernel function in the range of 1 to 20.

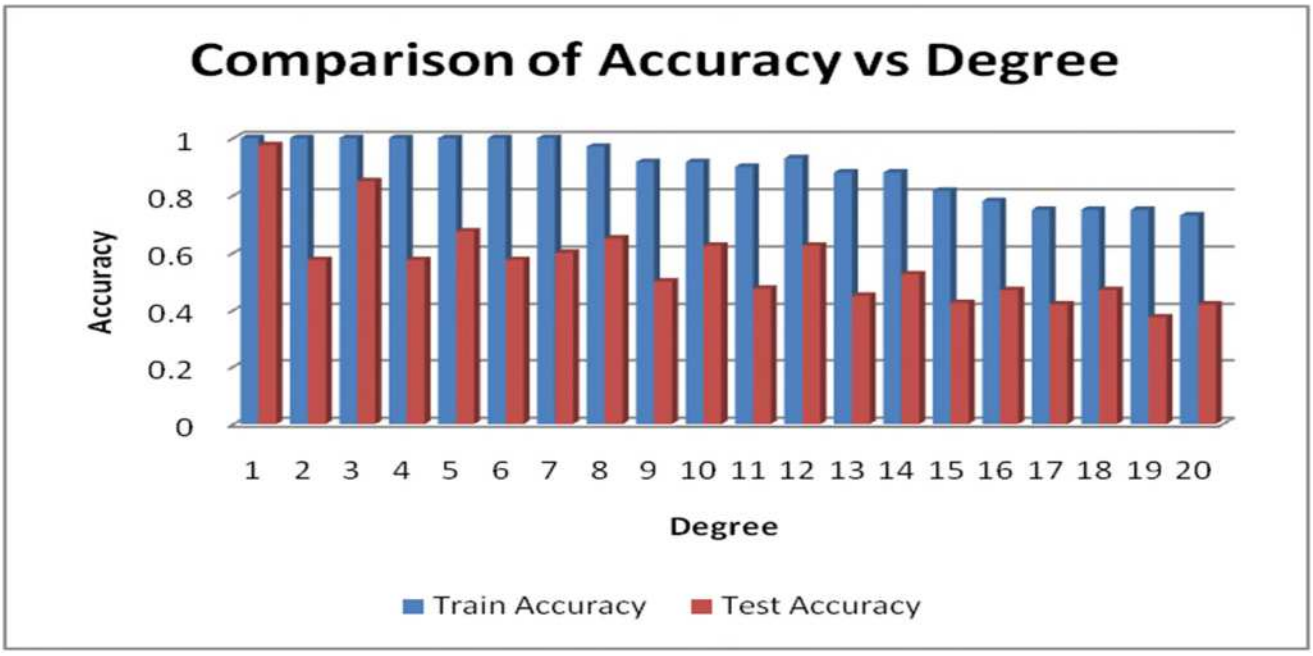


Fig.8. Comparison of Accuracy for different degrees of polynomial kernel function

The grid search cross validation is incorporated to evaluate the performance of different models. The results of different models are as shown in the Table 5. The results show that the Gaussian Naïve Bayes and Bernoullie Naïve Bayes outperforms the other approaches in terms of standard deviation and mean values. The box plot of the accuracy of the different models is shown in the Fig. 9. It can be seen from the figure that the random forest based model outperforms the other approaches.

Table 5. Evaluation of prediction models using cross validation

Model	Cross Validation Results Mean	Cross Validation Results Standard Deviation
Logistic Regression	0.925	0.114564
Gaussian Naïve Bayes	0.975	0.075000
Bernoulli Naïve Bayes	0.975	0.075000
Random Forest Classifier	0.95	0.100000
Support Vector Machine (Linear Kernel)	0.90	0.122474
Support Vector Machine (Polynomial Kernel)	0.90	0.122474
Support Vector Machine (Radial Basis Kernel)	0.90	0.122474

Support Vector Machine (Sigmoid Kernel)	0.90	0.122474

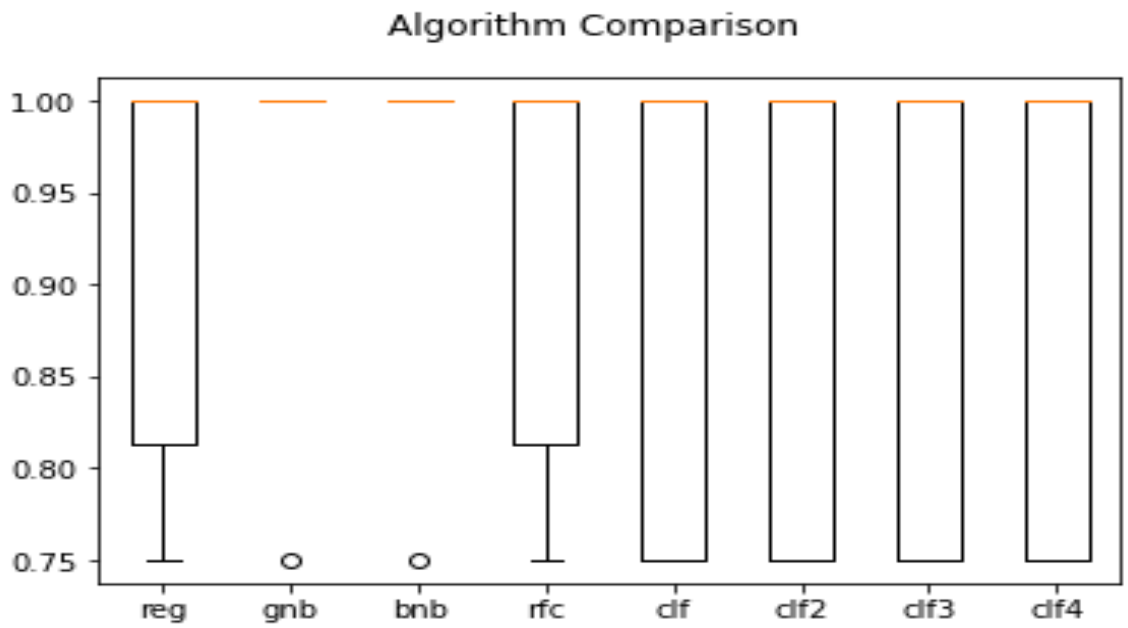


Fig. 9 Comparison of different models

8. Conclusion and Future Scope

The paper incorporates the machine learning based approaches to predict the node status. The most relevant features contributing to the node status prediction have been identified using the correlation analysis. The standard scalar based preprocessing technique has been applied to prepare the input dataset. The different machine learning based models using Logistic Regression, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Random Forest based classification, different kernel functions based SVM models has been designed based on the historical data of coverage probability and trust values. The performance of these models has been evaluated on the basis of different performance metrics. The results show that the Random Forest based model provides the best accuracy for the considered network settings. The models had been fine tuned using 5 fold cross validation technique and the best hyper parameters have been designed. To analyze the network performance in terms of network lifetime and throughput is our future scope. The performance of the proposed approach can be further enhanced by utilizing the deep learning based concepts.

9. Funding

There is no funding received for this work.

10. Competing Interest

The authors declare that there is no conflict of interest.

11. Author Contributions

The authors declare that all the authors have contributed equally in carrying out this research work.

12. Data Availability

The dataset is not publicly available, but it may be provided on reasonable request.

13. References

[1] P. Rawat, K.D. Singh, H. Chaouchi , J.M. Bonnin, Wireless sensor networks: a survey on recent developments and potential synergies, J. Supercomput 68 (1) (2014) 1–48 .

[2] I. F. Akyildiz , W. Su , Y. Sankarasubramaniam , E. Cayirci , Wireless sensor networks: a survey, Comput. Networks 38 (4) (2002) 393–422.

[3] J. Yick, B. Mukherjee, D. Ghosal , Wireless sensor network survey, Comput. Net- works 52 (12) (2008) 2292–2330.

[4] Mulligan, R. and Ammari, H.M. (2010) 'Coverage in wireless sensor networks: a survey', Network Protocols and Algorithms, Vol. 2, No. 2, ISSN: 1943-3581

[5] Thai, M.T., Wang, F. and Du, D-Z. (2008) 'Coverage problems in wireless sensor networks: designs andanalysis', International Journal of Sensor Networks, May, Vol. 3, No. 3, pp.191–200.

[6] M.A. Alsheikh , S. Lin , D. Niyato , H.-P. Tan, Machine learning in wireless sensor networks: algorithms, strategies, and applications, IEEE Commun. Surv. Tutorials 16 (4) (2014) 1996–2018.

[7] D.K. Sah, T. Amgoth, Parametric survey on cross-layer designs for wireless sensor networks, Comput Sci. Rev. 27 (2018) 112–134.

[8] J.B. Predd, S.B. Kulkarni , H.V. Poor , Distributed learning in wireless sensor net- works, IEEE Signal Process. Mag. 23 (4) (2006) 56–69.

[9] T.M. Mitchell, Machine Learning, first ed., McGraw-Hill, Inc., New York, NY, USA, 1997.

[10] T.O. Ayodele , Introduction to Machine Learning, first ed, InTech, 2010 .

[11] P. Langley, H.A. Simon, Applications of machine learning and rule induction, Com- mun. ACM 38 (11) (1995) 54–64.

[12] D. Praveen Kumar, Tarachand Amgoth, Chandra Sekhara Rao Annavarapu, Machine learning algorithms for wireless sensor networks: A survey, Information Fusion, Volume 49, 2019,Pages 1-25,

[13] N. Jaggi, and A. A. Abouzeid, Energy-Efficient Connected Coverage in Wireless Sensor Networks, Proc. 4th Asian International Mobile Computing Conference (AMOC), pp. 77-86, January 2006 .

[14] XiaochunXu and Sartaj Sahni, Approximation Algorithms for Wireless Sensor Deployment, April 21, 2006.

[15] Z. Zhou, S. Das and H. Gupta. Connected k-coverage problem in sensor networks. In Proc. of International Conference on Computer Communications and Networks (ICCCN'04), Chicago, IL, October 2004, pp. 373–378.

[16] Wang, J.; Zhong, N. Efficient point coverage in wireless sensor networks. J. Comb. Optim. 2006, 11, 291–304.

[17] M. Cardei, M. T. Thai, and Y. Li. Energy-efficient target coverage in wireless sensor network. Proc. Infocom' 05, May 2005.

[18] Cardei, M., Du, D. Improving Wireless Sensor Network Lifetime through Power Aware Organization. Wireless Netw 11, 333–340 (2005). https://doi.org/10.1007/s11276-005-6615-6.

[19] .Cardei, I., & Cardei, M. (2008). Energy-efficient connected-coverage in wireless sensor networks. International Journal of Sensor Networks, 3(3), 201-210.

[20] Chaturvedi, P. and Daniel, A.K. (2017). 'A novel sleep/wake protocol for target coverage based on trust evaluation for a clustered wireless sensor network', Int. J. Mobile Network Design and Innovation, Vol. 7, Nos. 3/4, pp.199–209.

[21] V. Narayan, A. K. Daniel and A. K. Rai, "Energy Efficient Two Tier Cluster Based Protocol for Wireless Sensor Network," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 2020, pp. 574-579, doi: 10.1109/ICE348803.2020.9122951.

[22] Narayan, V., & Daniel, A. K. (2019). Novel protocol for detection and optimization of overlapping coverage in wireless sensor networks. *Int. J. Eng. Adv. Technol*, *8*.

[23] Narayan V., Daniel A.K. (2021) RBCHS: Region-Based Cluster Head Selection Protocol in Wireless Sensor Network. In: Singh Mer K.K., Semwal V.B., Bijalwan V., Crespo R.G. (eds) Proceedings of Integrated Intelligence Enable Networks and Computing. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-33-6307-6_89.

[24] Narayan V., Daniel A K (2021). A Novel Approach for Cluster Head Selection using Trust Function in WSN, Scalable Computing: Practice and Experience, Vol 22, No. 1, https://doi.org/10.12694/scpe.v22i1.1808

[25] Zahra Taghikhaki, Nirvana Meratnia, Paul J.M. Havinga, "A Trust-based Probabilistic Coverage Algorithm for Wireless Sensor Networks", 2013 International Workshop on Communications and Sensor Networks (ComSense-2013), Procedia, Computer Science 21 (2013), 455 – 464.

[26] Jinfang Jiang, Guangjie Han, Feng Wang, Lei Shu, Mohsen Guizani, "An Efficient Distributed Trust Model for Wireless Sensor Networks", IEEE Transactions on Parallel and Distributed Systems, 2014.

[27] Hyo Sang Lim, Yang Sae Moon, Elisa Bertino, Provenance based Trustworthiness Assessment in Sensor Networks, DMSN'10, September 13, 2010, Singapore.

[28] Heinzelman, W. R., Chandrakasan, A. & Balakrishnan, H. (2000). Energy-efficient communication protocol for wireless micro sensor networks, HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8, IEEE Computer Society, Washington, DC, USA, p. 8020.

[29] A. Farhat, C. Guyeux , A. Makhoul , A. Jaber , R. Tawil , A. Hijazi , Impacts of wireless sensor networks strategies and topologies on prognostics and health management, J. Intell. Manuf. (2017) 1–27.

[30] W. Sun, X. Yuan, J. Wang , Q. Li , L. Chen , D. Mu , End-to-end data delivery reliability model for estimating and optimizing the link quality of industrial WSNs, IEEE Trans. Autom. Sci. Eng. 15 (3) (2018) 1127–1137.

[31] X. Chang , J. Huang , S. Liu , G. Xing , H. Zhang , J. Wang , L. Huang , Y. Zhuang , Accu- racy-aware interference modeling and measurement in wireless sensor networks, IEEE Trans. Mob. Comput. 15 (2) (2016) 278–291.

[32] W. Kim , M.S. Stankovi , K.H. Johansson , H.J. Kim , A distributed support vector machine learning over wireless sensor networks, IEEE Trans. Cybern 45 (11) (2015) 2599–2611.

[33] J. Shu , S. Liu , L. Liu , L. Zhan , G. Hu , Research on link quality estimation mechanism for wireless sensor networks based on support vector machine, Chin. J. Electr. 26 (2) (2017) 377–384.

[34] W. Elghazel , K. Medjaher , N. Zerhouni , J. Bahi , A. Farhat , C. Guyeux , M. Hakem , Random forests for industrial device functioning diagnostics using wireless sensor networks, in: Aerospace Conference, 2015 IEEE, IEEE, 2015, pp. 1–9.

[35] B. Yang , Y. Lei , B. Yan , Distributed multi-human location algorithm using naive bayes classifier for a binary pyroelectric infrared sensor tracking system, IEEE Sens. J. 16 (1) (2016) 216–223.

[36] J. Qin , W. Fu , H. Gao , W.X. Zheng , Distributed k -means algorithm and fuzzy c -means algorithm for sensor networks based on multi agent consensus theory, IEEE Trans. Cybern. 47 (3) (2017) 772–783.

[37] E. Ancillotti , C. Vallati , R. Bruno , E. Mingozzi , A reinforcement learning-based link quality estimation strategy for RPL and its impact on topology management, Com- put. Commun. 112 (2017) 1–13.

[38] H. Chen , X. Li , F. Zhao , A reinforcement learning-based sleep scheduling algorithm for desired area coverage in solar-powered wireless sensor networks, IEEE Sens. J. 16 (8) (2016) 2763–2774.

[39] Y. Xu , O. Ding , R. Qu , K. Li , Hybrid multi-objective evolutionary algorithms based on decomposition for wireless sensor network coverage optimization, Appl. Soft Comput. 68 (2018) 268–282.