International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-11 Issue-3, February 2022

Application of Advanced Machine Learning and Artificial Neural Network Methods in Wireless Sensor Networks Based Applications

Suvarna. S. Patil, B. M. Vidyavathi

Abstract*: Wireless sensor networks (WSN) have become very interesting areas of study in computer science these days. WSN is a group of sensors that are used in the physical world. It's easy to see how small these sensors are. They can detect physical wonders and help you deal with them. One of the most important reasons for distributing WSN-built-up applications is to make a decision about what to do next, which has been difficult because WSNs have limited processing power, limited storage space, and a lot of quickly changed data. This makes it important to look into new and appropriate data mining methods that can extract learning from a huge amount and a wide range of information that WSNs send out all the time. Machine Learning algorithms like Random Forest Regression and Artificial Neural networks are used in wireless sensor network-based applications like this one. This article also presents a comprehensive comparison of results from the application of Random Forest Regression and Artificial Neural Network algorithms to WSN data.

Keywords*: Wireless Sensor Networks, Data Mining, Random Forest Regression and ANN Algorithm.

I. INTRODUCTION

Wireless Sensor Network (WSN) is a wireless network that includes a great number of little sensor gadgets which are tightly positioned and beneficial in various zones such as detecting moisture [1], affluence, supervising link, object following, info scrutiny, etc. In utmost appliances, effectual and instantaneous supervising are vital necessities. In limited appliances, should perform this vast quantity of data administering at extremely great rapidity. This consequently has an advanced necessity for novel effective structures, phases, and required tools for managing this huge sensory data that is naturally systematized or amorphous. The growth of WSN was originally inspired by military uses like combat zone inspection, country-wide safety, medicinal facilities, home apparatuses, environmental examination, and several areas in which human arbitration is significantly dangerous [2]. The wireless sensor network is an appropriated insightful organization framework. It is made out of numerous miniature sensor hubs sent in the location region, which have the capacity of remote correspondence and processing. WSNs architecture is shown in Fig 1.

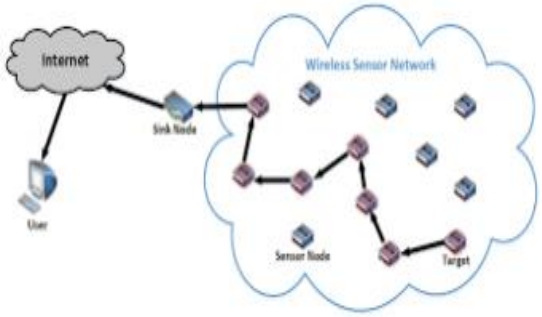

Fig 1: WSNs Architecture

There are a lot of basic things that make up Wireless Sensor Networks. They have a sensor hub group, memory, limited battery power, single or multiple base stations (BS), hub dynamicity, single-jump broadcast or multisource broadcast, data recurrence, information idleness, and more. With new technology, sensor nodes are getting smaller, but wireless networks are getting bigger, with thousands of nodes now. Nonetheless, there will be millions of nodes forthcoming. Trading with the enormous data generated from these sensor networks generates an extremely dangerous test, one reasonably advanced by the usage of Data Mining methods. Frequent pattern mining, followed by consecutive pattern mining, clustering, and classification are the main data mining methods used for WSNs. They are shown in Fig 2. All of these methods are used in both an integrated and dispersed way in WSNs, too. Data mining in sensor networks is a way to find examples and designs that can be used in real-world situations because of the continuous, fast, and logically possible flow of information by sensor networks. In this case, all the information can't be stored and must be used right away. Data mining algorithms should be able to process a lot of information quickly enough. Traditional data mining algorithms are only meant to look at static information. They use multistep and multi-scan mining algorithms to look at static data sets. So, traditional data mining methods can't be used to keep an eye on the huge amounts, high dimensionality, and dispersed nature of information that comes from WSNs.

Manuscript received on January 31, 2022.
Revised Manuscript received on February 02, 2022.
Manuscript published on February 28, 2022.
* Correspondence Author
Mrs. Suvarna.S. Patil*, Assistant Professor, Department of ECE, RYMEC, Ballari, India Email id:: suvarnav001@gmail.com
Dr. B.M. Vidyavathi, Professor and Head, Department of Artificial Intelligence and Machine Learning, BITM, Ballari, India Email id: vidyahm1@gmail.com

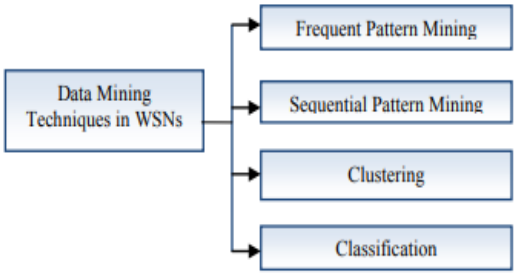

Fig 2: Data Mining Methods

A. Challenges

i) Resource Restriction: Sensor nodes are resources in phrases of power, memory, communique bandwidth, plus computational energy. The primary dispute confronted from data mining approaches for WSNs is assuring mining accuracy, condition while retaining good supply ingesting of WSNs to minimal.

ii) Rapid and Enormous Arrival of Data: The characteristic of WSNs info is that it's over the high speed. In numerous spaces, learning arrives quicker compared to skilful mining. Besides, spatiotemporal inserting of sensor information performs a fundamental position in WSNs function. This might at times reason numerous established information preparing strategies for performing incompetently on spatiotemporal sensor knowledge. The task for data mining procedures is to adapt to the enduring, fast, and adjusting learning streams and discover how to contain client connection throughout rapid data arrival.

iii) Web-based Mining: In WSNs, condition learning is topographically dispensed, inputs arrive reliably, and later information devices could substitute the result focused on more established data observably. Most data mining methods that break down learning in disconnected strategy don't meet the necessity of dealing with dispersed circle data. Subsequently, a wander for information mining strategies is the correct method to framework apportioned spilling information on the web.

iv) Displaying changes of mining outcomes after some time: When the information creating wonder is altering after some time, the separated model ought to be cutting-edge. Considering the congruity of data streams, a few specialists have expressed that taking pictures of the other of mining is essential than the mining result. The examination inconvenience is the best move toward to model this adjustment in the result

v) Data Revolution: Considering how sensor nodes are restricted as far as bandwidth, rebuilding formed information over the system isn't practical. Favourable position structure change is an important inconvenience. Yield is exchanged to the base station after disengaging the mannequin and examples locally by Wireless Sensor Networks info. Assignment for information mining framework is ideal for effectively speaking to data and scholarly examples over system for broadcast.

vi) Energetic Network Topology: Sensor organizes conveyed in, most likely cruel, uncertain, heterogeneous, and dynamic situations. Besides, sensor hubs may exchange among certain areas at any factor after some period. Such dynamicity and conglomeration build up the unpredictability of outlining a fitting data mining process for Wireless Sensor Networks.

II. LITERATURE SURVEY

The most important reason to show the WSNs-built-up applications is to make the ongoing solution, which has been very difficult because of the limited resources, limited processing, and the huge amount of quickly changed information that WSNs make. This makes us want to look into new and appropriate data mining methods that can get information from a lot of and different data coming in from WSNs.

In [4,] a number of different grouping and evaluation methods are used to look at different current data mining strategies for WSNs. Based on problems with the current process, an adaptive data mining structure for Wireless Sensor Networks is being planned for an upcoming study.

In the dispersed data extraction process [5,] the writer came up with a way to gather information about sensor nodes. This method reduces the number of messages that are sent out during broadcast; at the same time, the cost of the node buffer increases and the time it takes to send a few critical messages rises. Taherkordi et al. in [6] said that they would use a spread-out method to get sensor data. The K-Mean algorithm was shown in a dispersed form and short information was directed to sink. This will cut down on the number of communications. Thus, the amount of time and power sensor nodes use are limited. The main issues with this method are that the cluster head needs a lot of extra memory. This, too, needs a lot of math power to make sure that the information gathered is concise before it is sent to the sink. Tianjin Wang and Zhen Yang came up with the Location-Aware Based Data Clustering (LABDC) algorithm. In the trial, it was found that this algorithm reduces the number of communications and broadcast prices. The main problem with this algorithm is how well it works when there are a lot of sensors sharing information. In [7], writers Longjiang Guo and Chunyu Ai projected an H-cluster algorithm. This utilizes the dispersed method to cluster sensor info. This algorithm uses a group of sensor info as input. In this, sensor info is information gathered at every sensor node from the beginning to the present. After condensing groups of input sensor info-group, the algorithm generates a group of cluster traits as output. Cluster algorithm is observed to be extra effective in little WSN, seeing limits like information damage, energy efficacy, and excellence of cluster-info to be controlled. Nonetheless, the problem in this technique is that, as the quantity of information brought from sensors upsurges viz. for big gage WSNs, a quantity of info loss also upsurges. Likewise, the energy efficacy of networks declines from the rising magnitude of networks. In [8], writer Tanveer et al. Gave tree-based info assembly known as sensor pattern tree (SP-tree). It is utilized for generating connotation instructions by sensor networks info from a single database scan. In this, the writer has premeditated for getting every event-sensing sensor info, and prefix-tree is built in every recognized type founded on that. Founded on this, the tree in the diagonal frequency pattern is updated.

Retrieval Number: 100.1/ijeat.C33940211322
DOI: 10.35940/ijeat.C3394.0211322
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The outcome in the document demonstrates that SP-tree accomplishes fine PLT in time and memory ingesting. Nonetheless, the transmission efficacy of structure declines for a comparatively great amount of sensors. S. Nithyakalyani and S. Suresh Kumar in [9] projected KMeans Data Relay (KMDR) clustering algorithm for assemblage sensor nodes from that place, decreasing the no of nodes communicating info to sink node declines transmission above and, as this way, upsurge network functioning. Conserve and Observe Modes (COM) algorithm decreases the no of nodes inside the cluster. With no negotiating exposure, it confronts main disputes like restricted transmission bandwidth, restrictions in power source, and space properties area. The commitment of K- MDR is to reduce power usage; at last, the recreation results show that the time proficiency of the accomplished. In [10], writer Brahim Elbhiri and Sanaa projected a new algorithm, Spectral Cataloguing for Vigorous Clustering in Wireless Sensor Networks (SCRCWSN). This spectral division technique utilizes graph theory methods to separate networks in an ideal quantity of clusters. Replication outcomes show that the projected algorithm upsurges the entire network's duration and offers extra energy efficacy dispersal contrasted to Low-Energy Adaptive Clustering Hierarchy (LEACH) method and Centralized LEACH (LEACH-C) one. The writers of [11] advise a mixture of Wireless Local Area Network (WLAN) and sensor-node knowledge to detect fire. The structure is included of multi-sensor nodes and IP cameras in a wireless mesh system for detection and verifying fire in countryside and jungle zones in Spain. Upon sensing fire from wireless multi-sensor node, warning produced from the node is broadcasted towards central server where software use works for choosing nearby wireless camera(s). Later immediate images by region are flowed to sink. Here, sensor nodes are positioned with great space amid every node in which info by sensors and cameras is gathered and managed at the base station. Nevertheless, our projected structure contemplates a gathered positioning plan in which spaces amid neighbouring sensors nodes are slightly small. Like this, we aim to detect forest fire extra rapidly and direct corresponding data to the base station as quickly as probably. In [12], writers present a comparative study of numerous Data Mining techniques on WSN fire recognition info utilizing the WEKA device. The aim was to realize which of those possess the finest cataloging precision of fuzzy logic produced info and is very suitable for precise use of fire recognition. Real sensing data has been used in this work, and simulation is done under the circumstances near actuality. WSN offer excellent checking at an extremely low unit cost as far as capital and work contributed. In [13] Mangundu et al., portrayed a WSN-empowered design for a tipping pail downpour check framework to gather and communicate ongoing information utilizing General Pocket Radio Service (GPRS) through a cell organization. The information is sent from remote Rain Gauge (RG) stations to a public web server called as the Weather Underground (WU). In [14] Prajwala et al., aimed is to build decision rules to predict rainfall using random forest algorithm with the help of 5 attributes, namely humidity, temperature, atmospheric pressure, wind direction with target variable being amount of rainfall. Data was collected for 192 days using TOT sensors. The adaptive boost random forest classifier gives a highest accuracy for prediction of rainfall. The decision rules governing the amount of rainfall is identified. The random forest regressor is used to identify the Mean Squared Error.

III. PROPOSED METHODOLOGY

The proposed research work is given below:

3.1 Random Forest Regression (RFR) model

RF algorithm is mostly a mixture of Bagging and random subspace [15] algorithms and was restricted from Leo Breiman as a "mixture of tree forecasters like every tree hangs on arbitrary vector values tried individually and has identical dispersal for every tree in a forest. This technique is occasioned by the number of developments in tree classifiers' precision [16]. Random forest is managed learning algorithm. The "forest" it constructs is a collective of decision trees, typically skilled by the "bagging" technique. Overall knowledge of bagging technique is that mixture of learning replicas upsurges complete outcome. Random forest enhances extra arbitrariness to model, whereas increasing trees. Instead of exploring very significant characteristics through the excruciating node, this quests for the finest characteristic amid an arbitrary subgroup of characteristics—these outcomes in a varied range, usual outcomes in an improved model.

3.1.1 Feature Importance

An additional illustrious characteristic of the random forest algorithm is that it is extremely simple for measuring the comparative significance of every characteristic on the forecast. Random Forest delivers immense device for this which amounts characteristic's significance from watching how much tree nodes that utilize that characteristic decrease contamination through every tree in the forest. This calculates grade instinctively for every characteristic after preparation and gauges outcomes, so the totality of every significance is equivalent to one.

3.1.2 Important Hyperparameters

Hyperparameters in random forests are either utilized to increase the projecting supremacy of the model or make the model quicker. Let's gauge at hyperparameters of sklearns in-built random forest purpose.

A. Increasing the predictive power

Initially, there is **n_estimators** hyperparameter, which is the only amount of trees algorithm creates previously captivating highest polling or captivating means of forecasts. Typically, a greater quantity of trees upsurges functioning, and brands forecasts extra steady, but this also decelerates calculation. One more significant hyperparameter is **max_features**, the highest number of characteristics random forest contemplates for splitting node. Random Forest offers numerous choices, all labeled in certification. The final significant hyperparameter is **min_sample_leaf.** This controls the least amount of leaves compulsory for splitting an interior node.

b. Increasing the model's speed

The **n_jobs** hyperparameter expresses several engine processors that are permitted for usage. If it possesses a quantity of 1, it could just utilize a single processor. Quantity of "-1" denotes limitlessness.

The **random_state** hyperparameter brands model's output is replicable. The model will continually yield identical outcomes when it comprises a certain number of random stages and provides identical hyperparameters and training info. Finally, there is **oob_score** (known as oob sample), which is a random forest cross-authentication technique. Here, about one-third of the info is not utilized for the training model and could evaluate its functioning. These examples are called out-of-bag examples. It's extremely identical to the leave-one-out-cross-authentication technique, nonetheless nearly no extra computational weight to it.

This classifier exploits alteration from inoculating arbitrariness limited selection and reduces prejudice from rising tree to extreme extent (no thinning). In RF, the root of tree i comprise examples by training subgroup S, organized by its subsequent class. Node is lethal if it covers examples of one solitary class or if the amount of examples signifying every class is alike. In another situation, this wants to be extra advanced (no pruning). For this reason, at every node, the characteristic which assurances the finest split is nominated as following.

I. Data attained from selecting characteristic is calculated via:

(a) Entropy of Shannon, which gauges amount of data.

$$Entropy\ (p) = -\sum_{k=1}^{c} P(k/p)X\ \log(P(k/p)) \dots\dots\dots\dots (1)$$

In which p is amount of instances related to location in tree, c is overall amount of classes, k/p signifies portion of instances related to location in tree and branded class k, P (k/p) is amount of elements branded class k at location p.

(b) Gini index that calculates scattering in population

$$Gini\ (x) = 1 - \sum_{j=1}^{n} P\left(k/p\right)^2 \dots\dots\dots\dots\dots (2)$$

In which x is arbitrary example, c is amount of classes, k/p signifies portion of instances related to location in tree and branded class k, P (k/p) is amount of elements branded class k at location p.

II. Finest split is subsequently selected from addition of data by rising tree at specified location, tallying to every characteristic as following:

$$Gain\ (p,t) = f(p) - \sum_{j=1}^{n} P_j\ X\ f\left(p_j\right) \dots\dots\dots\dots (3)$$

In which p matches to location in tree, t signifies trial at division n, P_j is amount of elements at location p and then move to location p_j, f(p) communicates to either Entropy(p) or Gini(p). Trait which gives advanced Gain is designated for splitting node. Ideal training of cataloguing difficulty can be NP hard. Tree groups possess plus of flowing algorithm by diverse preliminary positions and this can improve estimated near-optimal classifier [17].

C. Advantages of RFR

Main benefits of random forest are its flexibility. This is utilized for both relapse and cataloguing duties and this is simple for viewing comparative significance it allocates for input characteristics as well.Random forest is similarly extremely convenient algorithm since default hyper parameters it utilizes frequently create decent forecast outcome. Interrelating hyper parameters is reasonably forthright, and there's not that several of them as well. Main difficulties in machine learning are over fitting, nonetheless almost every time this won't occur because of random forest classifier. If sufficient trees in forest are present, classifier won't over fit model.

3.1.3 Evaluation metrics

Important stage in regression model is evaluating precision of model. Mean Squared Error (MSE), Mean absolute error (MAE), Root Mean Squared Error (RMSE), and

Mean absolute percentage errors (MAPE) are utilized for evaluating functioning of model in regression scrutiny.

a. Mean absolute error signifies mean of total variance amid real and projected estimates in dataset. It gauges mean of balances in dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \dots\dots\dots\dots (4)$$

Where,

$\hat{y}$ – predicted value of y

y - Mean value of y

b. Mean Squared Error signifies mean of squared variance among real and projected estimates in data set. This gauges alteration of balances.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y})^2 \dots\dots\dots\dots\dots (5)$$

c. Root Mean Squared Error is square root of Mean Squared error. This calculates typical difference of balances.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y})^2} \dots\dots (6)$$

d. Mean absolute percentage error (MAPE), is called as mean absolute percentage deviation (MAPD) as well, is gauge of estimating precision of prediction technique in statistics. This generally states precision as ratio described by rule:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{A_t - F_t}{A_t}\right| \dots\dots (7)$$

In which A_t is real number and F_t is predicted number. Their alteration is separated from real number A_t. Complete value in this proportion is totalled for each predicted position in time and split from quantity of fitted points n.

e. Parameters of Random Forest Regression Model using 'ranger'.

Number of trees: Number of trees used for prediction.

Sample size: Dimensions of samples used.

Number of independent variables: Quantity of independent variables used.

Mtry: Quantity of variables to probably divide at every node. Default is (rounded down) square root of number of variables. Otherwise, solitary argument function reverting digit, given quantity of independent variables.

Target Node size: Target node size which is predicted.

Variable Importance mode: One of 'none', 'impurity', 'impurity corrected', 'permutation'. The 'impurity' degree is Gini index for cataloguing, alteration of replies for regression and amount of trial figures (see split rule) for endurance.

Splitrule: For cataloguing and possibility assessment "gini", "extratrees" or "hellinger" having default "gini". For regression "variance", "extra trees", "axstat" or "beta" having default "variance". For existence "logrank", "extratrees", "C" or "maxstat" having default "logrank".

oob.error: Calculate OOB projection fault. Adjust to FALSE for saving calculation period, e.g. for large existence forests.

r.squared: Known as explained variance or coefficient of determination (regression only). Calculated on out of bag info.

3.2 ANN (Artificial Neural Network) Algorithm

ANN is the type of calculating having plan stimulated by the nervous system, precisely, the human brain [18]. They demonstrate the role of neurons and synapses in brain guide to novel calculating methods to optimize difficulty like localization. With the ability of interrelated neurons to exhibit compound behaviour and perform different sensible data processing brands, ANN is perfect for augmenting current algorithms in position discovering if not substitute. ANN's dispersed and parallel processing convulsions behaviour of the majority wireless sensor networks. One significant characteristic of ANN is its capability of adaption. Synapses in the human brain are motives why humans distinguish, sense, and comprehend. Imitating synapses brands, ANN displays adaptable topography in continuous programs, which comprise its compliance and elasticity. WSNs projects challenge scalability, and having ANNs compliance characteristics, WSN can perform localization extra compliantly. With their resemblance in topology, WSN and ANN create a decent tandem. The goal is to introduce computational intelligence to WSN to improve its functionality and have a greater chance of survival. With ANN, WSN can be adaptive to changes and can exhibit more intelligent behaviour [19]. They were producing numerous linear equations at numerous positions. These perceptions are known as neurons or nodes, as they are elementary structures in normal neural networks inside the body.

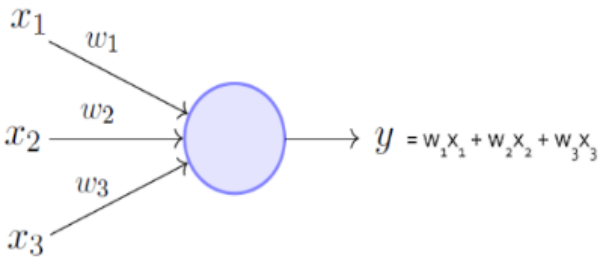

Fig 3: Neural network with Multiple Perceptron's

In the above outline, the primary perpendicular group of three neurons in the input level. Following two perpendicular groups of neurons is a portion of the main level generally denoted as concealed levels, and the last single neuron is output level. The neural network in the above outline is three-leveled. This is due to the input layer is usually not calculated as the share of network levels. Every neuron in input level denotes characteristic (column) in input info (i.e., x1, x2, x3 etc.). Process in the above network is that input info is served to neuron groups, and every neuron gives output. Once more, every output is served to more neurons, which creates extra output that is once more served to output level. Fault computed at this output level is once more directed back in-network for extra refining outputs of every neuron that are once more served to neuron in the output level for producing advanced output compared to earlier. As elucidated in the five-stage procedure above, this procedure is reiterated till we obtain output having minor fault [20].

Implementing ANN for Regression

Every neuron in ANN, excluding input level, creates output. Output is dependent on the type of function utilized. This work is usually mentioned as 'Activation Function' [21]. As ANN is mostly utilized for cataloging functions, the sigmoid function or additional alike cataloguing algorithms are usually utilized as activation tasks. Nevertheless, as we are currently attempting to resolve linear regression difficulty, our activation work in this is a non-entity, but 'Simple Linear Equation' of equation 4 is given below–

$$y=w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \ldots. w_nx_n \ldots\ldots (8)$$

In which x1, x2, x3..xn are liberated features in input info,w1, w3… we are weights (Coefficients) to subsequent features, and w0 is bias. Since our output must merely be a solitary linear layer, ANN must be organized by only one simple neural network layer. This one NN layer's output is a linear line, so this neuron is positioned at the output level. Concealed levels are compulsory when we attempt classifying objects utilizing manifold lines (or curves). So, any concealed levels are not required.Sum of squared errors (SSE) and Mean Squared Error (MSE) are utilized to evaluate the model's functioning in neural network study.SSE is the truly weighted totality of squared errors if the choice of heteroscedastic errors is non-equivalent to a continuous alteration.

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y})^2 \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

Where,

$\hat{y}$ – Dependent Variables

y – Average of Dependent variables

a. Mean Squared Error (MSE) signifies mean of squared change amid real and projected quantities in data set. This gauges change of residues.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \widehat{y})^2 \ldots\ldots\ldots\ldots\ldots (10)$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

3.3 *Dataset*

Intel Dataset: For our review, we utilized the openly accessible Intel Lab dataset [22] which contains information gathered from 54 sensors sent in the Intel Berkeley Research lab. Mica2Dot sensors with climate loads up gathered time-stepped geography data, alongside moistness, temperature, light and voltage esteems every 31 sec. The information was gathered utilizing the TinyDB in-network question preparing framework. The dataset comprises of around 2.3 million readings gathered from these sensors. The organization of the dataset is as per the following: date, time, age, mote ID, temperature, dampness, light, and voltage. The sensor ids range from 1-54.

IV. RESULTS

Experimental results of RFR and ANN models using Intel Labs weather data of WSNs in order to predict humidity, in turn weather are as shown below. Comparative analysis of the proposed techniques is determined using four important metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean absolute percentage errors (MAPE).

4.1. *RFR*

Table 1: Statistics of residuals for RFR

MAE	MSE	RMSE	MAPE
2.1186639	22.8123491	4.7762275	0.2070949

Table 1 shows the statistics of the residuals for RFR model.

Ranger result (RFR Model using 'ranger')

Call:Ranger (formula = humidity ~ ., data = traindat, num.trees = 201, mtry = 2, min.node.size = 3, sample.fraction = 0.8, significance= "impurity")

Table 2: Ranger Result

Type	Regression
Quantity of trees	201
Sample dimension	140002
Quantity of independent variables	3
Mtry	2
Target node dimension	3
Variable importance mode	Impurity
Split rule	Variance
OOB prediction error (MSE)	22.61981
R squared (OOB)	0.9130058

Table 2 shows Ranger function results with various type settings.

b. Evaluation Metrics (Normalized data)

Table 3: Statistics of residuals for RFR with Normalized data

MAE	MSE	RMSE	MAPE
0.0410154	0.0086232	0.092861	---

Table 3 depicts the statistics of residuals for RFR with the normalized dataset. From the results of tables 1 and 3, it is clear that the MAE error has reduced from 2.1186639 to o.o41015358. Hence normalization depicted reduced average error.

A. Ranger Result

Call: ranger (formula = humidity ~., data = traindat, num.trees = 201, mtry = 2, min.node.size = 3, sample.fraction = 0.8, significance= "impurity")

Table 4: Ranger result (Random Forest Regression Model using 'ranger')

Type	Regression
Quantity of trees	201
Sample dimension	140002
Quantity of independent variables	3
Mtry	2
Target node dimension	3
Variable importance mode	Impurity
Split rule	Variance
OOB prediction error (MSE)	0.008402533
R squared (OOB)	0.9139464

Table 4 shows Ranger function results for RFR with various type settings for the normalized dataset. As the data is normalized a lot of variation in the MSE value is observed referring to table 2 and table 4, like 22.61981 to 0.008402533. Also R squared value has increased from 0.09130058 to 0.9139464.

4.2 ANN Results

Table 5: Statistics of Residuals For ANN

SSE_Train	SSE_Test	MSE_Train	MSE_Test
1633.5303	713.7722	0.0233	0.0238

Table 5 lists the statistical values of residuals for the ANN model.

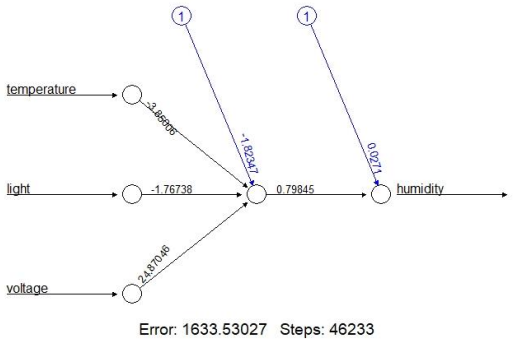

Fig 4: ANN Graph

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The ANN graph in fig 4 depicts that there are three neurons in the input layer and one neuron in the hidden and output layers respectively. Three important key attributes defined for predicting humidity are temperature, light and voltage. The values represented on the perceptron of the graph indicate weightage and biases for the particular sample numbers.

4.3 Comparison Between RFR and ANN Models

RFR algorithm performs better than ANN as shown in Table 6 and accuracy graph is given below.

Table 6: Comparison between RFR and ANN Algorithm

Input	Model	Prediction Accuracy	MAE
Intel Lab data of Wireless Sensor Networks	Random Forest Regression	96%	0.041015358
	Artificial Neural Networks	90%	0.100322235

RFR results analysis depicted better prediction accuracy of 96%, whereas ANN showed 90% using the key parameters as temperature, voltage and light as input to the models.

V. CONCLUSION

Variation in humidity parameter of the weather results in rainfall, affecting many lives. The main purpose of this paper is to normalize the Intel Lab data of WSNs and hence perform humidity prediction using RFR and ANN models with the key variables as temperature, light and voltage for decision making applications like, flood forecasting, landslide detection and many more, to save lives and economy. This work has presented variant machine learning models to predict humidity and in turn rainfall. Among the two presented models better accuracy is observed with various evaluation metrics. The recorded results clearly show that the RFR model is statistically more superior over ANN. As a future work deep neural network model can be used to enhance the prediction accuracy.

REFERENCES

1. Azhar Mahmood, Ke Shi, Shaheen Khatoon, Mi Xiao "Data Mining Techniques for Wireless Sensor Networks: A Survey" International Journal of Distributed Sensor Networks July 2013.
2. Samir Ifzarne, Hiba Tabbaa, Imad Hafidi and Nidal Lamghari, Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks Journal of Physics: Conference Series, Volume 1743, The International Conference on Mathematics & Data Science (ICMDS) 2020 29-30 June 2020 Khouribga, Morocco.
3. S.Stankovic, O.Rakocevic, N. Kojic, D.Milicev, "A Classification and Comparison of Data Mining Algorithms for Wireless Sensor Networks", lCIT2012, 978-1-4673-0342-2012 IEEE.
4. C. Sudha and A. Nagesh, "A Comprehensive Survey on Data Mining Techniques in Wireless Sensor Networks", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-6, June 2018.
5. A. Boukerche and S. Samarah, "An efficient data extraction mechanism for mining association rules from wireless sensor networks," in Proceedings of the IEEE International Conference on Communications (ICC '07), pp. 3936–3941, June 2007.
6. A. Taherkordi, R. Mohammadi, and F. Eliassen, "A communication efficient distributed clustering algorithm for sensor networks," in Proceedings of the 22nd International Conference on Advanced Information Networking and Applications Workshops/Symposia (AINA '08), pp. 634–638, March 2008.
7. Tianjing Wang, Zhen Yang, "A Location-Aware-Based Data Clustering Algorithm in Wireless Sensor Networks", 1-4244-2424-5/08, 2008 IEEE.
8. S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee, "Efficient mining of association rules from wireless sensor networks," in Proceedings of the 11th International Conference on Advanced Communication Technology (ICACT '09), pp. 719–724, February 2009.
9. S. Nithyakalyani, S. Suresh Kumar , "Data Relay Clustering Algorithm for Wireless Sensor Networks: A Data Mining Approach", Journal of Computer Science 8(8): 1281-1284, 2012 ISSN 1549-3636
10. "A New Spectral Classification for Robust Clustering in Wireless Sensor Networks", WMNC'2013, 978-1-4673-5616-9/13, 2013 IEEE.
11. Jaime Lloret, Miguel Garcia, Diana Bri, and Sandra Sendra. A wireless sensor network deployment for rural and forest fire detection and verification. Sensors, 9(11):8722–8747, 2009.
12. Mirjana Maksimovi´c and Vladimir Vujovi´c. Comparative analysis of data mining techniques applied to wireless sensor network data for fire detection. JITA Journal of Information Technology and Application (Banja Luka)-APEIRON, 6(2), 2013.
13. E. M. Mangundu, J. N. Mateus, G. -. L. Zodi and J. Johson, "A wireless sensor network for rainfall monitoring, using cellular network: A case for namibia," 2017 Global Wireless Summit (GWS), 2017, pp. 240-244, doi: 10.1109/GWS.2017.8300469.
14. T R, Prajwala and Ramesh, Dr D and Venugopal, H, Modeling and Forecasting of Rainfall using IoT sensors and Adaptive Boost Classifier for a Region (January 18, 2021). ICICNIS 2020.
15. Kamal Medjaher, Noureddine Zerhouni, Jacques Bahi, Ahmad Farhat, et al.. Random Forests for Industrial Device Functioning Diagnostics Using Wireless Sensor Networks. IEEE AEROSPACE CONFERENCE, 2015., 2015, Big Sky, United States. pp.1-9. ffhal-01304669
16. Hasan, M.A.; Nasser, M.; Ahmad, S.; Molla, M.K. Feature Selection for Intrusion Detection Using Random Forest. J. Inf. Secur. 2016, 7, 129–140. [CrossRef]
17. L. N. de Castro, "Fundamentals of natural computing: an overview," Phys. Life Rev., vol. 4, no. 1, pp. 1–36, 2007.
18. C.-S. Chen, "Artificial Neural Network for Location Estimation in Wireless Communication Systems," Sensors, vol. 12, no. 12, pp. 2798–2817, 2012.
19. Dela Cruz, Roselia. (2018). Artificial Neural Network-based Localization in Wireless Sensor Networks.
20. A. Payal, C. S. Rai and B. V. R. Reddy, "Artificial Neural Networks for developing localization framework in Wireless Sensor Networks," 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014, pp. 1-6, doi: 10.1109/ICDMIC.2014.6954228.
21. Tan, X.; Su, S.; Huang, Z.; Guo, X.; Zuo, Z.; Sun, X.; Li, L. Wireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm. Sensors 2019, 19, 203. https://doi.org/10.3390/s19010203.
22. Intel Lab Data. http://berkeley.intel-research.net/labdata/

AUTHORS PROFILE

Suvarna S Patil, obtained her Bachelor's degree in Electronics and Communication Engineering from Dr P G Halakatti College of Engineering, BLDEA's Bijapur, Karnataka in 2000 and Master's dgree from Ballari Institute of Technology, Ballari, Karnataka in 2006. Joined academics in the year 2001. Currently, she is Assistant Professor in Electronics and Communication Engineering Department at Rao Bahaddur Y Mahabaleswarappa Engineering College, Ballari. Her specializations include Computer Networking, Wireless Sensor Networks and Data Mining.

Dr. B M Vidyavathi, received doctorate in computer science and Engineering from Visvesvaraya University, Belagavi, Karnataka in the year 2010. She was working as a professor in Department of Computer science and engineering, BITM, Ballari, karnataka since 2000 and currently heading the Department of Artificial Intelligence and Machine Learning. Her area of interest includes Data mining, Pattern recognition, Software Engineering, Artificial Intelligence. She is a senior member of Computer society of India, Life member of Indian Society for Technical Education (LMISTE). Member of International Association of Engineers (IAENG), senior member of International Association of Computer Science and Information Technology (IACSIT). She has published 60 papers in national, international conferences and journals.