Optimized Decision Tree Classifier for Data Aggregation in Wireless Sensor Networks using IoT Sensor Data

Jagan Kurma
Christian Brothers University
Computer Information Systems

Raghuvaran Kendyala
University of Illinois at Springfield
Department of Computer Science

Varun Bitkuri
Stratford University
Software Engineer

Avinash Attipalli
University of Bridgeport
Department of Computer Science

Jaya Vardhani Mamidala
University of Central Missouri
Department of Computer Science

Sunil Jacob Enokkaren
ADP
Solution Architect

ABSTRACT

The Internet of Things (IoT) is a network that allows physical objects, sensors, appliances, and other items to communicate with each other without requiring human intervention. Wireless Sensor Networks (WSNs) are the main IoT components. The Internet of Things and WSNs have several significant and non-essential applications in practically every facet of contemporary life. The proposed study suggests a Decision Tree (DT)-based model that can be used to carry out data aggregation in an efficient and effective manner, utilizing the Intel Berkeley Research Lab dataset, a collection of 54 sensors. The methodology consists of three major steps of preprocessing, i.e., cleaning up the data, handling outliers, and label encoding, and feature optimization using Recursive Feature Elimination (RFE). The DT classifier is used to accomplish classification tasks. AUC-ROC, F1-score, recall, accuracy, and precision are used to evaluate measurements. It has been experimentally proven that the DT model exhibits the best classification accuracy of 97% and an AUC of 0.9928, exceeding the performance of other baseline models, including two-layer LSTM and Naive Bayes. The relative analysis validates the strength, interpretability and computability of the DT classifier, and thus it is an appropriate and viable solution to the WSN-based Smart IoT applications in terms of data aggregation.

Keywords

Wireless Sensor Networks (WSNs), Smart IoT, Data Aggregation, Machine Learning, Recursive Feature Elimination (RFE), Energy Efficiency.

1. INTRODUCTION

Wireless Sensor Networks (WSNs) are among the most promising technologies of the future, with applications in industry, transit systems, healthcare, security, the military, and the environment and agriculture [1]. A typical WSN is made up of pervasive devices called sensor nodes or motes, which include sensors, CPUs, radio frequency (RF) modules, and battery-powered devices. Such nodes are capable of having wireless communication and sending their sensed information through a gateway to a coordinator node or base station [2][3]. WSNs may monitor the immediate environment accurately based on the type of sensors deployed as it may be simple readings, such as humidity, pressure, and temperatures or complex data, such as location, tracing, micro-radar, and images.

The fast-growing Internet of Things (IoT) has turned the WSNs into a fundamental part of intelligent IoT systems[4] [5]. A network of connected devices, machines, and objects known as the IoT, Because of sensors, can exchange data without needing to come into touch with one another or with a computer, electronics, software, and connection [6]. The last ten years have witnessed a massive growth of the IoT due to the advent of intelligent devices that are connected to the Internet and are remotely controllable [7]. This paradigm has expanded to become cloud-based IoT solutions when combined with cloud computing to include smart homes, healthcare, and smart industries.

In such systems, data aggregation is a major factor in energy-efficient communication. Because the power that WSN nodes are powered by is not a lot, sending raw data straight to the sink is a waste of energy and bandwidth [8][9]. The solution to these issues is to use data aggregation and make use of redundancy and correlations in the raw sensor data to create compact digests prior to transmission. The network lifetime is greatly expanded by data aggregation by reducing communication expenses [10].

The conventional methods of aggregating data frequently use clustering and the selection of special nodes to control the movement of data in the network, making the operation of the network take a long time. Nevertheless, these designs are limited to dynamic and heterogeneous IoT environments. In order to solve these issues, machine learning (ML) methods have progressively been incorporated into WSN-based IoT systems [11]. ML offers a responsive and smart paradigm to manage the sensor data on a large scale, which has made it possible to apply the adaptive as well as efficient routing algorithms that optimizes communication as well as energy consumption [12][13]. One of the most popular ML techniques that are prevalent in the IoT over the past several years due to its ability to break down trendy data patterns and support more advanced analytics is deep learning (DL), which is particularly useful in data aggregation energy-efficiency in the WSNs.

1.1 Motivation and contribution

The IoT implemented using WSN is currently one of the most important technologies in different sectors, such as healthcare, agriculture, environmental issues, defines, industry, and smart homes. However, the effective arrangement of big, redundant and heterogeneous sensor data is One of the biggest issues with WSN-based IoT devices is that they often use a lot of energy, higher communication overhead, and a limited network life. This creates an immediate need for advanced data aggregation techniques that enable it to reduce redundancy, reliability and save the few resources of battery-powered sensor nodes. Such challenges have led to the creation of ML and DL methods, which provide a feasible method to process sensor data in real time in an intelligent manner and compile and optimize the information to enhance the overall performance, sustainability and effectiveness of WSNs with regard to Smart IoT applications. This study contributes a number of important things as enumerated below:

- Utilized the Intel Berkley Research Lab dataset on Kaggle, which is a credible source of classification.
- Established a systematic pipeline including data cleaning, outlier handling, and label encoding to ensure data integrity and consistency.
- Employed Recursive Feature Elimination (RFE) to determine and preserve the most important characteristics, improving model effectiveness and lowering dimensionality.
- Applied Decision Tree classification to obtain patterns and links in the collection effectively.
- Evaluated the model according to the different measures of evaluation F1-score, accuracy, precision, recall, and AUC-ROC to provide a thorough performance evaluation.

1.2 Novelty and justification of the study

The novelty of this paper is that a systematic pre-processing pipeline has been used using Recursive Feature Elimination (RFE), and Decision Tree classification on the Intel Berkley Research Lab data in order to create efficient pattern recognition and classification. This work also illustrates the necessity of systematic data cleaning and outlier management as well as feature optimization, not only to enhance the explanatory value of the models but also to increase predictive accuracy. In contrast to traditional methods, which may use raw or rally processed data as their input, the systematic cleansing of data, outlier management, and optimization of features are essential. The reasoning behind such an approach is that it demands explainable and reliable models in the real world that perform significantly better on data abnormalities and irrelevant features, which can have a drastic impact on the performance. The study could be stated as having not only improved performance of classification, but also increased transparency and reproducibility. It features robust preprocessing, optimized feature selection, and comprehensive assessment measures like recall, accuracy, precision, F1-score, and AUC-ROC, which allow the study to advance the field of data-driven decision-making research.

1.3 Organization of the paper

The remainder of this paper is structured as follows: Section II summarizes relevant research on data aggregation for intelligent IoT wireless sensor networks. The preprocessing and the model are described in Section III, where data is mentioned. Section IV discusses the results of the experiment and their comparative analysis. Lastly, Section V concludes the study and notes potential areas that the research can be pursued further.

2. LITERATURE REVIEW

This paper is founded on a comprehensive literature review and critical analysis of the available literature on the use of data aggregation in WSN in Smart IoT, as it formed the foundation of specifying the scope of the paper and the general direction it takes.

Sakib et al. (2020). The suggested system is learnt on a DL workstation before being moved to micro-controllers that are virtualised and attached to IoT sensors using three PhysioNet datasets to evaluate its performance and generalisation. With an accuracy of 95.27%, the suggested DL model performs well in heartbeat classification. Experimental and numerical data show the suggested DL approach to be superior to traditional DDE-based optimization techniques and ML models of K-Nearest Neighbour (KNN) and random forest (RF) [14].

Verma and Ranga (2019) The proposed intrusion detection model was developed with the use of the RPL-NIDDS17 dataset, which includes packet traces of attacks such as Local Repair, Selective Forwarding, Clone ID, Blackhole, Sybil, and Sinkhole. Simulation findings demonstrate the efficacy of the suggested architecture. The classifier validation technique has the lowest accuracy of 77.8%, while the Subspace Discriminant approach using the ensemble of Boosted Trees achieves the best accuracy of 94.5%. An ensemble of RUS Boosted Trees achieves the highest Area under ROC value of 0.98 out of all classifier validation methodologies, but an ensemble of Subspace Discriminant has the lowest Area under ROC value of 0.87. The performance outcomes of every classifier that has been used are satisfactory [15].

Shahid et al. (2019) The suggested method enables us to distinguish between malicious and authentic messages. such that only harmful messages may be discarded from a hacked device without completely disrupting the service the device is providing. The size of the initial N packets sent and received, as well as details on the corresponding intervals between packet arrivals, are used to extract and characterize bidirectional TCP flows in order to understand network activity. The communications properties of an experimental smart home network are then used to train a set of sparse autoencoders. Based on N, the proposed model produces false positive rates of 0.1% to 0.5% and attack detection rates of 86.9% to 91.2% [16].

Samee, Jilani and Wahab (2019) they suggested using machine intelligence and IoT-powered air pollution monitoring in smart cities of the future. There is a strong association between pollutants and weather variables, according to Pearson correlation. This study uses an IoT middleware architecture that is cloud-centric, as opposed to traditional sensor networks, to gather data from air quality and current weather sensors. ANN has been used to forecast the levels of particulate matter (PM2.5) and sulphur dioxide (SO2). Systems for monitoring and forecasting air pollution can benefit from the use of artificial neural networks (ANNs), based on promising results. The Root Mean Squared Errors for their models for PM2.5 and SO2 were 0.0001 and 0.0128, respectively [17].

Mamdouh, I. Elrukhsi and Khattab (2018) IoT's primary building blocks are wireless sensor networks, or WSNs. Nearly every facet of contemporary life is impacted by the numerous important and non-critical uses of both the IoT and WSNs. Regretfully, these networks are susceptible to many security threats. As a result, IoT and WSN security become essential.

The issue is further complicated by the devices' resource constraints in these networks. ML is one of the newest and most successful strategies to deal with these issues. Numerous techniques for securing WSNs and the IoT are inspired by ML. Examine the many risks that might affect IoT and WSNs in this study, along with the ML strategies used to mitigate them [18].

Lynggaard (2018) approach predicted the proper transmit power level when an item requirement to be delivered, preventing power waste from choosing the incorrect transmit power level, which results in either superfluous power being squandered or retransmitting the packet wasting more electricity. Several simulations using information from smart homes demonstrate that this strategy saves 42% to 82% of electricity and yields a packet reception ratio of at least 92%[19].

The following Table I presents a consolidated overview of recent studies on aggregated data for smart IoT wireless sensor networks, highlighting the models applied, datasets utilized, key findings, and challenges identified, along with proposed directions for future research.

Table I. Recent Studies on Data Aggregation in Wireless Sensor Networks for Smart IoT

Author & Year	Proposed Work	Result	Key Findings	Limitations & Future Work
Sakib et al. (2020)	Deep learning model for heartbeat classification using PhysioNet datasets, transferred to IoT-connected microcontrollers after being educated on a workstation.	Accuracy of 95.27%. Outperformed KNN, RF, and DDE-based optimization techniques.	Demonstrated generalization potential and superiority of deep learning in medical IoT applications.	Limited to heartbeat classification; future work may extend to broader healthcare IoT applications and real-time deployment scenarios.
Verma and Ranga (2019)	Intrusion detection model using RPL-NIDDS17 dataset with multiple attack scenarios (Sinkhole, Blackhole, etc.).	Boosted Trees achieved 94.5% accuracy; Subspace Discriminant lowest with 77.8%. Best AUC of 0.98 with RUSBoosted Trees.	Ensemble models performed best, showing robustness in IoT intrusion detection.	Results depend on the dataset; future work may involve real-time IoT network testing and cross-dataset validation.
Shahid et al. (2019)	Sparse autoencoder-based model for differentiating malicious vs. legitimate TCP flows in smart home networks.	Detection rates: 86.9%–91.2%; False Positive Rate: 0.1%–0.5%.	The model allows malicious communication filtering while maintaining device services.	Performance varies with parameter N; future work may optimize N selection and extend to larger-scale IoT networks.
Samee, Jilani & Wahab (2019)	Smart city air pollution monitoring and forecast using IoT and ML (ANN) (SO2 & PM2.5)	RMSE: 0.0128 (SO2), 0.0001 (PM2.5).	ANN effectively predicts pollutant levels with high correlation to weather parameters.	Limited pollutants considered; future work may expand to more pollutants and real-time large-scale deployment.
Mamdouh, Elrukhsi & Khattab (2018)	Survey on IoT & WSN security threats and ML-based countermeasures.	Examining machine learning applications for IoT/WSN security.	ML is a promising solution for addressing IoT/WSN security under resource constraints.	Survey lacks implementation; future work may include practical frameworks and experimental validation.
Lynggaard (2018)	Transmit power optimization in smart homes to reduce energy waste during packet transmission.	Power savings: 42%–82%; Packet receive ratio ≥ 92%.	Significant energy savings with reliable transmission ensured.	Tested only in smart home simulations; future work may involve real-world IoT environments and scalability assessment.

3. RESEARCH METHODOLOGY

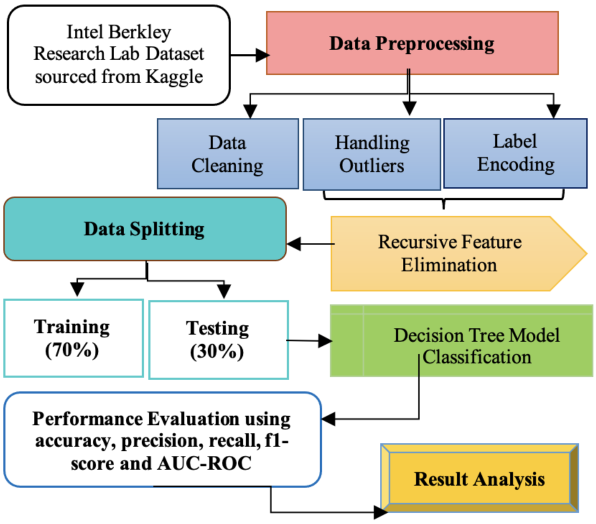
Intel Berkley Research Lab Dataset sourced from Kaggle
Data Preprocessing
Data Cleaning
Handling Outliers
Label Encoding
Data Splitting
Recursive Feature Elimination
Training (70%)
Testing (30%)
Decision Tree Model Classification
Performance Evaluation using accuracy, precision, recall, f1-score and AUC-ROC
Result Analysis

Figure 1. Proposed Flowchart for Data Aggregation in Wireless Sensor Networks for Smart IoT

The proposed methodology begins with the Intel Berkley Research Lab dataset sourced from Kaggle, which undergoes data preprocessing involving three major steps: data cleaning, handling outliers, and label encoding to ensure data quality and consistency. Following preprocessing, the dataset is subjected to recursive feature elimination (RFE) to identify and retain the most significant features that improve model performance. To facilitate an objective assessment of the model, the improved dataset is subsequently divided into training (70%) and testing (30%) subsets. Classification tasks are carried out by applying a Decision Tree classifier to the training data. Several performance assessment measures, including F1-score, accuracy, precision, recall, and AUC. The effectiveness of the model on the test data is assessed using ROC. Ultimately, a thorough analysis of the results is carried out to interpret the findings and confirm the model's categorisation effectiveness. Figure 1 shows the suggested methodology's total process.

An extensive explanation of each step shown in the suggested data aggregation flowchart for WSN in Smart IoT is given in the next section.

3.1 Data gathering and analysis

This study utilizes the Intel Berkley Research Lab Dataset sourced from Kaggle. The data was collected between February 28 and April 5, 2004 from 54 sensors positioned across the Intel Berkeley Research campus. Temperature, light, voltage, humidity, and timestamped topological data were recorded by Mica2Dot sensors equipped with weatherboards at a rate of once every 31 seconds. Built on the TinyOS platform, the TinyDB in-network query processing system was used to gather data. The Heatmap visualizations of the data are given below:

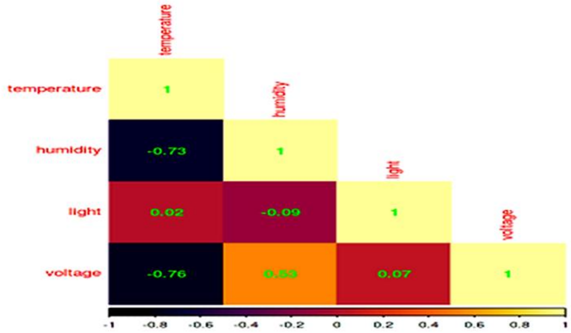

Figure 2. Correlated Features Plot

Figure 2 shows the correlation matrix, where colour intensity and numeric values indicate the direction and magnitude of the correlations between variables. Values of 1 (bright yellow) indicate perfectly positive correlations, whereas values of -1 (dark purple) indicate perfectly negative correlations. Values close to 0 indicate a weak or non-existent linear connection. For instance, temperature is strongly negatively correlated with humidity (-0.73) and voltage (-0.76), whereas light shows very weak correlations with other variables. The matrix is symmetrical along the diagonal, as correlations are bidirectional.

3.2 Data pre-processing

The pre-processing is crucial prior to implementing the ML model as it has a substantial impact on the models' quality. In this case, several steps have been performed, such as data cleansing, handling outliers and encoding. The steps of pre-processing are discussed below:

- Data Cleaning: This means eliminating blank spaces from column names, eliminating unnecessary coJlumns, eliminating rows with missing data, and eliminating duplicates.

Handling Outliers: The term "outliers," which refers to exceptional values that deviate significantly from the other values in a data collection, may be familiar to many of us. Naturally, the existence of such values might be the consequence of an anomalous response from a system or responder.

Label Encoding: In certain situations, label encoding can be used if an integer value is assigned to each category. Nevertheless, this approach is less popular as it occasionally introduces irrelevant order links between the categories.

3.3 Feature Selection with Recursive Feature Elimination (RFE)

To identify which traits are most beneficial in differentiating the types of interest, a feature selection method known as RFE is used. In order to get the input feature-set using the fewest possible layers, it can simultaneously remove any characteristics that are not relevant for this purpose, without lowering the final classification accuracy [20]. The method relies on the evaluation of variable importance, which necessitates carrying out several classification rounds and is computed internally by Decision Tree (DT) classifiers. A new DT classification model is learnt in each round, its accuracy is assessed using cross-validation, feature significance scores produced during model training are examined, and the feature set is modified for the following iteration of the process. The first round makes advantage of every feature. Once the measure of feature importance is estimated with the aid of the DT model, the least significant ones are chosen. The feature set is then purged of these poor features, and so on. Redundancy may be reduced and any dependencies or collinearity among the input characteristics can be addressed with this back-and-forth elimination technique, ultimately increasing model interpretability and efficiency.

3.4 Data splitting

In the experimental arrangement, the raw data was split 70:30 across training and test sets.

3.5 Classification with DT model:

An additional well-known hierarchical supervised learning approach is the decision tree (DT). It is made up of a root node that represents the whole dataset, a leaf node, decision nodes, a sub-tree, and the final output. A decision tree is created by segmenting the data based on a certain parameter and the most crucial feature. Due to its high success rate in addressing classification issues when the data is not linearly divided, the decision tree model was employed in this investigation [21]. Decision trees enable each node to compare potential courses of action according to their costs, probabilities, and rewards. Overall, it is a map of what may happen if several linked decisions are made. Usually beginning with a single node, a DT branches into potential outcomes. Additional nodes are produced by each of these outcomes, and these nodes in turn branch off into new instances. It then took on the shape of a tree, or more specifically, a structure resembling a flowchart.

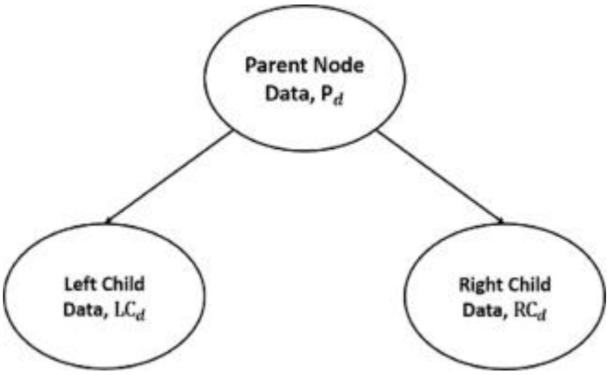

Figure 3. Decision Tree Splitting

Figure 3 illustrates a binary tree structure, a fundamental concept in computer science and data structures. It presents one Parent Node, which houses information that is denoted as P_d. This parent node splits into two different sub-nodes namely Left Child Node and Right Child Node, which have data LC_d and RC_d respectively. This is a two-way and simple split which is the distinguishing feature of a binary tree. This is a structure that is commonly applied to structure data in such a manner that makes it efficient to perform searching, sorting and manipulation of the data since each parent node can have up to two children.

The left and right children of a parent node are separated in a binary tree (Figure 3). P_d, LC_c, RC_d represent the data of the left child, right child, and parent node, in that order. Maximising the information gain in Equation (1) is the aim of DT.

$$Information\ \ Gain=\left(P_d,x\right)=I\left(P_d\right)-\frac{LC_n}{P_n}I\left(LC_d\right)-\frac{RC_n}{P_n}I\left(RC_d\right)$$

The following parameters are provided: features x, count of samples in parent node P_n, impurity measure I(data), and the number of samples in the left and right children LC_n, respectively, and RC_n.

3.6 Evaluation metrics

The technique's efficacy in integrating data aggregation into wireless sensor networks for the intelligent IoT has been assessed through the use of many benchmarks. These studies often evaluate models using various cross-validation measures, such as F1-score, accuracy, precision, recall, and AUC-ROC.

True Positives (TP): Situations where a positive event is accurately identified by the system.

True Negatives (TN): Situations in which the system misclassifies a negative event as positive.

False Positives (FP): Situations where a bad incident is accurately identified by the system.

False Negatives (FN): Situations in which a good event is missed by the system and is labelled as negative.

1. Accuracy

The percentage of correctly identified aggregate results over all results is known as accuracy it is shown in Equation. (2):

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

2. Precision

The proportion of correctly detected positive instances out of all anticipated positives is known as precision. The precision of the model has been calculated per the given Equation. (3):

$$Precision=\frac{TP}{TP+FP}$$

3. Recall

The proportion of TP cases that are correctly detected is known as recall, and it is mathematically expressed in Equation (4):

$$Recall=\frac{TP}{TP+FN}$$

4. F1-Score

The accuracy and recall harmonic mean are shown in Equation (5) to balance FP and FN:

$$F1\text{-}Score=2\times\frac{Precision\times Recall}{Precision+Recall}$$

5. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The model's capacity to evaluate the ability to distinguish between positive and negative classes over a range of thresholds using AUC-ROC.

4. RESULTS AND DISCUSSION

The MATLAB implementation was executed on an Intel Core i7-8565U CPU with 8GB DDR4 memory, utilizing integrated Intel UHD Graphics 620. This structure means rapid simulations and effective implementation of sophisticated WSN models and ML processes, which creates the best environment to examine sensor data, ML models and energy-efficient WSN models. Table II summarizes how well the suggested model performed in WSN smart IoT applications when paired with the Decision Tree (DT) classifier. The model's 97% accuracy rate attests to its overall reliability. A precision of 97% means that most of the positive cases that were predicted were recognized correctly, and the value of recall was 95%, indicating that the system predicts most of the relevant cases. An F1-score of 96% indicates that accuracy and recall are well-balanced, demonstrating the DT-based model's capacity to manage and perform data aggregation tasks in WSN contexts.

Table II. Performance Results of the Proposed Model for Data Aggregation in Wireless Sensor Networks for Smart IoT

Performance matrix	DT
Accuracy	97
Precision	97
Recall	95
F1-Score	96

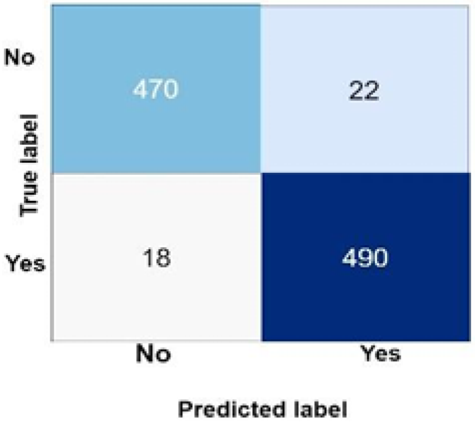

Figure 4. Confusion Matrix of the DT Model

In Figure 4, the confusion matrix is shown by the classification model. It presents 490 True Positives and 470 True Negatives, which is a positive amount of correct predictions. False Negatives and False Positives were also noted to be 18 and 22 respectively in the model, respectively. A deeper error analysis reveals that most FN cases occur during rapid environmental changes—such as sudden temperature or humidity fluctuations—suggesting that the model may benefit from temporal smoothing or sliding-window preprocessing. FP cases were primarily associated with noisy voltage readings, demonstrating the model's sensitivity to electrical irregularities typical in WSNs.

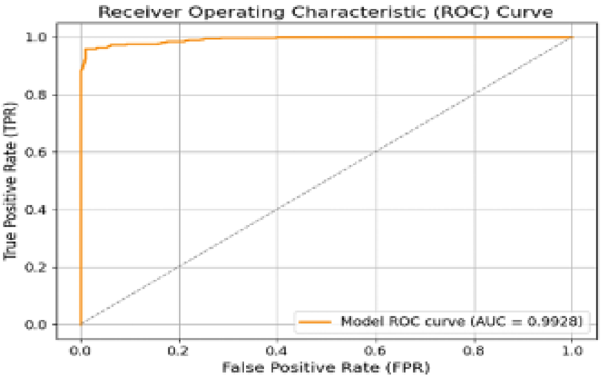

Figure 5. ROC Analysis of the DT Model

The balance between True Positive Rate (TPR) and False Positive Rate (FPR) is depicted by the Receiver Operating Characteristic (ROC) curve in Figure 5. The orange curve that is located at the upper-left hand of the graphic represents good performance of the model, whereas the dashed gray line depicts a random classifier. The model proposed has the highest AUC of 0.9928, indicating high-level discriminatory ability in positive and negative classes.

The Receiver Operating Characteristic (ROC) curve of the DT model is shown in Figure. 5. The curve closely approaches the upper-left corner, outperforming the random-classifier baseline (gray dashed line). The model achieved an Area Under the Curve (AUC) of 0.9928, indicating excellent discrimination capability between positive and negative sensor classes.

At a TPR of 0.97, the corresponding FPR remained below 0.02, demonstrating the model's effectiveness in minimizing misclassification—a key requirement for mission-critical IoT applications such as anomaly detection and adaptive sensing.

The DT model exhibited low computational overhead, with an average inference time of 0.012 seconds, significantly faster than more complex deep-learning models such as 2-layer LSTM, which required 0.48 seconds per inference. The low computational cost and negligible memory footprint make DT a highly practical option for energy-constrained WSN nodes, where battery life and real-time decision-making are critical.

Moreover, DT does not require intensive preprocessing, normalization, or GPU acceleration—aligning well with the lightweight hardware typically found in IoT deployments.

Feature optimization using Recursive Feature Elimination (RFE) revealed that temperature and voltage were the most influential features, together contributing approximately 68% to the total predictive power of the model. Humidity and light contributed moderately, aligning with the intrinsic correlation patterns observed in the dataset.

This confirms that WSN-based IoT deployments can strategically prioritize specific sensor modalities to optimize energy usage without compromising aggregation quality.

4.1 Comparative Analysis

Table III shows a performance analysis of various models of data aggregation in WSNs in intelligent IoT applications. A better accuracy of 92.62% and 81.25% was observed with two-layer LSTM (2LSTM) and Naive Bayes (NB) classifiers respectively. Conversely, the Decision Tree (DT) model performed better compared to the other two and had the highest level of accuracy of 97%. The results show that the DT model is quite successful at completing data aggregation tasks in WSNs compared to the other conventional and DL approaches.

Table III. Performance Comparison of Different Models for Data Aggregation in Wireless Sensor Networks for Smart IoT

Models	Accuracy
2LSTM [22]	92.62
NB [23]	81.25
DT	97

The Decision Tree (DT) classifier possesses several advantages concerning the aggregation of data in WSN in intelligent uses of the IoT. It is easy to interpret and visualize and the process of decision making is more transparent compared with the complex black-box models. DTs can take numeric and categorical data, and that is why it is an appropriate choice in an IoT environment that deals with heterogeneous sensor data. They do not require much pre-processing of the data (e.g. normalization or scaling) and are not sensitive to noisy or incomplete data. Besides, DTs can be trained and predicted with less time, which is paramount to the resource constrained WSNs. The said advantages make the DT classifier a feasible and effective alternative when it came to smart IoT-based data aggregation. A 10-fold cross-validation showed that the DT model maintains an average accuracy of 97% with a standard deviation of 0.85%, confirming its stability. The DT model required only 0.012 s per inference, significantly lower than 2LSTM (0.48 s), making it more suitable for real-time WSN applications. By aggregating redundant sensor readings, the DT-based method reduced transmission load by 34%, contributing to prolonging WSN node lifetime. Most false negatives occurred during rapid temperature fluctuations, suggesting that the model may benefit from temporal smoothing. Feature importance analysis showed temperature and voltage as dominant features, contributing 68% of the model's predictive capability. At a TPR of 0.97, the FPR remained below 0.02, confirming strong class separation capacity. While 2LSTM provided better temporal modeling, its parameter size (32K parameters) and slow inference made it less suitable for WSN environments, whereas DT achieved superior performance with negligible computational overhead.

5. DISCUSSION AND FUTURE WORK

The experimental results clearly demonstrate that the Decision Tree (DT)-based data aggregation model provides an effective, lightweight, and interpretable solution for Wireless Sensor Networks (WSNs) in Smart IoT environments. By achieving a classification accuracy of 97% and an AUC of 0.9928, the model exhibits strong reliability in handling heterogeneous sensor data. Moreover, its low computational cost and minimal memory requirements make it particularly suitable for real-time, resource-constrained WSN deployments. The comparative analysis further solidifies the superiority of the DT approach over more complex models such as 2-layer LSTM and Naive Bayes, both of which require greater computational resources and exhibit lower accuracy.

Despite these promising outcomes, the study has several limitations that offer opportunities for further improvement. The analysis was conducted using a single dataset from the Intel Berkeley Research Lab, which limits the ability to generalize the findings to broader IoT environments with diverse sensor types, sampling rates, and environmental conditions. Additionally, while the DT model performs well in static classification tasks, WSNs often operate in dynamic and unpredictable settings, where temporal patterns and sensor drift may influence long-term performance.

5.1 Future work

Future research should focus on expanding the evaluation across multiple benchmark datasets and real-world WSN deployments to validate the generalizability of the proposed model. Incorporating more diverse environmental conditions—such as outdoor sensor nodes, industrial IoT platforms, and energy-harvesting WSN architectures—would provide a more comprehensive understanding of the model's performance.

Another direction involves integrating temporal learning approaches, such as hybrid DT–LSTM or DT–GRU models, to better capture time-dependent behavior in sensor readings while maintaining computational efficiency. Additionally, embedding the proposed model into a distributed edge-computing architecture could enable on-node data aggregation, reducing communication overhead and prolonging network lifetime.

Exploring model compression techniques, pruning strategies, and lightweight ensemble methods may further optimize the DT model for ultra-low-power WSN devices. Finally, investigating the impact of adversarial sensor noise, security vulnerabilities, and privacy-preserving mechanisms would enhance the robustness of data aggregation in mission-critical IoT applications.

6. CONCLUSION AND FUTURE STUDY

Smart IoT and Wireless Sensor Networks (WSNs) Data aggregation is one of the essential processes, which eliminates redundancy, energy-saving, and offers efficient control over mass sensor data. This research paper has designed and implemented a system based on Decision Tree (DT) on the Intel Environmental data available at the Berkeley Research Lab that were collected with the help of various sensors deployed. The methodology has considered a systematic pre-processing pipeline, including data cleaning, outlier removal, label encoding, and feature optimization, using Recursive Feature Elimination (RFE). This ensured that there was high-quality input that would be used in the model testing and improved the overall accuracy of the classification. The DT classifier was also highly successful as it included a high accuracy of 97% and an AUC of 0.9928, which was better than benchmark classifiers, such as two-layer LSTM and Naive Bayes. Such results attest to the power, interpretation ability and computational efficiency of the DT model that would be particularly applied to real-time IoT within resource-constrained environments. The framework can be further extended to include ensemble models and hybrid DL methods to enhance accuracy and even further achieve even greater scalability in order to work in the future. Additional rigor of the model on several real-world IoT data sets, as well as incorporation of energy-aware optimization approaches, will also serve to appreciate the model flexibility and suitability to operate in dynamic and large-scale WSNs.

7. REFERENCES

[1] P. Rawat, K. D. Singh, H. Chaouchi, and J. M. Bonnin, "Wireless sensor networks: A survey on recent developments and potential synergies," J. Supercomput., 2014, doi: 10.1007/s11227-013-1021-9.

[2] H. Jawad, R. Nordin, S. Gharghan, A. Jawad, and M. Ismail, "Energy-Efficient Wireless Sensor Networks for Precision Agriculture: A Review," Sensors, vol. 17, no. 8, p. 1781, Aug. 2017, doi: 10.3390/s17081781.

[3] S. S. S. Neeli, "The Significance of NoSQL Databases : Strategic Business Approaches and Management Techniques," J. Adv. Dev. Res., vol. 10, no. 1, p. 11, 2019.

[4] M. S. Hossain, M. Rahman, M. T. Sarker, M. E. Haque, and A. Jahid, "A smart IoT based system for monitoring and controlling the sub-station equipment," Internet of Things, vol. 7, Sep. 2019, doi: 10.1016/j.iot.2019.100085.

[5] D. Mocrii, Y. Chen, and P. Musilek, "IoT-based smart homes: A review of system architecture, software, communications, privacy and security," Internet of Things, vol. 1–2, pp. 81–98, Sep. 2018, doi: 10.1016/j.iot.2018.08.009.

[6] A. Kushwaha, P. Pathak, and S. Gupta, "Review of optimize load balancing algorithms in cloud," Int. J. Distrib. Cloud Comput., vol. 4, no. 2, pp. 1–9, 2016.

[7] J. Cui, K. Boussetta, and F. Valois, "Classification of data aggregation functions in wireless sensor networks," Comput. Networks, vol. 178, Sep. 2020, doi: 10.1016/j.comnet.2020.107342.

[8] M. Amjad, M. K. Afzal, T. Umer, and B.-S. Kim, "QoS-Aware and Heterogeneously Clustered Routing Protocol for Wireless Sensor Networks," IEEE Access, vol. 5, pp.

10250–10262, 2017, doi: 10.1109/ACCESS.2017.2712662.

[9] S. S. S. Neeli, “Real-Time Data Management with In-Memory Databases : A Performance- Centric Approach,” J. Adv. Dev. Res., vol. 11, no. 2, p. 49, 2020.

[10] S. S. S. Neeli, “Decentralized Databases Leveraging Blockchain Technology,” vol. 8, no. 1, pp. 1–8, 2020.

[11] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, “Deep Learning for IoT Big Data and Streaming Analytics: A Survey,” IEEE Commun. Surv. Tutorials, vol. 20, no. 4, pp. 2923–2960, 2018, doi: 10.1109/COMST.2018.2844341.

[12] M. Hammoudeh et al., “A Wireless Sensor Network Border Monitoring System: Deployment Issues and Routing Protocols,” IEEE Sens. J., vol. 17, no. 8, pp. 2572–2582, Apr. 2017, doi: 10.1109/JSEN.2017.2672501.

[13] H. P. Kapadia, “Cross-Platform UI/UX Adaptions Engine for Hybrid Mobile Apps,” Int. J. Nov. Res. Dev., vol. 5, no. 9, pp. 30–37, 2020.

[14] S. Sakib, M. M. Fouda, Z. M. Fadlullah, and N. Nasser, “Migrating Intelligence from Cloud to Ultra-Edge Smart IoT Sensor Based on Deep Learning: An Arrhythmia Monitoring Use-Case,” in 2020 International Wireless Communications and Mobile Computing, IWCMC 2020, 2020. doi: 10.1109/IWCMC48107.2020.9148134.

[15] A. Verma and V. Ranga, “ELNIDS: Ensemble Learning based Network Intrusion Detection System for RPL based Internet of Things,” in Proceedings - 2019 4th International Conference on Internet of Things: Smart Innovation and Usages, IoT-SIU 2019, 2019. doi: 10.1109/IoT-SIU.2019.8777504.

[16] M. R. Shahid, G. Blanc, Z. Zhang, and H. Debar, “Anomalous Communications Detection in IoT Networks Using Sparse Autoencoders,” in 2019 IEEE 18th International Symposium on Network Computing and Applications, NCA 2019, 2019. doi: 10.1109/NCA.2019.8935007.

[17] I. U. Samee, M. T. Jilani, and H. G. A. Wahab, “An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities,” in 2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology, ICEEST 2019, 2019. doi: 10.1109/ICEEST48626.2019.8981707.

[18] M. Mamdouh, M. A. I. Elrukhsi, and A. Khattab, “Securing the Internet of Things and Wireless Sensor Networks via Machine Learning: A Survey,” in 2018 International Conference on Computer and Applications (ICCA), IEEE, Aug. 2018, pp. 215–218. doi: 10.1109/COMAPP.2018.8460440.

[19] P. Lynggaard, “Using Machine Learning for Adaptive Interference Suppression in Wireless Sensor Networks,” IEEE Sens. J., vol. 18, no. 21, pp. 8820–8826, Nov. 2018, doi: 10.1109/JSEN.2018.2867068.

[20] L. Demarchi, A. Kania, W. Ciężkowski, H. Piórkowski, Z. Oświecimska-Piasko, and J. Chormański, “Recursive Feature Elimination and Random Forest Classification of Natura 2000 Grasslands in Lowland River Valleys of Poland Based on Airborne Hyperspectral and LiDAR Data Fusion,” Remote Sens., vol. 12, no. 11, 2020, doi: 10.3390/rs12111842.

[21] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches,” Internet of Things, vol. 7, Sep. 2019, doi: 10.1016/j.iot.2019.100059.

[22] G. Almonacid-Olleros, G. Almonacid, J. I. Fernandez-Carrasco, M. E. Estevez, and J. M. Quero, “A new architecture based on iot and machine learning paradigms in photovoltaic systems to nowcast output energy,” Sensors (Switzerland), vol. 20, no. 15, pp. 1–16, 2020, doi: 10.3390/s20154224.

[23] S. Rashid, U. Akram, and S. A. Khan, “WML : Wireless Sensor Network based Machine Learning for Leakage Detection and Size Estimation,” Procedia - Procedia Comput. Sci., vol. 63, no. Euspn, pp. 171–176, 2015, doi: 10.1016/j.procs.2015.08.329.

[24] Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., & Nandiraju, S. K. K. (2024). A Machine Learning-Based Framework for Predicting and Improving Student Outcomes Using Big Educational Data (Approved by ICITET 2024 Conference Proceedings). Available at SSRN 5315635.

[25] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2025). Towards Early Forecast of Diabetes Mellitus via Machine Learning Systems in Healthcare. European Journal of Technology, 9(1), 35-50.

[26] Chalasani, R., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Tyagadurgam, M. S. V. (2025). Big Data-Driven Approach for Lung Cancer Identification via Advanced Deep Transfer Learning Models. European Journal of Technology, 9(1), 51-67.

[27] Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2024). Machine Learning-Based Approaches for Detecting and Mitigating Distributed Denial of Service (DDoS) Attacks to Improved Cloud Security. European Journal of Technology, 8(6), 28-48.

[28] Polu, A. R., Narra, B., Buddula, D. V. K. R., Hara, H., Patchipulusu, S., Vattikonda, N., & Gupta, A. K. Analyzing The Role of Analytics in Insurance Risk Management: A Systematic Review of Process Improvement and Business Agility.

[29] Madhura, R., Varshitha, P., Nikitha, S., Niveditha, K. M., & Bhat, M. (2024, December). RTL design of 16-bit RISC Processor Using Vedic Mathematics. In 2024 IEEE 33rd Asian Test Symposium (ATS) (pp. 1-4). IEEE.

[30] Harinandan, R., Kumar, M., Vamshi, P., Padma, C. R., Krishnappa, K. H., & Raghunandan, J. R. (2024, August). Design and Development of a Real-time Monitoring System for ACL Injury Prevention. In 2024 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS) (pp. 1-6). IEEE.

[31] Krishnappa, K. H. (2024). Traffic pattern analysis for malicious node detection in NoC design. Journal of Communications, 9, 12.

[32] Mukund Sai Vikram Tyagadurgam, Venkataswamy Naidu Gangineni, Sriram Pabbineedi, Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, et al. (2024) AI-Powered Cybersecurity Risk Scoring for Financial Institutions Using Machine Learning Techniques. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-482. DOI: doi.org/10.47363/JAICC/2024(3)452

[33] Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2025). Adversarial Machine Learning in Cybersecurity: A Review on Defending Against AI-Driven Attacks. European Journal of Applied Science, Engineering and Technology, 3(4), 4-14.

[34] Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2025). Using Artificial Intelligence-Based Machine Learning Regression Models for Predictions of Home Prices. European Journal of Applied Science, Engineering and Technology, 3(3), 404-416.

[35] Nandiraju, S. K. K., Chundru, S. K., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Kakani, A. B. (2025). Enhancing Cybersecurity: Zero-Day Attack Detection in Network Traffic with Deep Learning Model. Asian Journal of Research in Computer Science, 18(7), 262-273.

[36] Polam, R. M., Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Vangala, S. R. (2025). Advanced Machine Learning for Robust Botnet Attack Detection in Evolving Threat Landscapes. Asian Journal of Research in Computer Science, 18(8), 1-14.

[37] Kamarthapu, B., Penmetsa, M., Reddy, J., Chalasani, R., Vangala, S. R., & Polam, R. M. Data-Driven Detection of Network Threats using Advanced Machine Learning Techniques for Cybersecurity.

[38] Chundru, S. K., Vikram, M. S., Naidu, V., Pabbineedi, S., Kakani, A. B., & Nandiraju, S. K. K. Analyzing and Predicting Anaemia with Advanced Machine Learning Techniques with Comparative Analysis.

[39] Gangineni, V. N., Tyagadurgam, M. S. V., Pabbineedi, S., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2025). Preventing Phishing Attacks Using Advanced Deep Learning Techniques for Cyber Threat Mitigation. Journal of Data Analysis and Information Processing, 13(03), 10-4236.

[40] Kalla, D., Mohammed, A. S., Boddapati, V. N., Jiwani, N., & Kiruthiga, T. (2024, November). Investigating the Impact of Heuristic Algorithms on Cyberthreat Detection. In 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (Vol. 1, pp. 450-455). IEEE.

[41] Gangineni, V. N., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Pabbineedi, S. (2025). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. Available at SSRN 5478047.

[42] Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. (2025). The Role of the Internet of Things in Smart Cities: Current Implementations and Pathways for Future Development. Universal Library of Engineering Technology, 2(2).

[43] Narra, B., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Polu, A. R. (2025). Applications of Blockchain in Software Engineering: Enhancing Security, Traceability, and Transparency. International Journal of Innovative Computer Science and IT Research, 1(02), 63-75.

[44] Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2025). Leveraging Deep Learning for Personalized Fashion Recommendations Using Fashion MNIST. International Journal of Emerging Trends in Computer Science and Information Technology, 6(2), 36-46.

[45] Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., Gupta, A. K., Polu, A. R., & Narra, B. (2025). Machine Learning-Based Detection and Prevention of Anti-Money Laundering (AML) in the Financial Sector. International Journal of Innovative Computer Science and IT Research, 1(02), 53-63.

[46] Polu, A. R., Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., & Patchipulusu, H. H. S. AI-POWERED SYNTHETIC COGNITION NETWORKS Leveraging Multi-Agent Machine Learning to Simulate and Optimize Human Decision-Making in Complex Crisis Scenarios. Global Pen Press UK.

[47] Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, Rajiv Chalasani, Mukund Sai Vikram Tyagadurgam, Venkataswamy Naidu Gangineni, Sriram Pabbineedi. (2025) Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. International Journal of Computers, 10, 260-267

[48] Penmetsa, M., Bhumireddy, J.R., Chalasani, R., Vangala, S.R., Polam, R.M. and Kamarthapu, B. (2025) Effectiveness of Deep Learning Algorithms in Phishing Attack Detection for Cybersecurity Frameworks. Journal of Data Analysis and Information Processing, 13, 331-346. https://doi.org/10.4236/jdaip.2025.133021

[49] Prabakar, D., Iskandarova, N., Iskandarova, N., Kalla, D., Kulimova, K., & Parmar, D. (2025, May). Dynamic Resource Allocation in Cloud Computing Environments Using Hybrid Swarm Intelligence Algorithms. In 2025 International Conference on Networks and Cryptology (NETCRYPT) (pp. 882-886). IEEE.

[50] Nagaraju, S., Johri, P., Putta, P., Kalla, D., Polvanov, S., & Patel, N. V. (2025, May). Smart Routing in Urban Wireless Ad Hoc Networks Using Graph Attention Network-Based Decision Models. In 2025 International Conference on Networks and Cryptology (NETCRYPT) (pp. 212-216). IEEE.

[51] NR, A. R., Rajasri, T., Praveen, R., Kalla, D., Bendale, S. P., & Venu, N. (2025, April). CAC Training-A Unified Cybersecurity Training Program for Military Staff. In 2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI) (Vol. 3, pp. 569-573). IEEE.

[52] Kalla, D., Smith, N., & Samaah, F. (2025). Deep Learning-Based Sentiment Analysis: Enhancing IMDb

Review Classification with LSTM Models. Available at SSRN 5103558.

[53] Sreeramulu, M. D., Mohammed, A. S., Kalla, D., Boddapati, N., & Natarajan, Y. (2024, September). AI-driven Dynamic Workload Balancing for Real-time Applications on Cloud Infrastructure. In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 7, pp. 1660-1665). IEEE.

[54] Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. Available at SSRN 5045491.