iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70728

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Malicious Node Detection and Secure Data Storage in Wireless Sensor Networks Using Blockchain and Machine Learning

Sharmila Babu C S[1], Vyshali C[2], S. Saila[3]
Department of Computer Science &Engineering, Jyothy Institute Of Technology, Bengaluru, Karnataka, India

Abstract: ***Wireless Sensor Networks (WSNs) are widely used for monitoring and data collection in various environments. However, these networks are vulnerable to attacks from malicious nodes, which can compromise the integrity and reliability of the system. In this work, I propose a model that combines blockchain-based registration and authentication with machine learning for real-time malicious node detection. The system uses the Histogram Gradient Boost (HGB) classifier to identify threats and stores legitimate data in the Interplanetary File System (IPFS), with hashes recorded on the blockchain. To keep the system efficient, I use the Verifiable Byzantine Fault Tolerance (VBFT) consensus instead of Proof of Work (PoW). My results, based on the WSN-DS dataset, show that this approach not only improves detection accuracy but also reduces transaction costs compared to traditional methods.***

Keywords: Blockchain, histogram gradient boost, IPFS, malicious node detection, VBFT, WSN.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are made up of many small sensor nodes that collect and transmit data for applications like environmental monitoring, supply chain management, and military surveillance. These nodes are often deployed in open or hostile environments, making them easy targets for attackers. Malicious nodes can join the network, send false information, or disrupt communication, leading to serious security issues. Traditional solutions often rely on centralized authorities for authentication, which creates a single point of failure and is not ideal for resource-constrained WSNs. To address these challenges, I propose a decentralized system that leverages blockchain for secure node registration and data storage, and machine learning for detecting abnormal or malicious behavior. In this model, only nodes that are authenticated through the blockchain can participate in the network, and all data is securely stored and verified using IPFS and blockchain hashes. The machine learning component, specifically the HGB classifier, is trained to distinguish between normal and malicious nodes, allowing for real-time detection and response.

A. Motivation And Contributions

The main motivation for this work is the vulnerability of WSNs to internal attacks due to their open nature and limited resources. Random deployment of nodes often leads to security risks, and traditional centralized authentication can be easily compromised. Storing large volumes of data directly on the blockchain is costly, and consensus mechanisms like PoW are too resource-intensive for WSNs.

B. Key Contributions Of This Work

1) A blockchain-based decentralized authentication mechanism to prevent unauthorized access.
2) Integration of IPFS for efficient and secure data storage, with hashes recorded on the blockchain.
3) Use of VBFT consensus to reduce transaction costs and improve throughput.
4) Application of the HGB classifier for accurate detection of malicious nodes, outperforming other common classifiers.

II. RELATED WORK

Many researchers have proposed different solutions to address security in WSNs. Lightweight blockchain authentication schemes20 have been developed to ensure integrity and non-repudiation. Hybrid blockchain models21 and reinforcement learning-based routing22 have also been explored. Some works focus on key management, trust models, and privacy-preserving frameworks. However, most existing methods either rely on centralized entities or are too computationally heavy for WSNs.

©IJRASET: All Rights are Reserved

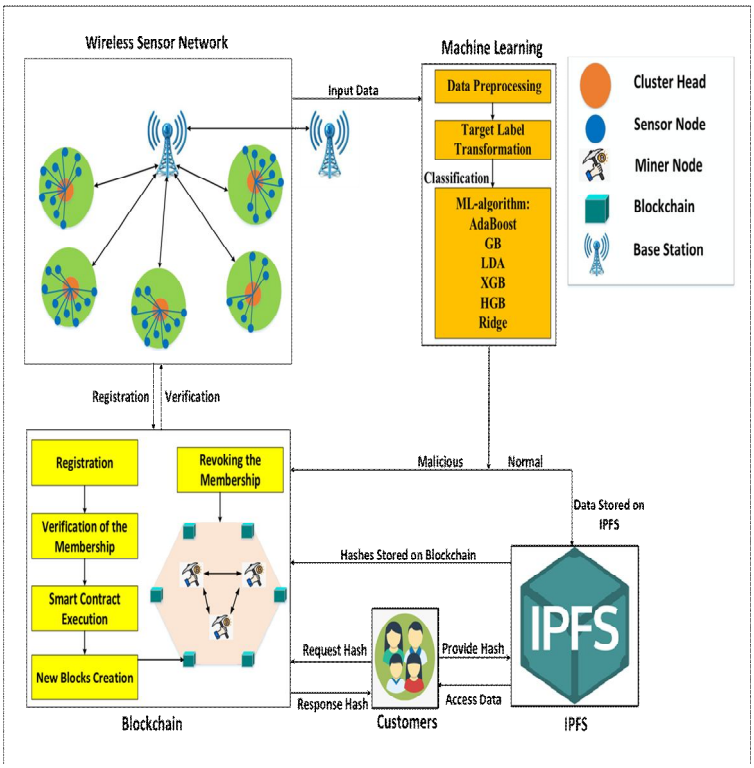


Figure 1Proposed network model for WSNs.

III. PROPOSED SYSTEM MODEL

In this section, I describe the network model and the assumptions made for the proposed system.

A. Assumptions

1) Sensor Nodes (SNs) and Cluster Heads (CHs) are resource-constrained and have unique identities.
2) Base Stations (BSs) have higher computational and storage capacity.
3) No energy holes are assumed in the network.
4) Recommended font sizes are shown in Table 1.

Limitations Already Addressed	Solutions Already Proposed	Validations Already Done	Limitations to be Addressed
Malicious nodes are present in the network [20]	Intrusion prevention framework is proposed	Data integrity and network connectivity	XoR function is not strong enough, blackhole and greyhole attacks may occur
Localization problem of unknown nodes [21]	Trust model based on blockchain is used	Feasibility, fairness and traceability	RSA slows down the encryption method in case of large data
Crowd sensing networks are vulnerable [22]	Confusion mechanism and blockchain based incentive mechanism are proposed	Energy consumption, delay	Not improved in route acquisition latency and packet delivery ratio
Centralization method is used for registration [23]	A hybrid blockchain based model is proposed	Processing time and transmission delay	Data duplication
Localization of WSN nodes, unknown nodes perform attacks on network [24]	A blockchain trust model is proposed	False negative rate, detection accuracy and energy consumption	No encryption and hashing algorithm are used for security
Dynamically routing and centralized registration [25]	Blockchain and reinforcement learning algorithm are used	Delay, energy consumption	Queue delay and processing delay
IoT node manufacturers are unable to agree on a simple central administrator [26]	BCR protocol is proposed	Packet drop ratio, packet delivery ratio, delay	Not improved in route acquisition latency and packet delivery ratio
Balance energy consumption of sensor nodes and to improve WSN longevity [27]	Dynamic hierarchical protocol based on combinatorial optimization is proposed	A hierarchy-based connection mechanism to construct a hierarchical network structure	Processing delay and computational complexity
Reduced lifetime of ultra-dense WSNs [28]	Unsupervised learning approach	Residual energy and computational complexity	Increased computational complexity

Table 1 Summarized related work table.

B. Problem Formulation

WSNs are susceptible to malicious nodes due to random deployment and open access. Centralized authentication is not reliable, and storing all data on the blockchain is expensive and inefficient.

C. System Model Description

The network is organized into SNs, CHs, and BSs. SNs collect data and send it to CHs, which then forward the data to BSs. Both SNs and CHs are registered and authenticated on the blockchain, which is managed by the BSs. The BSs use the HGB classifier to detect malicious nodes. If a node is found to be malicious, its registration is revoked. Otherwise, its data is stored in IPFS, and the hash is recorded on the blockchain.

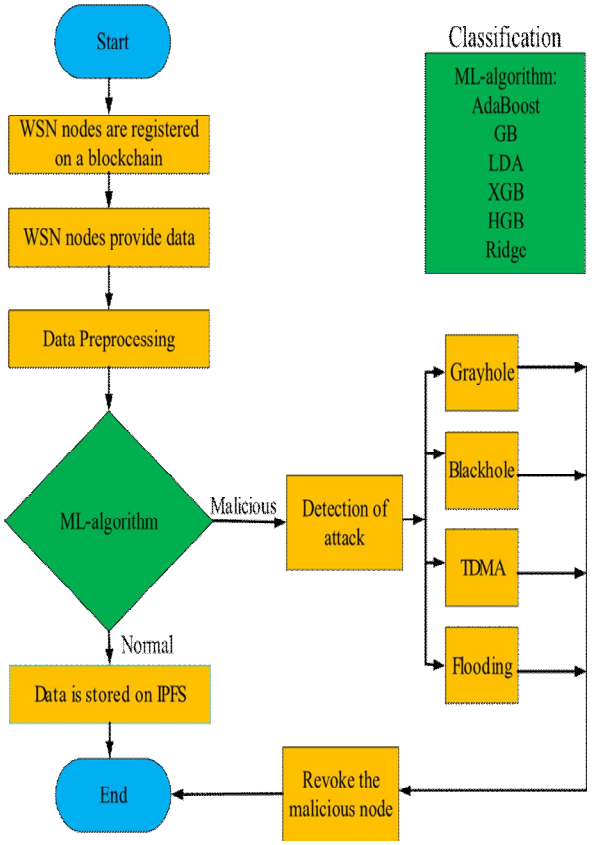


Figure 2 Proposed model's workflow.

D. Registration

SNs and CHs must register on the blockchain before they can participate in the network. Registration requests are sent to the BSs, which handle authentication.

E. Sensor Nodes

SNs are deployed randomly and are responsible for data collection. Their credentials are stored on the blockchain to ensure authenticity.

F. Cluster Heads

CHs act as intermediaries, aggregating data from SNs and forwarding it to the BSs. They have more resources than SNs but less than BSs.

G. Base Station

BSs are the most powerful nodes in the network. They process data, run the HGB classifier, and manage the blockchain and IPFS integration.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

H. Customers

Customers (end users) must also register on the blockchain to access data. They request data hashes from the blockchain and retrieve data from IPFS.

I. Malicious Node Detection Using Machine Learning Classifiers

The HGB classifier is used at the BS to distinguish between normal and malicious nodes. Other classifiers like AdaBoost, GB, XGB, LDA, and Ridge were also tested, but HGB provided the best results.

In our system, Base Stations (BSs) manage a blockchain network to authenticate Sensor Nodes (SNs) and Cluster Heads (CHs) using unique Node IDs. Data flows from SNs to CHs, which forward it to BSs for analysis. At the BS, a Histogram Gradient Boosting (HGB) classifier evaluates whether nodes are malicious. If malicious activity is detected, the blockchain instantly revokes the node's registration to prevent network harm. For legitimate data, we use IPFS—a decentralized storage system—to split data into chunks and generate secure hashes. These hashes are then recorded on the blockchain for tamper-proof auditing.

To optimize efficiency, we replaced Proof of Work (PoW) with Verifiable Byzantine Fault Tolerance (VBFT), reducing transaction costs by 20-30% while maintaining robust security. This streamlined approach ensures real-time threat response and scalable data integrity without overburdening resource-limited nodes.

Algorithm 1 Pseudo Code of the Proposed Model

Step-1: All nodes are registered on the BSs, where blockchain is implemented.
Step-2: SNs collect data from the surrounding area and relay it to CHs, while CHs forward the data to the BSs [64].
Step-3: The BSs are trained on an ML classifier, HGB, that classifies the data and sends it either to a malicious node or a normal node.
Step-4: Data is stored on the IPFS if the BS classifies the node as a normal node. Otherwise, the BS revokes the node's registration.
Step-5: The IPFS provides hashes that are stored on the blockchain implemented on the BSs.

In our system, we use a machine learning classifier at the base stations (BSs) to distinguish between normal and malicious nodes in the network. To ensure robust detection, I compared six different classifiers: AdaBoost, Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Linear Discriminant Analysis (LDA), Ridge, and Histogram Gradient Boost (HGB). After classification, any node identified as malicious is promptly removed from the network by the BS. For example, AdaBoost, introduced by Freund and Schapire, is an ensemble method that combines several weak learners to form a strong classifier, making it both fast and efficient for tasks like ours. It works by repeatedly adjusting the weights of incorrectly classified samples, gradually improving accuracy with each round. This approach, along with the other classifiers, helps ensure that only legitimate nodes remain active in the network, maintaining its security and reliability.

Algorithm 2 AdaBoost Algorithm

Initialize the observation weights $w_i = 1/N$, i = 1, 2, …, N.
for m=1 to M
(a) Fit a classifier $G_m(x)$ to training data using weight w_i.
(b) Compute $err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$.
(c) Compute $\alpha_m = \log((1 - err_m)/err_m)$.
(d) Set $w_i \leftarrow w_i.exp[\alpha_m.I(y_i \neq G_m)]$, i = 1, 2, …, N.
endfor
Output $G_m(x) = \text{sign}[\sum_{m=1}^{M} \alpha_m G_m(x)]$.

Gradient Boosting (GB) is a supervised machine learning algorithm developed by Friedman in 2001. Unlike AdaBoost, GB builds an ensemble model by combining multiple weak learners (typically decision trees) in a sequential manner. It starts with a base model and iteratively corrects errors by training new trees on the residuals (differences between predictions and actual values) of previous ones. Each new tree focuses on reducing the remaining errors, and their predictions are combined step-by-step to form a stronger overall model. This additive approach, guided by a loss function that minimizes errors, allows GB to achieve high accuracy. However, its time complexity (O(ftnlog n)O(ftnlogn)) increases with dataset size and tree depth, making it slightly resource-intensive compared to simpler classifiers.

Algorithm 3 Gradient Boost Algorithm

Initialize model with constant value: $F_0x = argmin_r \sum_{i=1}^{n} L(y_i, r)$.
for m = 1 to M
Compute residual $r_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$ for i = 1, …, n
Train regression tree with feature x against r and create terminal nodes reasons R_{jm} for j = 1, …, m
Compute $r_{jm} = argmin_r \sum_{x_i \varepsilon R_{jm}} L(y_i, F_{(m-1)(x_i)} + r)$ for j = 1, …, m
Update the model: $F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{j_m} r_{jm} 1 (x \in R_{jm})$
endfor
Output $f_{(x)} = F_m(x)$.

J. *Dataset Description*

The WSN-DS dataset is used, containing 18 features and five classes (normal, Grayhole, Blackhole, TDMA, Flooding). The dataset is highly imbalanced, so the SMOTE technique is used for balancing.

K. *Data Sampling*

To address the imbalance in the dataset, SMOTE is applied to oversample the minority classes, ensuring fair training for the classifiers.

IV. RESULTS AND DISCUSSION

A. *Blockchain Results' Discussion*

I evaluated the system using both PoW and VBFT consensus mechanisms. VBFT significantly reduced transaction costs, especially for the registration function. IPFS was tested for file upload and download times, and it performed efficiently even for large files.

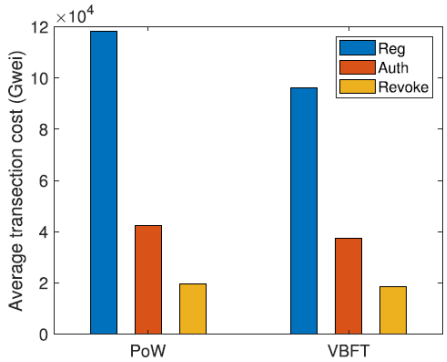


Figure 3 Comparison of PoW and VBFT consensus mechanism in terms of transaction cost.

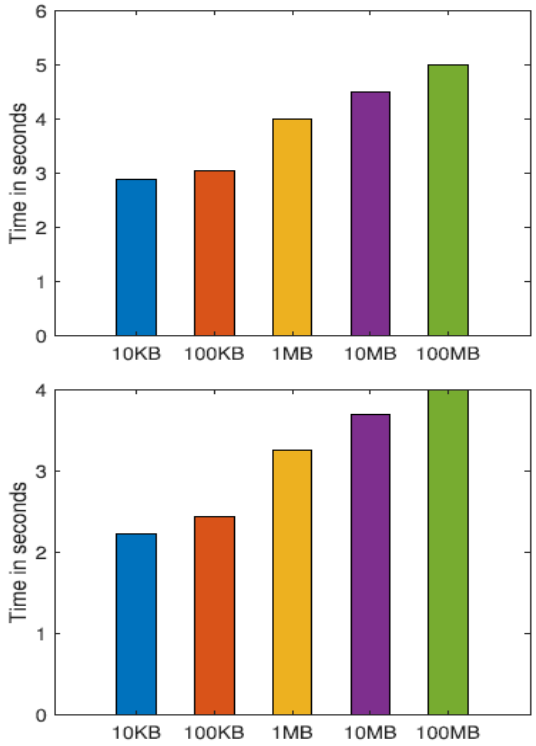


Figure 4 (a) Comparison of time consumed in uploading files on IPFS.(b) Comparison of time consumed in downloading files from IPFS.

B. Analysis Of Ml Results

The HGB classifier achieved the highest accuracy, precision, recall, and F1-score compared to other classifiers, especially on the balanced dataset. The ROC curve showed excellent classification performance.

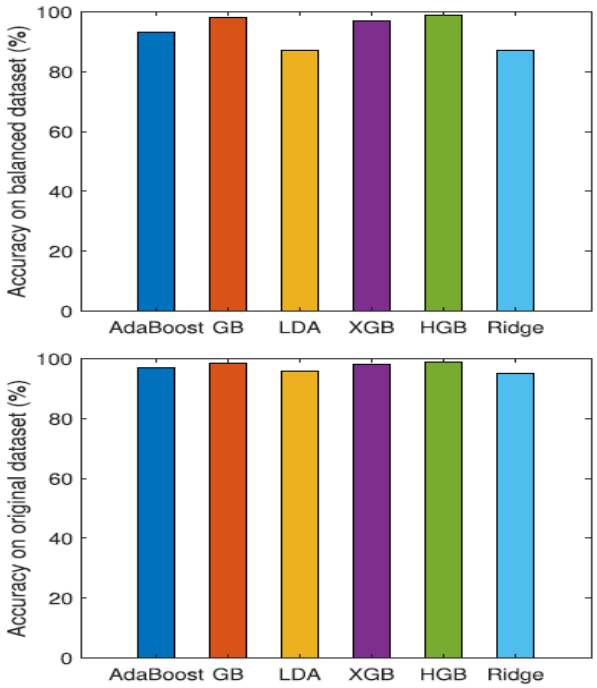


Figure 5(a) Accuracy of classifiers on the balanced dataset. (b) Accuracy of classifiers on the original dataset.

Precision measures how reliable positive predictions are, while recall checks if all actual positives are identified. In tests, the HGB classifier excelled, achieving **99% precision** on balanced data and **98%** on imbalanced data—outperforming rivals like AdaBoost, GB, and XGB by **2–16%**. This edge comes from HGB's binning method, which organizes data features into evenly distributed buckets before training weak learners. This approach reduces noise and ensures more accurate predictions, making HGB the top choice for reliably detecting malicious nodes in both balanced and real-world imbalanced scenarios.

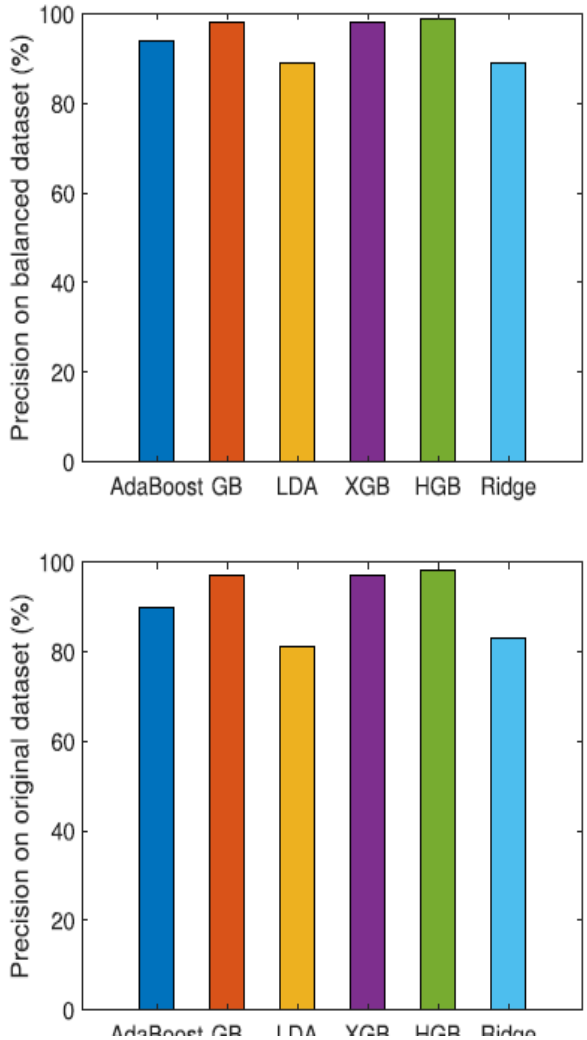


Figure 6 (a) Precision of classifiers on the balanced dataset.(b) Precision of classifiers on the original dataset.

C. Comparative Analysis Of The Proposed Model

In our tests, we compared six machine learning classifiers—AdaBoost, GB, XGB, LDA, Ridge, and HGB—to detect malicious nodes in WSNs. To address data imbalance in the WSN-DS dataset, we used SMOTE, a technique that oversamples minority attack classes (like Grayhole and Flooding) to ensure fair model training. The results showed that HGB outperformed all others, achieving the highest micro- and macro-F1 scores (99% on balanced data, 97% on raw data). For example, HGB outshined GB by 2–4%, AdaBoost by 8–10%, and Ridge by 14–16% [Tables 4–5]. This superiority stems from HGB's **binning method**, which organizes data features into evenly distributed buckets before training weak learners, improving both accuracy and efficiency. While traditional classifiers like LDA and Ridge struggled with imbalanced data, HGB's ensemble approach—combining multiple decision trees—excelled even on the original dataset [Figs. 8–9]. By integrating HGB with SMOTE, our model ensures reliable malicious node detection, making it ideal for securing resource-constrained WSNs

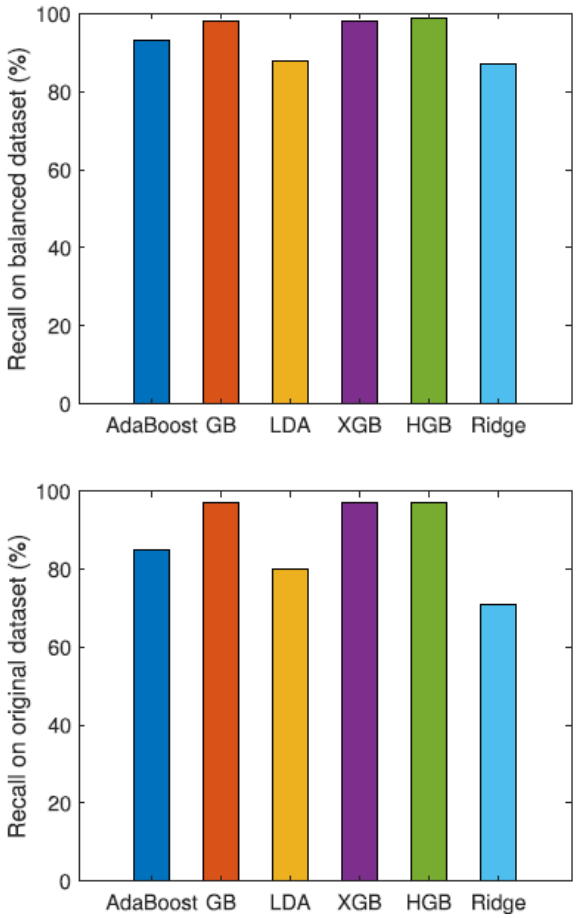


Figure 7(a) Recall of classifiers on the balanced dataset. (b) Recall of classifiers on the original dataset.

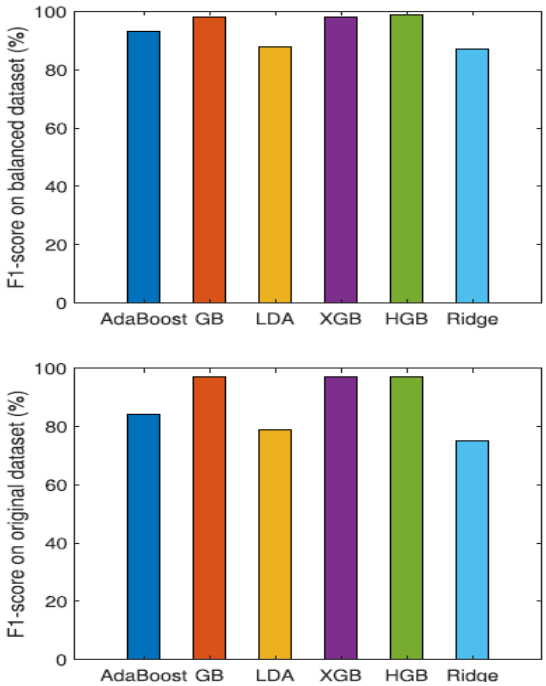


Figure 8 (a) F1-score of classifiers on the balanced dataset. (b) F1-score of classifiers on the original dataset.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. FEASIBILITY OF THE PROPOSED MODEL

The proposed approach is practical for real-world WSN deployments. Computationally intensive tasks are handled by BSs, so SNs and CHs are not overburdened. The use of blockchain and IPFS ensures data integrity and security without excessive overhead.

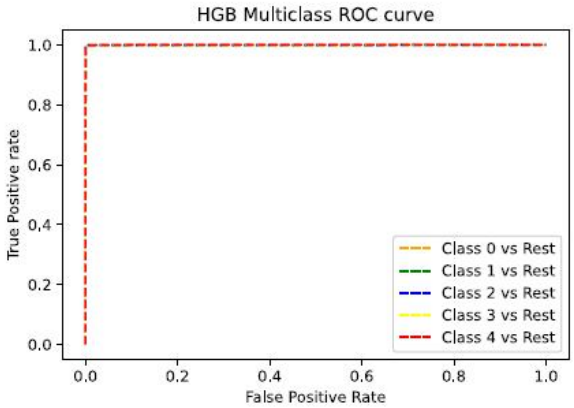


Figure 9 ROC curve of classifiers on the balanced dataset.

VI. CONCLUSION

This work presents a secure and efficient framework for detecting malicious nodes in WSNs using machine learning and blockchain. The combination of HGB for detection and VBFT for consensus achieves high accuracy and low transaction costs. In the future, I plan to explore ensemble models and smart contract vulnerability analysis to further enhance the system.

REFERENCES

[1] Nouman, U. Qasim, H. Nasir, A. Almasoud, M. Imran, and N. Javaid, "Malicious Node Detection Using Machine Learning and Distributed Data Storage Using Blockchain in WSNs," IEEE Access, vol. 11, pp. 6105–6122, Jan. 2023. DOI: 10.1109/ACCESS.2023.3236983.

[2] [2]L. Xiong, N. Xiong, C. Wang, X. Yu, and M. Shuai, ''An efficientlightweight authentication scheme with adaptive resilience of asynchronization attacks for wireless sensor networks,'' IEEE Trans. Syst.,Man, Cybern., Syst., vol. 51, no. 9, pp. 5626–5638, Sep. 2021, doi:10.1109/TSMC.2019.2957175.

[3] [3] H. Wang, P. Tu, P. Wang, and J. Yang, ''A redundant and energyefficientclusterhead selection protocol for wireless sensor network,'' in Proc. 2nd Int. Conf. Commun. Softw. Netw., 2010, pp. 554–558, doi:10.1109/ICCSN.2010.46

[4] [4] S. A. Sert, E. Onur, and A. Yazici, ''Security attacks and countermeasures in surveillance wireless sensor networks,'' in Proc. 9th Int. Conf. Appl. Inf.Commun. Technol. (AICT), Oct. 2015, pp. 201–205.

[5] [5] S. A. Sert, C. Fung, R. George, and A. Yazici, ''An efficient fuzzy pathselection approach to mitigate selective forwarding attacks in wirelesssensor networks,'' in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE),Jul. 2017, pp. 1–6.

[6] [6] R. Alkhudary, ''Blockchain technology between Nakamoto and supply chain management: Insights from academia and practice,'' SSRN Electron.J., pp. 1–12, Jul. 2020, doi: 10.2139/ssrn.3660342.

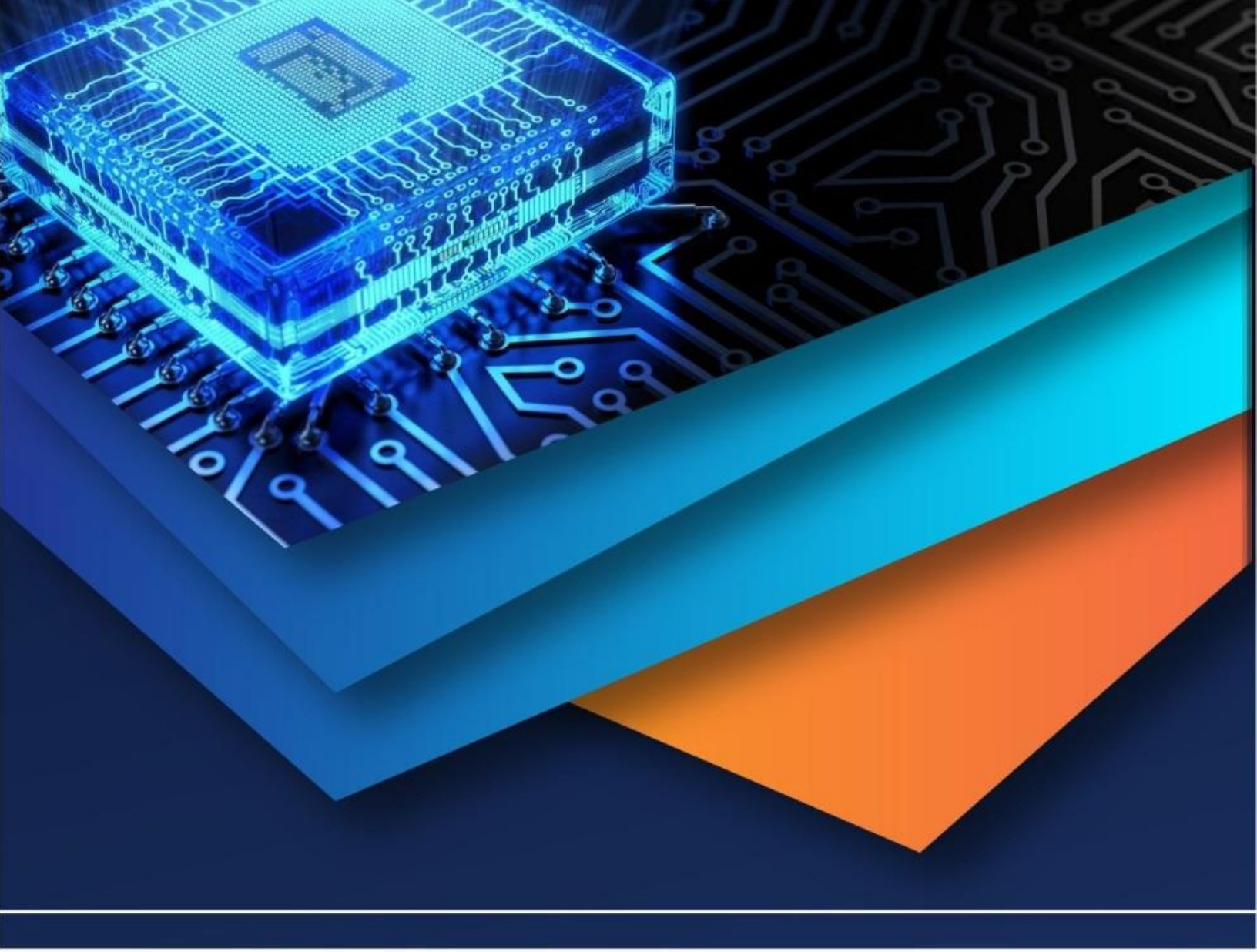
[7] [7]Z. Abubaker, N. Javaid, A. Almogren, M. Akbar, M. Zuair, andJ. Ben-Othman, ''Blockchained service provisioning and malicious nodedetection via federated learning in scalable internet of sensor things networks,''Comput. Netw., vol. 204, Feb. 2022, Art. no. 108691.

[8] [8]A. S. Yahaya, N. Javaid, M. U. Javed, A. Almogren, and A. Radwan,''Blockchain based secure energy trading with mutual verifiable fairness in a smart community,'' IEEE Trans. Ind. Informat., vol. 18, no. 11,pp. 7412–7422, Nov. 2022.

[9] [9]O. J. Pandey, V. Gautam, S. Jha, M. K. Shukla, and R. M. Hegde, ''Time synchronized node localization using optimal H-node allocation in a small world WSN,'' IEEE Commun. Lett., vol. 24, no. 11, pp. 2579–2583,Nov. 2020, doi: 10.1109/LCOMM.2020.3008086.

[10] [10] Z. Abubaker, A. U. Khan, A. Almogren, S. Abbas, A. Javaid, A. Radwan,and N. Javaid, ''Trustful data trading through monetizing IoT data usingBlockChain based review system,'' Concurrency Comput., Pract. Exper.,vol. 34, no. 5, p. e6739, Feb. 2022.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)