

(RESEARCH ARTICLE)



EXPLAINABLE MACHINE LEARNING FOR NETWORK INTRUSION DETECTION USING LIGHTGBM AND SHAP: AN IMPLEMENTATION STUDY ON HIKARI-2021

Kingsley I. Chibueze ^{1,*}, Okika Stephen Sunday ², Onyeabo Uzoamaka Agatha ³, Emeka Augustine Chinachi ⁴ and C.O. Ugwoke ⁵

¹ Department of Computer Engineering, Faculty of Engineering and the Built Environment, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria.

² Department of Electrical and Electronic Engineering, Faculty of Engineering and the Built Environment, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria.

³ Department of Biomedical Engineering, Faculty of Engineering and the Built Environment, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria.

⁴ Department of Mechanical Engineering, Faculty of Engineering and the Built Environment, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria.

⁵ Department of Civil Engineering, Faculty of Engineering and the Built Environment, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria.

* Corresponding Author

ORCID Details

Kingsley I. Chibueze: <https://orcid.org/0009-0008-9387-0297>

World Journal of Advanced Engineering Technology and Sciences, 2026, 20(03), 103–108

Publication history: Received on 08 June 2026; revised on 16 September 2026; accepted on 18 September 2026

Article DOI: <https://doi.org/10.30574/wjaets.2026.20.3.0442>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Machine-learning intrusion detection systems can provide strong predictive performance while offering limited evidence for why an individual network-flow record was classified as suspicious. This study examines the integration of SHapley Additive exPlanations (SHAP) with a LightGBM-based network intrusion detector trained on HIKARI-2021. The work is extracted from an implemented intelligent hybrid intrusion detection system (IDS), but this paper isolates the explainability contribution rather than treating the complete hybrid architecture as its subject. HIKARI-2021 contains 555,278 records and 83 original columns in the project dataset; after preprocessing, 81 features were retained. LightGBM was configured with 250 estimators, a learning rate of 0.05, 48 leaves, class weighting, and a tuned decision threshold of 0.7987. On the held-out test set, the classifier achieved 90.60% accuracy, 94.63% weighted precision, 90.60% weighted recall, 92.01% weighted F1-score, and an ROC AUC of 0.9522. Suspicious-class precision was 40.50% and recall was 81.89%. SHAP was integrated into the suspicious-prediction pathway to attribute predictions to input features, while a fallback textual explanation was implemented when SHAP output was unavailable. The source experiment did not include a formal quantitative evaluation of explanation fidelity, stability, completeness, or analyst usefulness. Accordingly, this paper reports SHAP as an implemented explainability mechanism and does not claim that SHAP itself improves predictive accuracy or reduces false positives.

Keywords: Explainable Artificial Intelligence; SHAP; Lightgbm; Intrusion Detection; HIKARI-2021; Feature Attribution; Network-Flow Classification.

1 INTRODUCTION

Machine-learning intrusion detection increasingly relies on models that learn complex relationships among network-flow features. Such models can detect patterns that are difficult to encode as deterministic rules, but their predictions may be difficult for security analysts to interpret. In operational security, an alert is not simply a class label: analysts need evidence about why a record was flagged before deciding whether to investigate or escalate it.

Explainable artificial intelligence (XAI) addresses this problem by providing information about the relationship between model inputs and outputs. Neupane et al. (2022) identified explainability as an important direction in intrusion detection research. Gaspar et al. (2024) demonstrated the use of SHAP and LIME for interpreting IDS decisions and emphasized that explanation generation should be accompanied by validation. Chen et al. (2024) further demonstrated the use of SHAP information in feature-selection and intrusion-detection analysis for encrypted traffic.

SHAP is relevant to network-flow classification because it provides feature-level contribution values for individual predictions. A suspicious prediction can therefore be accompanied by evidence about which input variables contributed toward that decision. This differs from conventional global feature importance, which describes model behavior in aggregate and does not necessarily explain a particular alert.

This paper focuses on the explainability layer of the implemented LightGBM IDS developed from the project report. It does not repeat the complete hybrid-IDS contribution addressed in Paper 1. Instead, it examines how SHAP is integrated with LightGBM, how the explanation is represented for an alert, the predictive context in which the explanations are generated, and what remains experimentally unverified.

The objectives were to: (i) define the SHAP-based explanation workflow; (ii) integrate SHAP with suspicious LightGBM predictions; (iii) distinguish local from global feature-attribution use; and (iv) identify the evidence and limitations of the implemented explainability mechanism.

2 LITERATURE REVIEW

XAI seeks to make machine-learning decisions understandable to humans. In IDS, this is particularly important because alerts may trigger incident-response actions and false alarms consume analyst attention. Neupane et al. (2022) surveyed explainable IDS methods and highlighted challenges involving model diversity, explanation quality, and stakeholder needs.

Younisse et al. (2022) investigated SHAP explanations for intrusion-detection convolutional neural networks and showed how feature-attribution information can be used to examine why a model produces particular outputs. Their work supports the use of SHAP as an explanatory layer rather than as an independent classifier.

For an individual observation, SHAP assigns contribution values to features relative to a model baseline. Positive and negative contributions indicate movement toward or away from the target prediction, respectively. Local explanations concern individual alerts, whereas global interpretation can be obtained by aggregating feature contributions over a defined sample.

Gaspar et al. (2024) compared SHAP and LIME in an IDS setting using perturbation-based analysis. Their study is important to the present work because it demonstrates that generating an explanation and validating the quality of that explanation are separate methodological questions. An IDS should therefore not be described as fully explainable merely because an XAI library can generate an output.

Chen et al. (2024) proposed an explainable feature-selection and intrusion-detection approach for encrypted traffic using SHAP and causal information, evaluating CNN and random-forest models on CICIDS2017 and NSL-KDD. Their objective was to use model interpretation to reduce the feature space while retaining detection performance. The present study differs by using SHAP primarily as a post-prediction explanation mechanism for LightGBM on HIKARI-2021.

Recent XAI-IDS research has emphasized that explanation quality should be assessed using multiple dimensions. Arreche et al. (2024) introduced the E-XAI framework for evaluating black-box XAI methods using dimensions including descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness. This provides an important

methodological boundary for the present study: implementation of SHAP is evidence of explainability functionality, but it is not, by itself, evidence of explanation fidelity or analyst usefulness.

2.1 Research Gap

The literature contains surveys of explainable IDS, SHAP-based intrusion-detection studies, and XAI evaluation frameworks. A practical integration gap remains between generating feature attributions and associating those attributions directly with the alert generated by a conventional network-flow classifier. The present implementation addresses this integration point by attaching SHAP feature attribution to suspicious LightGBM predictions within the IDS workflow and retaining the explanation with the detection result for later review.

3 MATERIALS AND METHODS

3.1 Dataset and Preprocessing

HIKARI-2021 was used as the training dataset. The dataset was introduced by Ferriyan et al. (2021) as a network intrusion-detection dataset containing benign traffic and encrypted synthetic attack traffic. In the project implementation, 555,278 records and 83 original columns were used. The fields Unnamed: 0, Unnamed: 0.1, originh, responh, traffic_category, and uid were removed, leaving 81 features. Numerical missing values were median-imputed; categorical missing values were replaced by the mode and ordinal-encoded. The binary target used 0 for normal traffic and 1 for suspected traffic.

An 80:20 stratified split produced 444,222 training records and 111,056 test records. The SHAP layer uses the same preprocessed feature representation supplied to LightGBM so that the resulting attribution corresponds to the variables actually presented to the predictive model.

3.2 LightGBM Configuration

The explained classifier used the configuration reported in the project implementation. Table 1 summarizes the principal settings.

Table 1: LightGBM configuration used for the predictions explained by SHAP.

Parameter	Configuration
Estimators	250
Learning rate	0.05
Leaves	48
Class weighting	Applied
Decision threshold	0.7987
Input features	81
Train/test split	80:20 stratified

3.3 SHAP Explanation Workflow

SHAP is invoked after the LightGBM model produces a suspicious machine-learning prediction. The preprocessed feature vector is passed to the explanation mechanism, which assigns feature-level contributions to the prediction. The application can expose the most influential features as evidence accompanying the alert. Where SHAP is unavailable, the system can produce a fallback textual summary based on traffic characteristics such as protocol, port numbers, duration, TCP flag counts, packet rate, and payload volume.

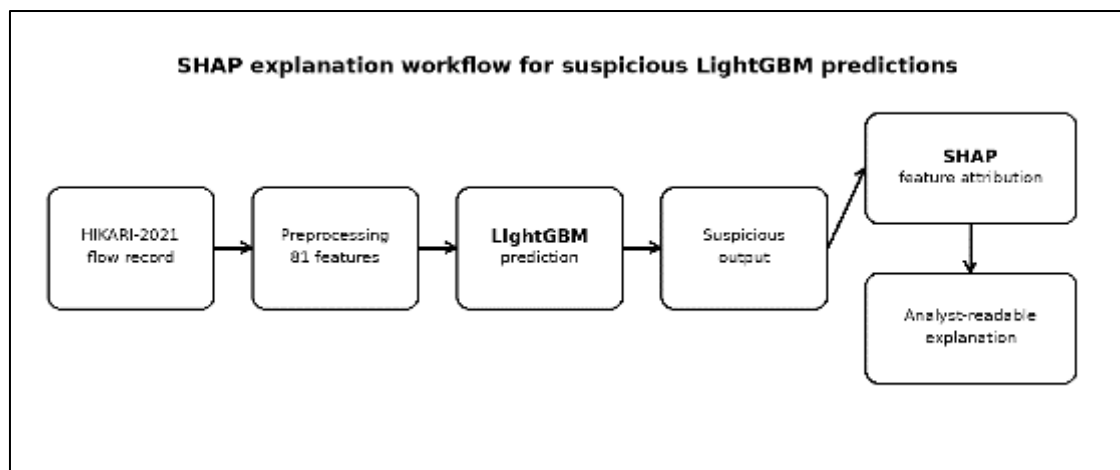


Figure 1: SHAP explanation workflow implemented around the LightGBM intrusion detector.

The workflow separates prediction from explanation. SHAP does not replace the LightGBM classifier and does not determine the initial class label; rather, it explains the feature contribution associated with a prediction already produced by the classifier.

3.4 Local and Global Interpretation

A local SHAP explanation is intended for an individual alert and identifies the features that most strongly contributed to that particular prediction. Global interpretation can be obtained by aggregating the magnitude of SHAP contributions over a defined sample and ranking features by average influence. The source project does not provide a reproducible global SHAP aggregation or a verified numerical ranking of HIKARI-2021 features. Therefore, no global feature ranking is invented in this paper.

3.5 Explanation Storage and Analyst Review

The implemented detection-result structure includes a verdict, confidence score, severity level, and an explanation field. The explanation field is designed to store SHAP or fallback natural-language evidence. This makes the explanation part of the alert record rather than a detached diagnostic artifact and supports subsequent analyst review.

4 RESULTS AND DISCUSSION

4.1 Predictive Performance of the Explained Model

The LightGBM classifier achieved 90.60% accuracy, 94.63% weighted precision, 90.60% weighted recall, and 92.01% weighted F1-score. The suspicious class achieved 40.50% precision, 81.89% recall, and 54.20% F1-score. The ROC AUC was 0.9522. These results define the predictive context in which the SHAP explanations are generated; they should not be interpreted as a quantitative evaluation of SHAP itself.

The suspicious-class results are particularly relevant to the role of XAI. Although recall was 81.89%, precision was only 40.50%, indicating a substantial false-positive burden. SHAP can provide evidence for investigating these alerts, but the current experiment does not measure whether explanations actually reduce investigation time, improve analyst agreement, or reduce false-alarm burden.

4.2 Confusion-Matrix Context

Table 2 presents the confusion matrix of the explained LightGBM classifier. The classifier correctly identified 6,174 suspicious records and missed 1,365 suspicious records. It also produced 9,071 false positives among benign records.

Table 2: Confusion matrix of the LightGBM classifier on the held-out test set.

Actual / Predicted	Benign	Suspicious
Benign	94,446	9,071
Suspicious	1,365	6,174

These values reinforce the distinction between predictive performance and explanation quality. SHAP can explain a prediction, but it does not change the confusion matrix unless a separate decision process uses the explanations to alter model outputs.

4.3 Contribution of SHAP to the IDS Workflow

The implemented SHAP layer changes the output from a class label alone to a class label accompanied by feature-level evidence. This is useful because two suspicious records can receive the same class label for different combinations of traffic features. A local explanation can reveal those differences and provide an evidence trail for analyst review.

The fallback explanation mechanism provides continuity when SHAP cannot be generated, but it should not be treated as mathematically equivalent to SHAP attribution. A rule-based textual summary is an interpretation aid, whereas SHAP values are derived from the model's predictive behavior.

4.4 Comparison with Recent Explainable IDS Studies

Table 3: Comparison of the present explainability-focused implementation with selected recent XAI-IDS studies.

Study	Focus	XAI contribution	Difference from present work
Neupane et al. (2022)	X-IDS survey	Reviews XAI methods and challenges	Survey rather than HIKARI-2021 implementation
Younisse et al. (2022)	CNN-based IDS	SHAP-based interpretation of predictions	Different model and datasets
Gaspar et al. (2024)	MLP-based IDS	SHAP/LIME comparison with perturbation analysis	Does not evaluate the present LightGBM/HIKARI-2021 pipeline
Chen et al. (2024)	Encrypted-traffic IDS	SHAP-guided feature selection	Different objective and datasets
Arreche et al. (2024)	XAI evaluation	Multidimensional XAI evaluation framework	Evaluation framework rather than this implementation
Present study	LightGBM + HIKARI-2021	SHAP attached to suspicious alert pathway	Formal quantitative XAI-quality evaluation remains future work

4.5 Limitations

The main limitation is evidential rather than architectural. The project demonstrates that SHAP was implemented and used to generate feature-based explanations, but it does not provide enough experimental output to quantify explanation fidelity, stability, sparsity, completeness, analyst agreement, or explanation latency. It also does not provide a reproducible sample of SHAP values from which a global feature ranking can be independently verified.

Accordingly, this paper does not claim that SHAP improves classification accuracy or reduces false positives. Such claims require a controlled comparison of IDS operation with and without explanations, together with formal XAI evaluation. This boundary is important because predictive performance and explanation quality are distinct properties.

5 CONCLUSION

This study presented an explainability-focused analysis of a LightGBM network intrusion detector trained on HIKARI-2021. SHAP was integrated into the suspicious-prediction pathway to provide feature-level attribution for individual machine-learning decisions, while a fallback textual explanation was implemented when SHAP output was unavailable.

The underlying LightGBM model achieved 90.60% accuracy, 94.63% weighted precision, 90.60% weighted recall, 92.01% weighted F1-score, and an AUC of 0.9522. Suspicious-class recall was 81.89% and precision was 40.50%. The principal contribution of this paper is the integration of model explanation into the IDS alert workflow; the available experiment does not establish a causal improvement in predictive performance attributable to SHAP.

Future Work

- Evaluate SHAP fidelity and stability using controlled perturbations of influential network-flow features.
- Measure explanation sparsity, completeness, robustness, and computational overhead for LightGBM on HIKARI-2021.
- Publish reproducible local SHAP examples for representative benign and suspicious records.
- Aggregate SHAP values over a defined test sample to establish a verifiable global feature-influence ranking.
- Conduct an analyst study comparing alerts with and without explanations using triage time, interpretation accuracy, and confidence measures.
- Compare SHAP with LIME and counterfactual explanations on the same LightGBM/HIKARI-2021 model.
- Evaluate explanation consistency under distribution shift and on live network-flow data.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS*Acknowledgments*

The authors acknowledge all individuals and organizations who contributed to the conduct of the study and preparation of this manuscript.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

- [1] Arreche, O., Guntur, T. R., Roberts, J. W., & Abdallah, M. (2024). E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection. *IEEE Access*, 12, 23954–23988. <https://doi.org/10.1109/ACCESS.2024.3365140>
- [2] Chen, X., Liu, M., Wang, Z., & Wang, Y. (2024). Explainable deep learning-based feature selection and intrusion detection method on the Internet of Things. *Sensors*, 24(16), 5223. <https://doi.org/10.3390/s24165223>
- [3] Ferriyan, A., Thamrin, A. H., Takeda, K., & Murai, J. (2021). Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic. *Applied Sciences*, 11(17), 7868. <https://doi.org/10.3390/app11177868>
- [4] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access*, 12, 30164–30175. <https://doi.org/10.1109/ACCESS.2024.3368377>
- [5] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10, 112392–112415. <https://doi.org/10.1109/ACCESS.2022.3216617>
- [6] Younis, R., Ahmad, A., & Abu Al-Haija, Q. (2022). Explaining intrusion detection-based convolutional neural networks using Shapley additive explanations (SHAP). *Big Data and Cognitive Computing*, 6(4), 126. <https://doi.org/10.3390/bdcc6040126>