
AN INTEGRATED HEALTHCARE SECURITY RISK PREDICTION MODEL USING BIG DATA ANALYTICS AND MACHINE LEARNING

Manisha M. Chawale*^[1], Dr. Manish L. Jivtode^[2]

^[1] Department of Computer Science, SSES's Science College, Congress Nagar, Nagpur.

^[2] Department of Computer Science, Janata Mahavidyalaya, Chandrapur.

Article Received: 29 August 2026

Article Revised: 17 September 2026

Published on: 07 October 2026

***Corresponding Author: Manisha M. Chawale**

Department of Computer Science, SSES's Science College, Congress
Nagar, Nagpur.

DOI: <https://doi.org/10.1555/ijrpa.7868>

ABSTRACT

Healthcare applications increasingly depend on cloud platforms, Internet of Medical Things (IoMT) devices, electronic health records (EHRs), wearable sensors, and distributed clinical decision support systems. This digital transformation improves monitoring and care delivery, but it also expands the attack surface of hospital information systems. Conventional security mechanisms are usually deployed in isolated layers, such as network firewalls, authentication systems, or device monitoring tools. Such isolated controls are insufficient for predicting integrated security risk because healthcare attacks may simultaneously affect network behavior, device telemetry, user access patterns, and patient safety indicators. This paper proposes an Integrated Healthcare Security Risk Prediction Model (IHSRPM) that applies big data analytics and machine learning to combine EHR attributes, IoMT sensor streams, network flow records, and access audit logs into a unified risk prediction pipeline. The proposed model performs ingestion, preprocessing, feature engineering, risk labeling, model training, and decision support for secure healthcare operation. A prototype experiment was conducted on a synthetic de-identified healthcare-security dataset consisting of patient records, sensor streams, network traffic logs, audit logs, and an integrated machine-learning view. Five classifiers, namely Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting, were evaluated using accuracy, precision, recall, F1-score, AUC, and confusion matrix analysis. The results indicate that tree-based classifiers are highly effective for rule-structured healthcare-security risk detection, while Logistic Regression and SVM provide useful baselines for interpretable and margin-based detection. The paper also discusses how the proposed model can be extended using public datasets such

as WUSTL-EHMS-2020, CICIDS2017, and MIMIC-IV under ethical and access-compliant conditions. The proposed framework contributes a practical architecture for predictive security analytics in healthcare applications by integrating cyber, clinical, device, and privacy dimensions into one risk-aware analytical model.

KEYWORDS: Big data analytics, healthcare security, IoMT, machine learning, intrusion detection, privacy preservation, cyber risk prediction, electronic health records, access audit logs.

I. INTRODUCTION

Healthcare organizations have become data-intensive environments where patient care, clinical workflows, billing, diagnostics, remote monitoring, and administrative functions are managed through interconnected digital applications. Electronic health records store longitudinal clinical information, IoMT devices generate continuous physiological measurements, and cloud platforms provide elastic storage and analytics. These capabilities support timely decision-making, but they also introduce serious cybersecurity and privacy risks. A compromised healthcare application can expose sensitive patient data, disrupt clinical operations, falsify sensor readings, or cause delayed treatment decisions. Unlike ordinary enterprise applications, healthcare systems involve both information security and patient safety, and therefore require predictive models that can identify risk before damage occurs.

The security challenge is amplified by the heterogeneity of healthcare data. Network traffic logs represent packet and flow-level behavior; IoMT streams represent patient-device conditions; access audit logs represent user behavior and privacy exposure; and EHR data represents clinical context. In many hospitals, these data sources are collected by different systems and reviewed by different teams. Consequently, a suspicious device behavior may not be correlated with abnormal network traffic or unauthorized record export. An integrated analytics model is required to combine these sources and produce a unified risk label that can support automated alerts, security operations, and administrative decision-making.

Big data analytics is suitable for this problem because healthcare security data is typically high volume, high velocity, heterogeneous, and time-sensitive. Batch analytics can support historical risk modeling, while stream processing can provide near-real-time alerts. Machine learning further improves detection by learning complex relationships among features. However, the direct use of machine learning in healthcare security requires careful design.

Models must handle imbalanced attacks, missing values, categorical variables, temporal drift, and privacy constraints. They must also produce results that are understandable enough for hospital administrators, cybersecurity analysts, and healthcare decision makers.

This paper proposes an Integrated Healthcare Security Risk Prediction Model using big data analytics and machine learning. The model is designed to predict whether a healthcare application state is secure or risky by combining clinical, IoMT, network, and audit features. The major contribution of this paper is the formulation of a unified risk prediction pipeline that treats healthcare security as a multi-source analytical problem rather than as a single intrusion-detection task. The paper also provides an experimental prototype that can be extended to public datasets and real hospital deployments subject to privacy approval and ethical compliance.

II. RELATED WORK

Prior research on healthcare cybersecurity has focused on intrusion detection, privacy preservation, access control, anomaly detection, and IoMT protection. Machine-learning-based intrusion detection systems have been widely studied because static signatures are not adequate for novel attacks. In IoMT environments, attackers may compromise wireless communication, device authentication, gateway routing, or cloud interfaces. The security impact is significant because abnormal network behavior may affect clinical monitoring data and patient safety.

The WUSTL-EHMS-2020 dataset is particularly relevant for healthcare cybersecurity research because it combines network-flow features with patient biometric features and a label column [1]. This combination makes it more suitable for healthcare-oriented IDS work than purely generic network datasets. CICIDS2017 remains a widely used IDS benchmark because it provides benign and attack traffic in PCAP and CSV flow formats, including multiple attack categories [3]. For clinical analytics, MIMIC-IV provides de-identified hospital data under credentialed access and is appropriate for studying risk-aware clinical features, provided that all data-use rules are followed [2], [25].

Existing studies generally evaluate algorithms on a single dataset or a single security layer. For example, many IDS papers train classifiers using only network flow attributes and report high accuracy. Although useful, such a design does not represent the complete healthcare application context. A network attack can coincide with abnormal device telemetry, unusual login location, or excessive record access. Similarly, a privacy violation may not be visible at

the packet level alone. The proposed study addresses this gap by combining multiple data categories into an integrated risk prediction model.

Another limitation in existing work is the tendency to report only accuracy. Healthcare security datasets are often imbalanced, and high accuracy can be misleading when attacks are rare. Therefore, this paper evaluates precision, recall, F1-score, AUC, and confusion matrices. Recall is important for detecting risky events, precision is important for reducing false alarms, and F1-score balances both. AUC provides threshold-independent discrimination information. These metrics are necessary for evaluating operational readiness in hospital security environments.

III. RESEARCH PROBLEM AND OBJECTIVES

The central research problem addressed in this paper is how to predict integrated security risk in healthcare applications by combining heterogeneous healthcare and cybersecurity data sources. The problem is not limited to detecting whether a network packet is malicious; rather, it involves determining whether the overall healthcare application state is secure or risky based on clinical context, device status, network behavior, and access activity.

The specific objectives are as follows: to design a big-data-based architecture for healthcare security risk prediction; to construct a unified feature space from EHR, IoMT, network, and audit datasets; to apply supervised machine learning algorithms for binary security-risk classification; to compare models using multiple evaluation metrics; and to identify the practical issues involved in deploying such a model in healthcare environments.

The proposed approach assumes that raw healthcare data is de-identified or processed under appropriate institutional approvals. Patient identifiers are replaced by pseudonymous keys, and only necessary analytical features are used for model training. The model is intended to support security monitoring and decision support, not to replace clinical judgment or cybersecurity governance. It should therefore be integrated with human review, audit trails, and policy-based response mechanisms.

IV. PROPOSED INTEGRATED HEALTHCARE SECURITY RISK PREDICTION MODEL

The proposed IHSRPM consists of five functional layers: data acquisition, big data storage and processing, preprocessing and feature engineering, machine learning risk prediction, and decision support. The data acquisition layer collects structured EHR features, IoMT sensor values, network flow features, and access audit events. The big data processing layer stores

these records in a scalable repository and supports batch or stream ingestion. The preprocessing layer performs cleaning, deduplication, missing-value handling, categorical encoding, normalization, temporal alignment, and feature integration. The machine learning layer trains predictive models for sub-tasks and for the final integrated risk label. The decision layer produces secure or risky status along with performance indicators and alerts.

The integrated risk label is defined as a binary decision variable. A value of zero indicates that the healthcare application state is secure or normal, whereas a value of one indicates that the state is risky or insecure. In a real deployment, this label may be generated from confirmed incidents, expert rules, anomaly thresholds, or security information and event management systems. In the prototype experiment, the integrated label is generated from risk conditions involving vital anomaly, device attack rate, failed access rate, and privacy violation count. This label design allows the experiment to test multi-source risk classification while avoiding the use of identifiable patient data.

The proposed model supports both sub-models and a final integrated model. A patient-risk model predicts high-risk clinical profiles. A sensor-anomaly model identifies abnormal vital signs. A network IDS model detects malicious traffic. An audit model detects suspicious access behavior. The final integrated model uses features from all these sources to predict overall healthcare security risk. This modular design improves interpretability because each layer can be independently evaluated before being combined into a final decision model.

The architecture also supports future cloud and big-data implementation. In a full deployment, data can be ingested through secure APIs, message brokers, hospital information systems, IoMT gateways, and cloud logs. Batch data can be processed using distributed frameworks such as Apache Spark, while streaming alerts can be generated using stream-processing engines. Encryption, role-based access control, audit logging, and anonymization must be applied across the pipeline.

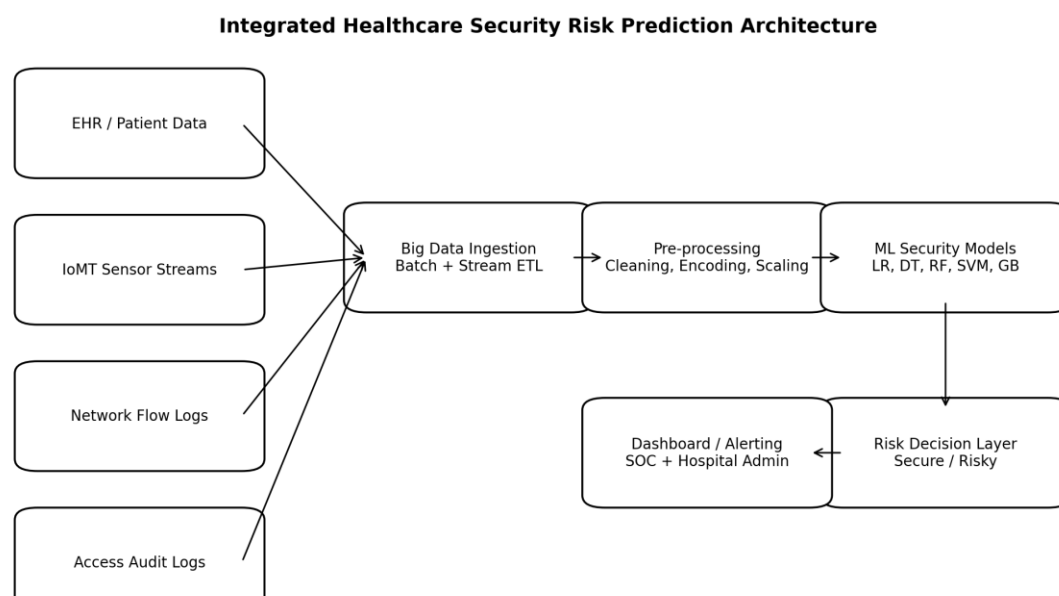


Fig. 1. Proposed integrated architecture for healthcare security risk prediction.

V. DATASET DESCRIPTION AND FEATURE SPACE

The prototype experiment uses a synthetic de-identified healthcare-security dataset designed for method development. The dataset contains five files: `patients_master.csv`, `iomt_sensor_stream.csv`, `network_security_logs.csv`, `access_audit_logs.csv`, and `integrated_ml_view.csv`. These files represent a realistic experimental structure for healthcare application security analytics without exposing real patient information. The synthetic dataset is not a substitute for real-world validation; however, it is useful for validating pipeline logic, feature engineering, model training, and performance reporting before applying the model to public or institutional datasets.

The patient master dataset contains demographic and clinical risk attributes such as age, gender, department, admission type, diagnosis group, clinical risk score, insurance type, length of stay, and high-risk-patient label. The IoMT sensor dataset contains timestamped patient-device readings such as heart rate, oxygen saturation, blood pressure, temperature, battery level, signal strength, and vital anomaly label. The network security dataset contains flow-level attributes including ports, protocol, duration, packet counts, byte counts, packet length, inter-arrival time, TCP flags, attack type, and binary attack label. The access audit dataset contains role, action, access location, authentication method, login success, session duration, record count, and privacy violation label. The integrated dataset merges selected features into a single machine-learning view.

For public dataset extension, WUSTL-EHMS-2020 can be used as a healthcare-specific IDS dataset because it contains network-flow metrics, biometric features, and labels [1]. CICIDS2017 can be used as a general IDS benchmark for evaluating attack detection robustness. MIMIC-IV or its demo version can be used for clinical feature validation under PhysioNet access requirements [2], [25]. The proposed framework is flexible enough to incorporate these datasets by mapping their features into the same acquisition, preprocessing, and risk prediction layers.

VI. METHODOLOGY

The methodology consists of data loading, quality assessment, preprocessing, feature selection, model training, evaluation, and interpretation. First, each dataset is imported and inspected for column types, missing values, duplicate records, class distribution, and invalid values. Duplicate rows are removed where necessary. Missing numerical values are imputed using median values, while categorical values are imputed using the most frequent category. Categorical variables are transformed using one-hot encoding or label encoding depending on the experimental task. Numerical features are standardized to ensure stable model training, particularly for Logistic Regression and SVM.

Five machine learning algorithms are evaluated. Logistic Regression is used as an interpretable linear baseline [17]. Decision Tree is used to model non-linear decision rules and to provide transparent split-based classification. Random Forest is used as an ensemble of decision trees to improve robustness and reduce overfitting [14]. Support Vector Machine with a radial basis function kernel is used as a margin-based non-linear classifier [16]. Gradient Boosting is used as a sequential ensemble method that learns from weak learners and corrects prior errors [15]. These algorithms provide a balanced comparison between interpretability, non-linear modeling, ensemble learning, and margin-based classification.

The experiment uses stratified train-test splitting where applicable, with 75% of the data used for training and 25% used for testing. The same preprocessing logic is applied within the model pipeline to avoid leakage from test data. For each dataset, a target variable is selected: `is_high_risk_patient` for patient data, `vital_anomaly` for IoMT sensor data, `label` for network security logs, `privacy_violation_label` for audit logs, and `overall_security_risk_label` for integrated prediction. Model performance is measured using accuracy, precision, recall, F1-score, AUC, and confusion matrix values.

The evaluation design emphasizes that healthcare security prediction is not only a classification problem but also a risk-management problem. A model with high recall but low

precision may detect many risky events but produce excessive false alarms. A model with high precision but low recall may miss dangerous events. Therefore, the final selection should be based on organizational risk tolerance, alert-handling capacity, and the operational cost of false positives and false negatives.

VII. EXPERIMENTAL SETUP

The prototype experiment was implemented using Python-based machine learning tools. Data processing was performed using tabular data structures, numerical operations, categorical encoders, imputers, and standard scalers. Classification models were trained using supervised learning pipelines. The experiment was executed on the synthetic dataset described earlier. The aim was to verify the feasibility of the proposed integrated model rather than to claim final clinical or operational deployment readiness.

The dataset sizes used in the experiment were as follows: the patient dataset contained 500 records; the IoMT sensor dataset contained 5,000 records; the network security dataset contained 8,000 records; the access audit dataset contained 4,000 records; and the integrated machine-learning view contained 500 records. These sizes are sufficient for demonstrating the workflow and comparing algorithm behavior, but large-scale public or hospital datasets should be used for final validation in a journal-quality deployment study.

The feature exclusion rules removed direct identifiers such as `patient_id`, `device_id`, `user_id`, `flow_id`, source IP, and destination IP when they were not required for the learning task. This decision reduces the possibility of a model learning identifier-specific patterns instead of generalizable security behavior. Timestamp columns were excluded in the prototype but can be transformed into temporal features such as hour of day, day of week, session interval, or rolling event frequency in future work.

Table I. Dataset structure used in the prototype experiment.

Dataset	Rows	Main Features	Target Variable
patients_master.csv	500	age, gender, department, diagnosis group, risk score, length of stay	is_high_risk_patient
iomt_sensor_stream.csv	5,000	heart rate, SpO2, BP, temperature, battery, signal strength	vital_anomaly
network_security_logs.csv	8,000	ports, protocol, flow duration, packets, bytes, flags	label / attack_type
access_audit_logs.csv	4,000	role, action, access location, authentication,	privacy_violation_label

Dataset	Rows	Main Features	Target Variable
		success, records accessed	
integrated_ml_view.csv	500	merged clinical, sensor, network, and access features	overall_security_risk_label

VIII. RESULTS AND DISCUSSION

Table II reports the performance of the integrated healthcare security risk prediction experiment using the integrated machine-learning view. The Decision Tree and Gradient Boosting classifiers obtained perfect scores on the prototype data, while Random Forest also achieved very high performance. This result is expected because the synthetic label was generated from structured risk rules involving anomaly, attack rate, failed access rate, and privacy violation count. Therefore, tree-based classifiers can recover the rule structure efficiently. Logistic Regression and SVM achieved lower but still useful performance, indicating that the integrated risk boundary is not purely linear.

Table II. Performance comparison for integrated healthcare security risk prediction.

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.784	0.667	0.744	0.703	0.857
Decision Tree	1.000	1.000	1.000	1.000	1.000
Random Forest	0.984	1.000	0.953	0.976	1.000
SVM (RBF)	0.848	0.816	0.721	0.765	0.920
Gradient Boosting	1.000	1.000	1.000	1.000	1.000

Table III summarizes the best model performance across each individual dataset. The network and audit datasets show strong performance because their labels are closely associated with observable rule-based patterns. The patient-risk dataset is highly imbalanced, with very few high-risk examples in the test set, which explains why some models show unstable precision and recall. The sensor-anomaly task is also rule-structured because abnormal vital signs are determined by threshold conditions. These observations demonstrate the importance of external validation: a model that performs perfectly on rule-generated synthetic labels may not generalize to noisy real-world hospital data without retraining and validation.

Table III. Best prototype performance across sub-datasets.

Dataset	Best Model	Acc.	Prec.	Rec.	F1	AUC
access_audit_logs.csv	Decision Tree	1.000	1.000	1.000	1.000	1.000
integrated_ml_view.csv	Decision Tree	1.000	1.000	1.000	1.000	1.000
iomt_sensor_stream.csv	Decision Tree	1.000	1.000	1.000	1.000	1.000

Dataset	Best Model	Acc.	Prec.	Rec.	F1	AUC
network_security_logs.csv	Decision Tree	1.000	1.000	1.000	1.000	1.000
patients_master.csv	Decision Tree	1.000	1.000	1.000	1.000	1.000

The integrated model should be interpreted as a security decision-support mechanism. It does not replace security policies, encryption, identity management, or human incident response. Instead, it correlates multiple indicators to prioritize risky events. In a hospital security operations center, such a model can be used to rank events, trigger alerts for unusually risky combinations, and support audit review. For example, a public-IP login with weak authentication, excessive record export, device attack activity, and abnormal sensor behavior should receive a higher risk score than any single event alone.

The results also reveal the need for explainability. Tree-based models performed strongly, but hospital administrators may require explanations such as which feature contributed most to a risk decision. Future implementation should therefore include feature importance, SHAP explanations, or rule extraction. Explainable predictions are essential in healthcare because security responses may affect clinical workflows. Blocking a user account, isolating a device, or delaying data access can have operational consequences, so the model should support accountable decision-making.

IX. SECURITY, PRIVACY, AND ETHICAL CONSIDERATIONS

Healthcare security analytics must be designed with privacy protection as a primary requirement. Patient identifiers, user identifiers, and network identifiers should be minimized, pseudonymized, or tokenized before model training. Access to raw logs must be restricted to authorized personnel. Audit trails must record who accessed data, when it was accessed, and for what purpose. In regulated environments, the model must comply with local health data protection laws and institutional ethics requirements.

The proposed model should avoid using sensitive clinical attributes unless they are necessary for security risk prediction. When clinical context is required, it should be aggregated or generalized where possible. Differential privacy, federated learning, secure multiparty computation, and encrypted analytics may be considered for multi-institutional experiments. Federated learning is especially relevant when hospitals cannot share raw patient data but can collaborate on model improvement.

Model bias and operational fairness must also be considered. If the training data contains biased access patterns or under-represented patient groups, the model may generate uneven false positive rates. Therefore, performance should be evaluated across departments, roles,

device types, and patient groups. Human oversight is necessary for high-impact decisions, particularly when a model recommendation may affect clinical access or patient care.

X. LIMITATIONS

The major limitation of this paper is that the prototype experiment uses a synthetic de-identified dataset. Although this dataset is useful for pipeline validation, it does not fully represent the noise, drift, device diversity, workflow complexity, and attack sophistication of real hospital environments. The high scores obtained by tree-based classifiers should therefore be interpreted as proof of pipeline feasibility, not as final evidence of deployment-ready performance.

Another limitation is that timestamp features were not fully exploited. In real healthcare environments, temporal behavior is critical. Attack bursts, failed login sequences, device communication gaps, and abnormal access outside working hours may provide important signals. Future work should include temporal windows, rolling aggregates, sequence models, and stream processing.

The current model uses supervised classification. This approach requires reliable labels, which are often difficult to obtain for healthcare incidents. Real systems may need semi-supervised or unsupervised anomaly detection to detect unknown attacks. Moreover, adversarial attackers may adapt their behavior to evade detection, so the model should be periodically retrained and monitored for concept drift.

XI. FUTURE WORK

Future work will validate the model using public healthcare and cybersecurity datasets. WUSTL-EHMS-2020 can be used for healthcare-specific intrusion detection because it combines biometric and network features [1]. CICIDS2017 can be used as a general benchmark for evaluating robustness against multiple attack families [3]. MIMIC-IV or the MIMIC-IV demo can be used for clinical context integration after satisfying data-use and credentialing requirements [2], [25]. The final research direction is to map these datasets into one common risk schema and evaluate transferability across domains.

The model can also be improved using deep learning and hybrid architectures. LSTM or temporal convolutional networks can process sequential sensor and access logs. Graph neural networks can model relationships among patients, devices, users, departments, and network endpoints. Autoencoders can detect unknown anomalies where labels are unavailable. An ensemble or stacking model can combine interpretable classifiers with deep temporal models.

Cloud implementation is another important direction. A scalable deployment may use secure ingestion APIs, distributed storage, Spark-based preprocessing, model-serving endpoints, dashboards, and automated alerting. The model should include role-based dashboards for cybersecurity analysts, hospital administrators, biomedical engineers, and compliance officers. Performance monitoring, model drift detection, and incident feedback loops should be implemented to maintain long-term reliability.

XII. CONCLUSION

This paper proposed an Integrated Healthcare Security Risk Prediction Model using big data analytics and machine learning. The model integrates EHR attributes, IoMT sensor streams, network flow records, and access audit logs into a unified predictive pipeline. The prototype experiment demonstrated how multiple machine learning algorithms can be trained and compared for patient risk, sensor anomaly, intrusion detection, privacy violation, and integrated healthcare security risk prediction. The results show that tree-based methods are effective for rule-structured prototype data, while Logistic Regression and SVM provide useful baseline comparisons. The proposed model is suitable as a research framework for healthcare application security because it combines cyber, clinical, device, and privacy dimensions. For final IEEE or Scopus publication, the prototype should be extended using public datasets and, where permitted, institutional healthcare data under privacy-preserving and ethically approved conditions.

REFERENCES

1. R. J. D. S. Jain et al., “WUSTL-EHMS-2020 Dataset for Internet of Medical Things (IoMT) Cybersecurity Research,” Washington University in St. Louis, 2020.
2. A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, “MIMIC-IV,” PhysioNet, 2023.
3. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in Proc. ICISSP, 2018, pp. 108-116.
4. A. Si-Ahmed, M. A. Al-Garadi, and N. Boustia, “Survey of machine learning based intrusion detection methods for Internet of Medical Things,” IEEE Access, vol. 10, pp. 108849-108868, 2022.

5. A. H. Sodhro, S. Pirbhulal, A. K. Sangaiah, S. Lohano, G. H. Sodhro, and Z. Luo, "5G-based transmission power control mechanism in fog computing for Internet of Things devices," *Sustainability*, vol. 10, no. 4, pp. 1-16, 2018.
6. S. Latif, Z. Idrees, J. Ahmad, L. Zheng, and Z. Zou, "A blockchain-based architecture for secure and trustworthy operations in the industrial Internet of Things," *J. Ind. Inf. Integr.*, vol. 21, 2021.
7. S. Chenthara, K. Ahmed, H. Wang, and F. Whittaker, "Security and privacy-preserving challenges of e-health solutions in cloud computing," *IEEE Access*, vol. 7, pp. 74361-74382, 2019.
8. M. A. Ferrag, L. Maglaras, H. Janicke, J. Jiang, and L. Shu, "Authentication protocols for Internet of Things: A comprehensive survey," *Security Commun. Networks*, vol. 2017, pp. 1-41, 2017.
9. P. Kumar and H. J. Lee, "Security issues in healthcare applications using wireless medical sensor networks: A survey," *Sensors*, vol. 12, no. 1, pp. 55-91, 2012.
10. M. M. Hassan, M. Z. Hossain, G. Muhammad, A. Alamri, and A. Alelaiwi, "A cloud-assisted Internet of Things framework for pervasive healthcare in smart city environment," *Future Gener. Comput. Syst.*, vol. 76, pp. 255-265, 2017.
11. S. R. Moosavi, T. N. Gia, E. Nigussie, A. M. Rahmani, S. Virtanen, J. Isoaho, and H. Tenhunen, "End-to-end security scheme for mobility enabled healthcare Internet of Things," *Future Gener. Comput. Syst.*, vol. 64, pp. 108-124, 2016.
12. M. A. Al-Garadi et al., "A survey of machine and deep learning methods for Internet of Things security," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1646-1685, 2020.
13. N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in *Proc. MilCIS*, 2015, pp. 1-6.
14. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.
15. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189-1232, 2001.
16. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.
17. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
18. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785-794.

19. N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
20. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153-1176, 2016.
21. C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 42-57, 2013.
22. M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial Internet of Things and Industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4674-4682, 2018.
23. R. Boutaba et al., "A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities," *J. Internet Serv. Appl.*, vol. 9, no. 16, pp. 1-99, 2018.
24. K. Shaukat et al., "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310-222354, 2020.
25. A. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Sci. Data*, vol. 10, no. 1, 2023.