ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Predictive Modeling of Cyber Incident Escalation Risk in Hospital Electronic Medical Record (EMR) Systems Using Ensemble Learning Models

Genevieve Donkor Armah[1]; Idoko Peter Idoko[2]; Yewande Iyimide Adeyeye[3]; Lawrence Anebi Enyejo [4]; Tony Isioma Azonuche[5]

[1] Department of Economics, Youngstown State University, Youngstown Ohio, USA
[2]Department of Electrical/ Electronic Engineering, University of Ibadan, Nigeria
[3] Department of Day Case Surgery, Dumfries and Galloway Royal Infirmary, Dumfries, United Kingdom
[4]Department of Telecommunications, Enforcement Ancillary and Maintenance, National Broadcasting Commission, Aso-Villa, Abuja, Nigeria
[5] Department of Project Management, Amberton University, Garland Texas, USA

Publication Date: 2026/02/24

Abstract: Hospital Electronic Medical Record (EMR) systems constitute mission-critical clinical infrastructure whose compromise can directly disrupt care delivery, threaten patient safety, and expose healthcare organizations to regulatory and financial risk. Contemporary security operations in hospitals remain largely reactive, relying on static severity scoring and post-incident classification that provide limited support for anticipating whether an observed cyber event will escalate into a high-impact incident. This study addresses this gap by proposing a predictive, time-bounded framework for modeling cyber incident escalation risk in EMR environments using ensemble learning methods. We develop a retrospective observational study leveraging multi-source, PHI-safe operational telemetry, including security alerts, identity and access logs, endpoint and network signals, EMR audit metadata, and incident management records. Escalation is formalized as a forward-looking outcome defined by severity reclassification, containment intensity, operational downtime, or confirmed data compromise within specified horizons (6, 24, and 72 hours). A comprehensive feature engineering strategy integrates event-level indicators, identity anomalies, endpoint behaviors, network propagation signals, and EMR workflow context. Multiple ensemble models bagging, boosting, and stacking are evaluated against baseline approaches under severe class imbalance, with emphasis on probability calibration and decision-aligned metrics. Results demonstrate that ensemble models substantially outperform baselines in identifying escalation-prone events, particularly at short and medium horizons, while calibrated probabilities enable actionable threshold-based triage. Explainability analysis reveals that escalation risk is driven by the interaction of identity misuse, lateral movement, privilege changes, and anomalous EMR access patterns rather than isolated signals. Operational case studies show how probabilistic escalation forecasting can reduce time-to-containment and unnecessary disruptions when embedded within human-in-the-loop security workflows. The study contributes an escalation-focused modeling framework, interpretable risk signals aligned with clinical operations, and deployment guidance for hospital security operations. Overall, the findings demonstrate that calibrated ensemble learning can meaningfully enhance proactive cyber risk management in EMR systems when integrated with disciplined governance and incident response practices.

Keywords: *Electronic Medical Records (EMR); Healthcare Cybersecurity; Incident Escalation Risk; Ensemble Learning; Predictive Analytics; Security Operations Center (SOC); Ransomware.*

How to Cite: Genevieve Donkor Armah; Idoko Peter Idoko; Yewande Iyimide Adeyeye; Lawrence Anebi Enyejo; Tony Isioma Azonuche (2026) Predictive Modeling of Cyber Incident Escalation Risk in Hospital Electronic Medical Record (EMR) Systems Using Ensemble Learning Models. *International Journal of Innovative Science and Research Technology*, 11(2), 1312-1347. https://doi.org/10.38124/ijisrt/26feb578

I. INTRODUCTION

- *Background: EMR Systems as High-Value, High-Availability Clinical Infrastructure*

Electronic Medical Record (EMR) systems have evolved into the central digital backbone of modern hospital operations, underpinning clinical decision-making, care coordination, billing, and regulatory reporting. By integrating patient histories, diagnostic results, medication orders, and

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

clinical workflows across departments, EMRs enable interoperability among clinicians, laboratories, pharmacies, and external health information exchanges, thereby supporting continuity and quality of care (Adler-Milstein & Huckman, 2013; HealthIT.gov, 2019; Aluso, 2021). Their role extends beyond data storage to real-time clinical operations, where system availability directly affects patient throughput, diagnostic accuracy, and treatment timeliness. As a result, EMRs are classified as mission-critical infrastructure within healthcare organizations, with downtime measured not merely in financial losses but in delayed or compromised patient care (Kruse et al., 2017; Aluso et al., 2024).

The high availability requirements of EMR systems make them particularly sensitive to cyber disruptions. Hospitals operate in environments characterized by legacy systems, heterogeneous medical devices, and continuous access demands, which together increase the attack surface and limit the feasibility of extended system shutdowns for security remediation (Kellermann & Jones, 2013; Aluso et al., 2026). Consequently, cyber incidents affecting EMRs, such as ransomware infections, credential compromise, or database corruption, often propagate rapidly across interconnected clinical systems. What may begin as a low-level security alert can escalate into a hospital-wide operational crisis if not contained promptly.

Importantly, cyber incident escalation in EMR environments should be understood as a patient-safety and service-availability problem rather than solely an information technology issue. Empirical evidence shows that cyberattacks on hospital information systems are associated with care delays, diversion of emergency patients, increased length of stay, and heightened risk of adverse clinical outcomes (Gordon et al., 2021; McLeod & Dolezel, 2018; Anim-Sampong et al., 2022; Animasaun et al., 2025). Regulatory and public health bodies have similarly emphasized that loss of access to electronic health records during cyber incidents can impair clinical judgment and disrupt coordinated care delivery, thereby posing direct risks to patient safety (World Health Organization, 2021; Aluso et al., 2022; Animasaun et al., 2026). Framing cyber incident escalation in this broader socio-technical context highlights the need for predictive, risk-aware approaches that prioritize early containment to protect both clinical operations and patient outcomes, rather than reactive responses focused narrowly on technical system recovery.

- *Problem Statement: Why "Incident Escalation Risk" Is a Distinct Predictive Target*

In hospital Electronic Medical Record (EMR) environments, a cybersecurity incident should not be viewed as a static or binary event but as a dynamic process that can evolve in severity over time. Incident escalation refers to the progression of an initially low-severity security event into a high-impact incident with broader organizational consequences, such as lateral system compromise, prolonged EMR downtime, unauthorized exposure of sensitive health data, or failure to contain ransomware before operational disruption occurs (Behl & Behl, 2017; Gordon et al., 2021; Animasaun et al., 2026). This progression is particularly acute in healthcare settings, where tightly coupled clinical systems, continuous access requirements, and constrained response windows create conditions in which small anomalies can rapidly amplify into systemic crises.

Despite this reality, most hospital cybersecurity practices remain anchored in static risk scoring and after-the-fact incident classification. Traditional risk assessment frameworks typically assign severity based on predefined alert categories, asset criticality, or compliance checklists, often without incorporating temporal context or evolving attacker behavior (ENISA, 2016; McLeod & Dolezel, 2018; Anokwuru, 2024). Similarly, incident classification is frequently performed post hoc, once the scope and impact are already known, limiting its utility for real-time decision-making. These approaches assume that severity is inherent at detection time, rather than an emergent property shaped by response delays, environmental conditions, and adversary persistence. As a result, security operations teams may underestimate early-stage events that later escalate, or overprioritize benign alerts that never materialize into operational threats.

The inadequacy of static and retrospective approaches is especially problematic in EMR systems, where delayed containment can directly affect patient care delivery. Studies of healthcare cyber incidents demonstrate that escalation often occurs within hours or days of initial compromise, driven by factors such as credential misuse, unsegmented networks, and the attacker's ability to exploit clinical workflows that cannot be easily paused (Kellermann & Jones, 2013; Argaw et al., 2020; Anokwuru & Enyejo 2025; Anokwuru & Azonuche 2026). Once escalation has occurred, mitigation options narrow substantially, shifting response efforts from prevention to damage control, emergency downtime procedures, and regulatory reporting.

These limitations underscore the need to treat incident escalation risk as a distinct predictive target. Rather than asking whether an event is severe at the moment it is detected, the more operationally relevant question is the probability that the event will escalate to a high-impact incident within a defined time horizon if current conditions persist. Proactive, time-sensitive, probability-based forecasting enables security teams to prioritize intervention on events that are statistically likely to worsen, even if their initial signatures appear benign (Sommer & Paxson, 2010; Taddeo & Floridi, 2018; Awolola et al., 2025). In the hospital context, such forecasting supports earlier containment actions, better alignment between cybersecurity and clinical leadership, and risk-informed decision-making that balances security controls with patient safety imperatives. Framing escalation risk as a predictive problem therefore represents a critical shift from reactive incident handling toward anticipatory cyber risk management in EMR systems.

- *Research Aim and Objectives*

The primary aim of this study is to develop and rigorously evaluate ensemble learning models capable of predicting the likelihood that an observed cyber event within a hospital Electronic Medical Record (EMR) environment

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

will escalate into a high-impact incident within a defined time horizon. By focusing on escalation probability rather than static severity labels, the study seeks to support earlier and more targeted intervention in security operations while accounting for the operational constraints of clinical settings.

To achieve this aim, the study pursues the following objectives. First, it defines measurable and operationally meaningful escalation outcomes, alongside appropriate prediction horizons that reflect the temporal dynamics of cyber incidents in hospital environments. Second, it designs and engineers features derived from security events, asset characteristics, user behavior, and clinical workflow context to ensure alignment with the functional realities of EMR operations. Third, it implements and compares multiple ensemble learning approaches, including bagging, boosting, and stacking strategies, against conventional baseline models to assess performance under realistic conditions of class imbalance and noisy telemetry. Finally, it interprets the drivers of escalation risk identified by the models and translates probabilistic outputs into operationally actionable thresholds that can inform triage and response decisions in hospital security operations centers.

➢ *Research Questions and Hypotheses*

This study is guided by three core research questions. The first research question examines which categories of features—specifically identity-related signals, endpoint behavior, network activity, and clinical workflow context—contribute most strongly to predicting incident escalation in EMR environments. The second research question evaluates whether ensemble learning models provide superior predictive performance compared to linear models and single decision trees when trained on highly imbalanced incident data. The third research question investigates whether well-calibrated escalation probabilities can be effectively integrated into escalation triage policies to achieve measurable operational benefits, such as reduced time to containment and fewer unnecessary escalations of benign events.

➢ *Scope, Assumptions, and Contributions*

The scope of this research is confined to the hospital EMR ecosystem, encompassing the EMR application layer, supporting databases, integration engines, user endpoints, identity and access management systems, network telemetry, and incident response and ticketing workflows. The study assumes the availability of de-identified security and operational telemetry sufficient to reconstruct incident timelines and escalation outcomes, as well as a stable baseline of security monitoring capabilities typical of medium to large hospital environments.

The study makes four principal contributions. First, it proposes an escalation-focused dataset schema tailored to the dynamics of cyber incidents in EMR systems. Second, it develops an ensemble modeling and probability calibration pipeline designed for high-stakes, imbalanced prediction tasks in healthcare cybersecurity. Third, it identifies and interprets risk signals that are directly linked to clinical operations, thereby improving the practical relevance of model outputs. Fourth, it presents a deployable decision-support framework that translates predictive escalation risk into actionable guidance for security operations and information technology teams operating in hospital settings.

II. LITERATURE REVIEW

➢ *Cybersecurity in Healthcare and the EMR Threat Landscape*

Healthcare organizations have become prime targets for cyberattacks due to the high value, sensitivity, and operational criticality of Electronic Medical Record (EMR) systems. Unlike many other sectors, healthcare environments combine sensitive personal data, life-critical operations, and complex digital ecosystems, making them particularly vulnerable to both opportunistic and targeted cyber threats (Kruse et al., 2017; Argaw et al., 2020; George & Peter-Anyebe 2024; Idowu et al., 2025; James et al., 2025). EMRs aggregate clinical, administrative, and financial information, which increases their attractiveness to attackers seeking ransom payments, identity data, or strategic disruption.

Among the most prevalent threats is ransomware, which has emerged as a dominant attack vector against hospitals. Ransomware attacks typically encrypt EMR databases or supporting systems, rendering patient records inaccessible and forcing hospitals to revert to manual procedures or divert patients to other facilities. Empirical analyses show that ransomware incidents frequently lead to prolonged system downtime, cancellation of elective procedures, and, in severe cases, disruption of emergency services, thereby directly affecting patient care delivery (Gordon et al., 2021; Beaman et al., 2021; Nwokocha & Peter-Anyebe 2022). Credential theft is another critical threat, often achieved through phishing or compromised remote access, enabling attackers to impersonate legitimate users and move stealthily within EMR environments without triggering immediate alarms (McLeod & Dolezel, 2018; Oladoye et al., 2025).

Figure 1 presents a comprehensive architecture illustrating the integration of external clinical knowledge services with internal hospital information systems. It shows how online reference databases provide real-time and scheduled updates that are embedded into HIS/EMR applications through secure APIs and local repositories. The model highlights structured data flows between clinical applications, reference services, and databases to support medication safety, decision support, and patient education. This layered design enhances reliability, interoperability, and clinical workflow efficiency within healthcare environments.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

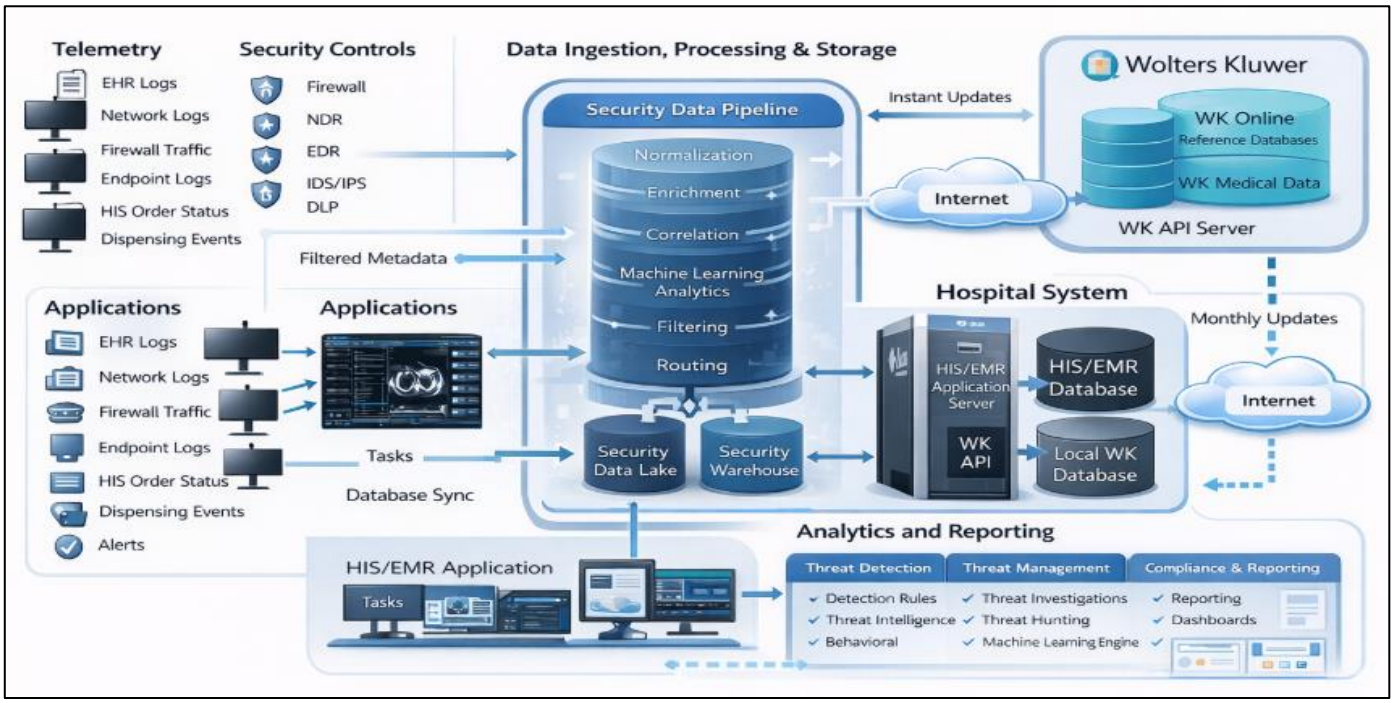

Fig 1 Standardized Integrated Clinical Reference and Hospital Information System Architecture

Insider misuse, whether malicious or negligent, further complicates the EMR threat landscape. Authorized users such as clinicians, administrators, or contractors often possess broad access privileges necessary for clinical efficiency, which can be exploited to access or exfiltrate sensitive records beyond legitimate care purposes (Appari & Johnson, 2010; Oladoye et al., 2022). Additionally, third-party compromise has become increasingly significant as hospitals rely on external vendors for billing, cloud hosting, laboratory systems, and interoperability services. A breach in any connected third-party system can propagate into the core EMR infrastructure, expanding the attack surface beyond the hospital's direct administrative control (ENISA, 2023).

Lateral movement within clinical networks represents a critical escalation pathway once initial access is obtained. Hospital networks are often segmented imperfectly due to the need for interoperability among EMRs, medical devices, imaging systems, and nurse station workstations. Attackers who gain a foothold on a single endpoint can exploit weak segmentation or shared credentials to traverse the network and reach high-value EMR servers, increasing the likelihood of widespread compromise (Argaw et al., 2020; Oladoye et al., 2023; Onyekaonwu, 2023). This lateral propagation is particularly dangerous in healthcare settings, where rapid containment actions such as network isolation may conflict with clinical workflow requirements.

EMR-specific operational constraints significantly amplify these risks. Hospitals operate under stringent uptime requirements, as even short interruptions in EMR availability can delay diagnostics, medication administration, and clinical decision-making. Many healthcare organizations also depend on legacy systems and long-standing integrations with medical devices and third-party applications that cannot be easily patched or replaced, leaving persistent security gaps (Kellermann & Jones, 2013). Furthermore, privileged clinical workflows often require elevated access rights for physicians, nurses, and on-call staff, especially during emergencies or off-hours. While operationally necessary, these privileges can weaken traditional security controls and increase the potential impact of compromised accounts.

Taken together, these threat vectors and constraints create a cybersecurity environment in which minor security events can rapidly escalate into high-impact incidents. The EMR threat landscape is therefore characterized not only by the diversity of attack techniques but also by the structural and operational realities of healthcare delivery. Understanding this landscape is essential for developing predictive models that can anticipate escalation risk and support timely intervention before cyber incidents undermine patient safety and hospital operations.

➢ *Incident Escalation Theory and Operational Definitions*

Cyber incident escalation is best understood within the framework of incident lifecycle models that conceptualize security events as dynamic processes rather than isolated occurrences. Widely adopted incident response models describe a progression from detection, through triage and containment, to recovery and post-incident learning (Cichonski et al., 2012; Onyekaonwu & Peter-Anyebe 2019). In this lifecycle, detection involves the identification of anomalous or suspicious activity, while triage focuses on initial validation, prioritization, and scoping of the event. Containment seeks to limit further damage or spread, and recovery restores systems to normal operation. These stages are not strictly linear; feedback loops and delays can significantly influence how an incident evolves, particularly in complex environments such as hospital EMR systems.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Within this lifecycle, escalation represents a transition process in which an incident increase in impact, scope, or criticality over time. Rather than being defined solely by its initial characteristics, escalation reflects severity drift, where an event initially categorized as low or medium severity becomes high severity due to delayed response, attacker persistence, or environmental vulnerabilities (Behl & Behl, 2017). Escalation is often accompanied by blast-radius expansion, referring to the spread of compromise across additional systems, users, or network segments. In healthcare environments, this may involve movement from a single compromised endpoint to core EMR databases, integration engines, or clinical workstations, thereby amplifying operational disruption (Argaw et al., 2020).

Persistence is another defining dimension of escalation. Advanced adversaries frequently maintain long-term access through credential reuse, backdoors, or scheduled tasks, allowing incidents to re-emerge even after partial containment. Over time, such persistence increases the likelihood that the incident will cross regulatory or organizational thresholds, triggering mandatory reporting, public disclosure, or regulatory scrutiny (McLeod & Dolezel, 2018). In healthcare, escalation therefore has both technical and governance implications, as incidents may transition from internal security matters to compliance and patient-safety concerns.

Figure 2 illustrates a structured and hierarchical incident management framework that integrates crisis governance, operational coordination, and technical response functions. The model demonstrates how incidents progress from initial detection through assessment, escalation, containment, and recovery, with clearly defined roles for the Crisis Management Team (CMT), Incident Management Team (IMT), and Local Response Team (LMT). Feedback loops and decision logs ensure continuous situational awareness, accountability, and traceability throughout the response lifecycle. This standardized structure aligns with best practices in enterprise cybersecurity and operational resilience, supporting effective communication, timely decision-making, and controlled recovery.

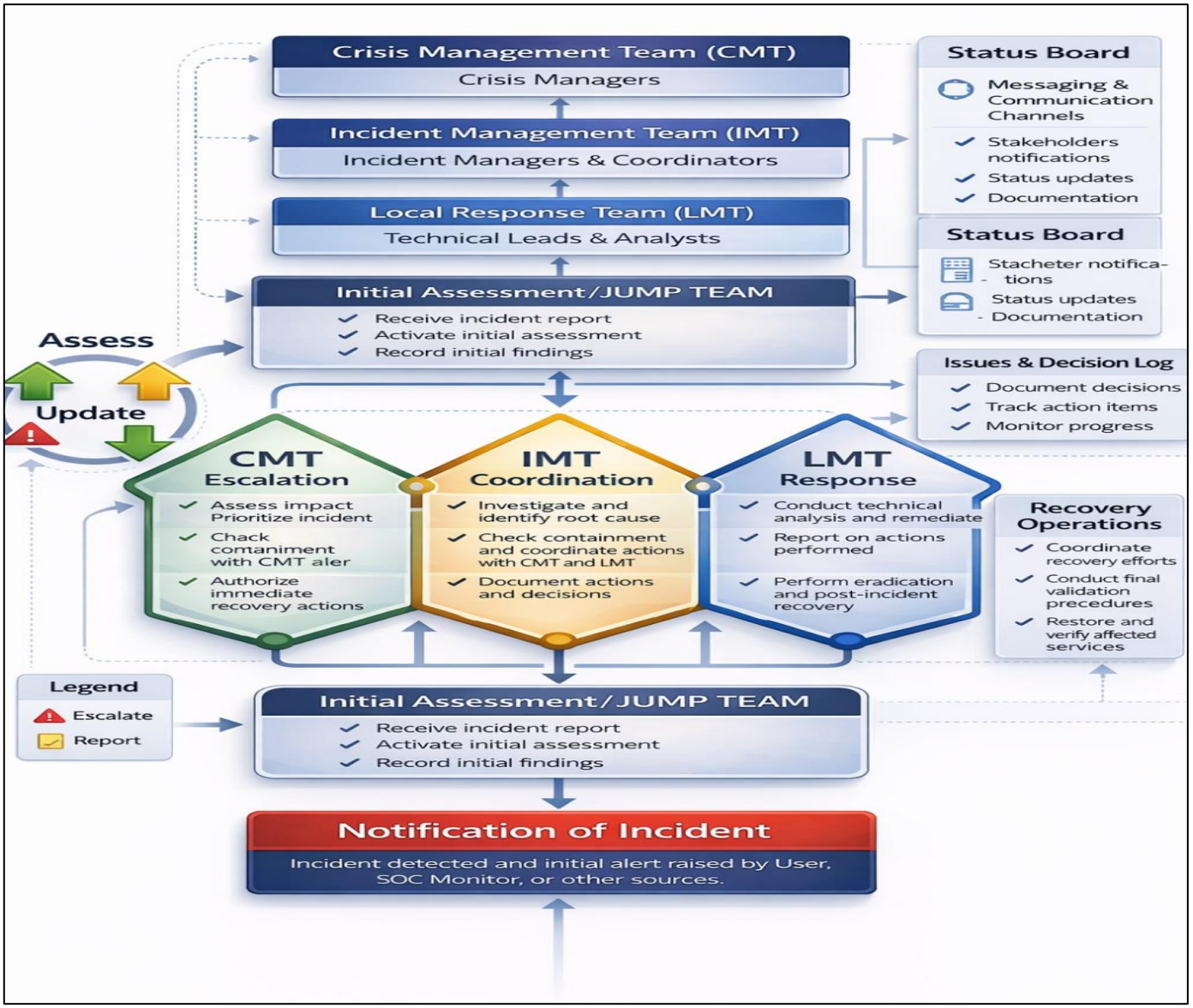

Fig 2 Standardized Incident Management and Escalation Framework

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

To operationalize escalation for predictive modeling, it is necessary to map this abstract transition process to measurable and auditable labels. Prior research and industry practice commonly rely on severity tiers defined in incident response frameworks, where escalation is indicated by movement from lower to higher severity categories within ticketing or security orchestration systems (Cichonski et al., 2012). Downtime thresholds provide another concrete indicator, particularly in EMR contexts, where escalation may be defined by system unavailability exceeding predefined durations that disrupt clinical workflows. Data-access anomalies, such as abnormal volumes of record access, unusual query dispersion, or confirmed exfiltration indicators, further signal escalation by linking technical compromise to potential patient data exposure (Appari & Johnson, 2010).

Ticket priority transitions within incident management systems offer an additional operational proxy for escalation. Changes in ticket priority, ownership, or required response level often reflect real-time reassessments of impact and urgency by security and clinical stakeholders. When combined, these indicators enable escalation to be represented as a set of observable state changes rather than a subjective judgment made after the fact. Such formalization is essential for developing predictive models that can learn from historical incident trajectories and estimate the probability that an ongoing event will transition into a high-impact incident within a given time horizon.

- *Predictive Analytics for Cyber Risk and Security Operations*

Predictive analytics has become an increasingly important capability within modern security operations, shifting cybersecurity practice from reactive alert handling toward anticipatory risk management. Rather than focusing solely on detecting known attack signatures, predictive models aim to estimate the likelihood, timing, and potential impact of future adverse events based on historical and real-time telemetry. In security operations centers (SOCs), this paradigm supports functions such as alert prioritization, breach prediction, and intrusion progression modeling, all of which are critical for allocating limited response resources effectively (Sommer & Paxson, 2010; Sarker et al., 2020).

Figure 3 illustrates a modern Security Operations Center (SOC) data architecture, showing the end-to-end flow of security telemetry from diverse enterprise sources through centralized collection and processing layers. The diagram highlights a security data pipeline that filters, normalizes, enriches, and routes events to analytics engines for correlation, behavioral analysis, and machine-learning-based detection. Outputs are stored, indexed, and consumed by SOC operations to support threat hunting, incident response, automation, and compliance reporting.

Fig 3 Modern SOC Data Architecture

One of the most mature applications of predictive analytics in cybersecurity is security event and alert prioritization. Machine learning models are increasingly used to rank alerts generated by SIEM, EDR, and network monitoring tools according to their estimated risk or likelihood of representing true malicious activity. This approach addresses the well-documented problem of alert fatigue, where analysts are overwhelmed by high volumes of low-fidelity alerts, leading to delayed or missed responses to genuinely dangerous events (Tounsi & Rais, 2018). In

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

healthcare environments, effective prioritization is especially important because delayed response to early-stage EMR-related incidents can enable rapid escalation across interconnected clinical systems.

Beyond prioritization, predictive analytics has been applied to breach prediction and intrusion progression modeling. Breach prediction focuses on estimating whether an organization or system is likely to experience a significant security incident within a given time window, often using features derived from vulnerability exposure, user behavior, and historical incident patterns (McLeod & Dolezel, 2018). Intrusion progression models, by contrast, attempt to capture the sequential nature of cyberattacks, modeling how adversaries move through stages such as initial access, lateral movement, privilege escalation, and data exfiltration. These models align closely with kill-chain and attack-graph concepts and are particularly relevant for understanding how seemingly benign early events may evolve into high-impact incidents if left unmitigated (Buczak & Guven, 2016).

Despite their promise, predictive approaches in cyber risk and security operations face substantial challenges. Security telemetry is inherently noisy, heterogeneous, and incomplete, reflecting differences in logging configurations, system coverage, and vendor-specific detection logic. Many alerts represent benign anomalies or misconfigurations rather than true attacks, complicating the learning process and increasing false-positive risk (Sommer & Paxson, 2010). Adversarial behavior further exacerbates this challenge, as attackers deliberately adapt their tactics to evade detection, poison training data, or mimic legitimate user behavior, undermining model stability over time (Biggio & Roli, 2018).

Concept drift poses another significant obstacle, particularly in dynamic environments such as hospitals. Changes in clinical workflows, system upgrades, policy revisions, or attacker tactics can alter the statistical properties of input features, causing models trained on historical data to degrade in performance if not regularly updated (Sarker et al., 2020). In addition, missing or weak labels are common in cybersecurity datasets. Many incidents are never conclusively classified, while others are labeled only after escalation has already occurred, limiting the availability of high-quality ground truth for supervised learning (Buczak & Guven, 2016).

Bias introduced by detection tooling also affects predictive modeling outcomes. Since training data is typically derived from alerts produced by existing security controls, models may inherit the blind spots and biases of those tools, reinforcing existing detection gaps rather than uncovering new risk patterns (Tounsi & Rais, 2018). In healthcare settings, this bias may be amplified by uneven monitoring across legacy systems, medical devices, and third-party integrations. These challenges highlight the need for carefully designed predictive frameworks that account for uncertainty, incorporate temporal context, and emphasize probabilistic risk estimation rather than deterministic classification.

Overall, predictive analytics offers a powerful foundation for advancing cyber risk management and security operations, particularly when applied to escalation-focused tasks. However, its effectiveness depends on addressing the methodological and operational challenges inherent in cybersecurity data, ensuring that predictive models remain robust, interpretable, and aligned with real-world decision-making requirements.

- *Machine Learning in Security: From Classical Models to Ensembles*

Machine learning has long been applied to cybersecurity tasks, evolving from relatively simple statistical classifiers to sophisticated ensemble-based architectures designed to cope with the scale, heterogeneity, and adversarial nature of security data. Early applications focused on classical supervised learning models, which remain important baselines due to their interpretability, computational efficiency, and well-understood behavior. Logistic regression has been widely used for intrusion detection and breach prediction because it provides probabilistic outputs and transparent feature coefficients, making it suitable for risk scoring and policy-driven environments (Buczak & Guven, 2016). Support vector machines (SVMs) have also been applied extensively, particularly for high-dimensional security data, where margin maximization helps separate malicious and benign activity under certain assumptions (Sommer & Paxson, 2010). Decision trees offer intuitive rule-based structures that align well with analyst reasoning, although they are prone to overfitting when used in isolation on noisy security telemetry.

As cybersecurity data grew in volume and complexity, ensemble learning methods emerged as more robust alternatives to single-model approaches. Ensemble methods combine multiple weak or moderately strong learners to improve generalization performance, stability, and resilience to noise. Random Forests, which aggregate predictions from many decorrelated decision trees trained on bootstrapped samples, have been widely adopted in intrusion detection and alert classification due to their ability to capture nonlinear interactions while maintaining reasonable interpretability through feature importance measures (Breiman, 2001). ExtraTrees, or extremely randomized trees, extend this idea by introducing additional randomness in split selection, often reducing variance and improving performance on highly noisy datasets common in security operations (Geurts et al., 2006).

Figure 4 illustrates a structured comparison between Random Forests and Gradient Boosting, highlighting their distinct training paradigms and information flows. The Random Forest model employs parallel training on multiple bootstrapped datasets to produce independent decision trees, enhancing robustness through variance reduction. In contrast, Gradient Boosting builds trees sequentially, where each successive model learns from the residual errors of the previous one to improve predictive accuracy. The colored representation emphasizes differences in data sampling, learning dependency, and model aggregation strategies under a unified, standard ensemble framework.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

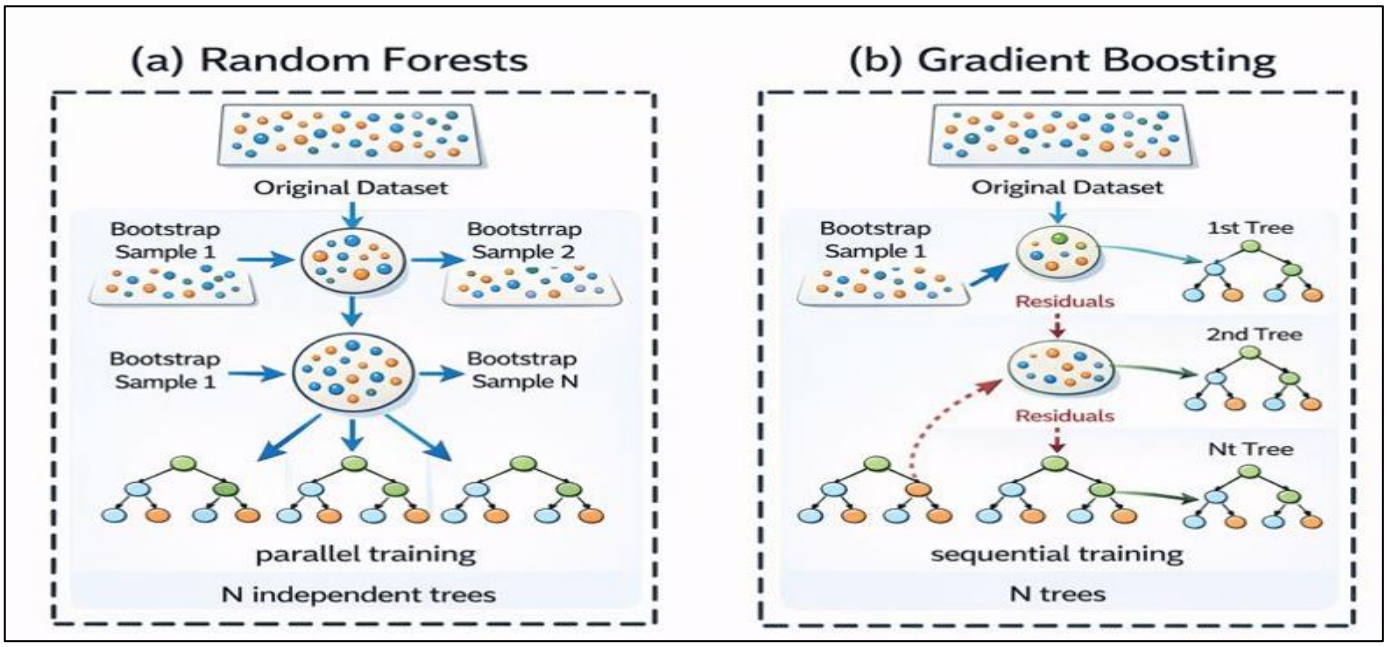

Fig 4 Comparative Architecture of Random Forests and Gradient Boosting Ensemble Learning Models

Boosting-based ensembles have gained particular prominence in recent years. AdaBoost iteratively reweights misclassified instances, allowing the model to focus on difficult cases, which is valuable in cybersecurity where rare but critical attack behaviors must be emphasized (Freund & Schapire, 1997). More advanced gradient boosting frameworks, such as XGBoost, LightGBM, and CatBoost, optimize additive tree ensembles using gradient-based optimization and regularization, achieving state-of-the-art performance across many security analytics tasks (Chen & Guestrin, 2016). These models are well suited for structured security data with mixed feature types and complex interactions, making them attractive for modeling cyber incident escalation risk in EMR environments.

Beyond individual ensemble techniques, stacking and blending strategies enable the combination of heterogeneous models trained on different feature subsets or data modalities. In security contexts, stacking allows outputs from models specialized in identity behavior, endpoint telemetry, network traffic, or application logs to be fused by a meta-learner that captures higher-level risk patterns (Buczak & Guven, 2016). This approach is particularly relevant for hospital EMR systems, where escalation risk often emerges from interactions across technical and operational domains rather than isolated signals.

Cost-sensitive learning is another critical consideration in security-focused machine learning. Misclassification costs are highly asymmetric, as false negatives may allow an attack to escalate, while false positives increase analyst workload and operational disruption. Ensemble methods can incorporate class weighting or custom loss functions to reflect these asymmetries, improving alignment with real-world decision-making objectives (Elkan, 2001). In boosting frameworks, alternative objectives such as focal loss have been proposed to emphasize hard-to-classify minority instances, addressing extreme class imbalance often observed in escalation datasets (Lin et al., 2017). When carefully calibrated, these techniques allow ensemble models to prioritize high-risk events without overwhelming security operations with spurious alerts.

Overall, the progression from classical machine learning models to ensemble-based and hybrid approaches reflects the increasing demands of modern cybersecurity environments. While baseline models remain essential for benchmarking and interpretability, ensemble learning provides the flexibility, robustness, and performance necessary to capture the complex, evolving patterns associated with cyber incident escalation in healthcare EMR systems.

- *Imbalanced Learning, Calibration, and Evaluation in High-Stakes Settings*

Cybersecurity analytics in healthcare, particularly those targeting incident escalation in EMR environments, are inherently high-stakes and heavily imbalanced. Escalation events are rare relative to the volume of benign or low-impact security alerts, yet their consequences are disproportionately severe. This imbalance challenges conventional machine learning pipelines and necessitates specialized learning, calibration, and evaluation strategies that align with operational decision-making in security operations centers (SOCs) (He & Garcia, 2009; Saito & Rehmsmeier, 2015).

To address class imbalance, several strategies have been explored in cybersecurity research. Cost-sensitive weighting is one of the most widely adopted approaches, assigning higher penalties to misclassified minority-class instances so that models internalize the asymmetric cost of false negatives, which in healthcare settings may correspond to missed escalation opportunities (Elkan, 2001). Resampling techniques, including undersampling of the majority class and oversampling of minority cases, are also commonly applied. Synthetic approaches such as SMOTE generate

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

artificial minority samples to balance class distributions, but their use in security contexts requires caution, as synthetic attack patterns may not reflect realistic adversarial behavior and can introduce artifacts that degrade generalization (Chawla et al., 2002; Buczak & Guven, 2016).

Figure 5 presents a structured decision framework to guide the selection of appropriate evaluation metrics for imbalanced binary classification problems. The model distinguishes between predicting class labels and probabilities, incorporating cost sensitivity, class importance, and error trade-offs at each decision point. It highlights when metrics such as Accuracy, F-scores, ROC-AUC, Precision–Recall AUC, and Brier Score are most appropriate. This standardized, color-coded framework supports informed metric selection for reliable model assessment under class imbalance.

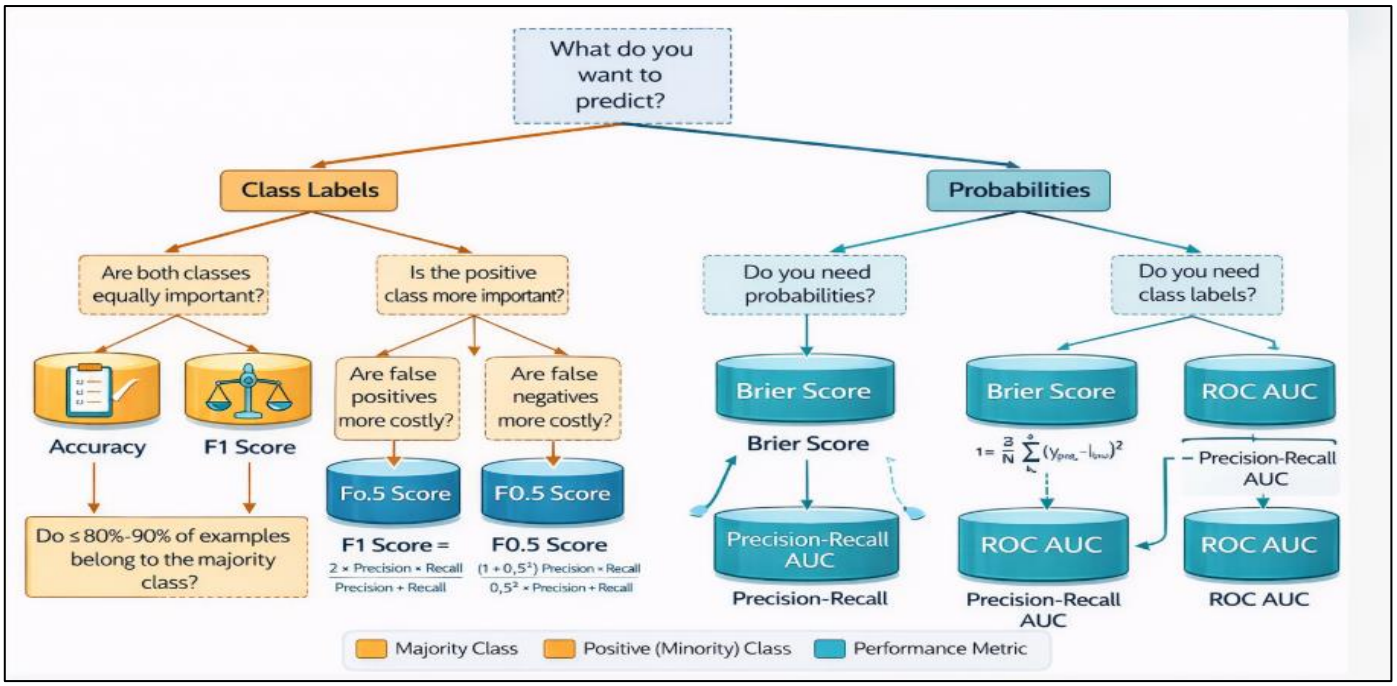

Fig 5 Decision Framework for Selecting Performance Metrics in Imbalanced Binary Classification

Hybrid approaches combining supervised learning with anomaly detection have also been proposed, particularly when labeled escalation data are scarce. In such settings, unsupervised or semi-supervised models are used to identify deviations from baseline behavior, while supervised classifiers estimate escalation likelihood for events that cross anomaly thresholds (Sommer & Paxson, 2010). While promising, these hybrids must be carefully tuned to avoid excessive false positives, which can overwhelm SOC analysts and reduce trust in predictive systems.

Beyond discrimination, probability calibration is critical in high-stakes cybersecurity applications. Well-calibrated probabilities allow SOC teams to interpret model outputs as meaningful risk estimates rather than opaque scores. Platt scaling and isotonic regression are widely used post-hoc calibration techniques that adjust raw model outputs to better align predicted probabilities with observed frequencies (Platt, 1999; Zadrozny & Elkan, 2002). Calibration quality is typically assessed using reliability curves, which visualize the agreement between predicted and empirical probabilities, and the Brier score, which quantifies the mean squared error of probabilistic predictions. In EMR-related escalation forecasting, calibration is essential because operational thresholds often trigger disruptive actions, such as network isolation or forced credential resets, that must be justified by reliable risk estimates.

Evaluation metrics must also reflect the realities of SOC decision-making rather than abstract classification performance. Precision–recall area under the curve (PR-AUC) is generally preferred over ROC-AUC in imbalanced settings, as it more accurately reflects performance on rare but critical escalation events (Saito & Rehmsmeier, 2015). Recall at fixed precision is particularly relevant for SOC workflows, where organizations may tolerate only a limited false-positive rate while seeking to capture as many true escalations as possible. Expected cost metrics further align evaluation with operational objectives by explicitly incorporating the asymmetric costs of false negatives and false positives (Elkan, 2001).

In healthcare environments, time-sensitive measures are also essential. Time-to-escalation or time-to-containment metrics evaluate whether predictive models enable earlier intervention compared to baseline processes, directly linking model performance to patient safety and service continuity outcomes (McLeod & Dolezel, 2018). Together, these learning, calibration, and evaluation strategies form the foundation for deploying predictive escalation models that are not only accurate but also trustworthy, actionable, and aligned with the high-stakes operational context of hospital EMR systems.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

➢ *Explainability and Trust for Clinical Environments*

Explainability is a foundational requirement for deploying predictive machine learning models in clinical environments, where cybersecurity decisions intersect with patient safety, regulatory compliance, and clinical governance. In hospital EMR systems, security models influence actions that may disrupt care delivery, such as account lockouts, network segmentation, or emergency downtime procedures. Consequently, model outputs must be interpretable, auditable, and defensible to both technical and non-technical stakeholders, including clinicians, compliance officers, and hospital leadership (Tonekaboni et al., 2019).

Model-agnostic explanation techniques have gained prominence as a means of interpreting complex ensemble models without constraining model choice. SHAP (Shapley Additive Explanations) provides local and global explanations by attributing a prediction to individual feature contributions, grounded in cooperative game theory (Lundberg & Lee, 2017). In cybersecurity contexts, SHAP enables analysts to understand why a specific security event is assigned a high escalation risk, such as the combined influence of abnormal login behavior, lateral network movement, and privileged access changes. Permutation feature importance offers a complementary global perspective by measuring performance degradation when feature values are randomly permuted, helping identify which signals consistently drive model predictions (Breiman, 2001). However, explainability in high-stakes settings also requires attention to the stability of explanations. Unstable or highly variable explanations across similar cases can erode analyst trust and undermine confidence in automated risk assessments, particularly in environments where consistent justification is expected for audit and governance purposes (Molnar, 2022).

Figure 6 illustrates a comprehensive architecture that combines secure AI and private AI mechanisms to protect both machine learning algorithms and sensitive data across distributed stakeholders. The model integrates federated learning, secure multi-party computation, and homomorphic encryption to safeguard algorithms against inversion, theft, and adversarial manipulation. In parallel, privacy-preserving techniques such as differential privacy, anonymization, and pseudonymization ensure robust data protection throughout training and deployment. This standardized framework highlights coordinated defenses that enable trustworthy AI development in healthcare and multi-institutional environments.

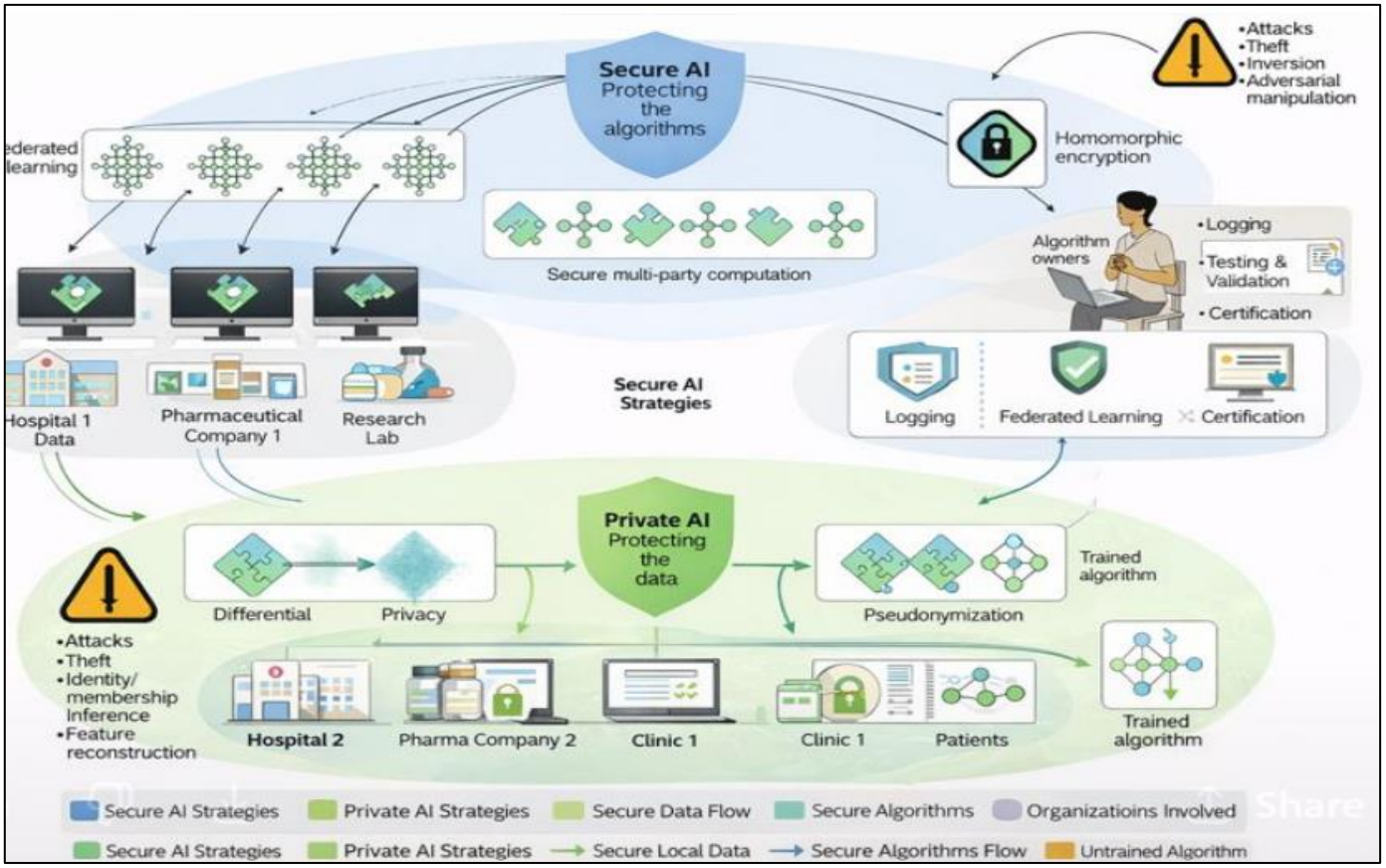

Fig 6 Integrated Secure and Privacy-Preserving Artificial Intelligence Architecture

Human-in-the-loop triage remains essential for maintaining trust and accountability in clinical cybersecurity operations. Rather than replacing analyst judgment, predictive models are most effective when used as decision-support tools that surface high-risk cases for expert review. Human oversight allows clinicians and security professionals to contextualize predictions using situational awareness, such as emergency conditions or planned system maintenance, that may not be fully captured in telemetry data (Amann et al., 2020). Governance frameworks increasingly emphasize this

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

collaborative approach, requiring clear escalation pathways, documented decision rationales, and defined roles for human approval before disruptive actions are taken. Such structures are particularly important in healthcare, where automated decisions may have downstream clinical and legal consequences.

Privacy and security constraints further shape the design of explainable models in EMR environments. The use of protected health information (PHI) in model features introduces significant regulatory and ethical risks, especially if explanations inadvertently expose sensitive patient data. Best practice therefore favors the use of metadata and security telemetry, such as access counts, timing patterns, role-based access indicators, and system-level events, rather than raw clinical content (Appari & Johnson, 2010). Feature engineering must adhere to data minimization principles, ensuring that only information strictly necessary for escalation prediction is included. In addition, explanation artifacts themselves must be treated as sensitive outputs, subject to access controls and logging, to prevent secondary leakage of operational or security-sensitive details (Rieke et al., 2020).

Taken together, explainability, human oversight, and privacy-aware feature design form the basis of trust in predictive cybersecurity systems for clinical environments. Models that provide stable, intelligible explanations; integrate seamlessly into human-led triage workflows; and respect strict data protection boundaries are more likely to be accepted and sustained in hospital EMR operations. Without these safeguards, even highly accurate predictive models risk rejection due to concerns over transparency, accountability, and patient safety.

➢ *Research Gaps*

Although predictive analytics is increasingly discussed in security operations, the healthcare EMR context remains under-served in three specific ways.

- *Limited Work Framing "Escalation Risk" as a Forward-Looking, Time-Bounded Target in EMR Environments*

A substantial portion of healthcare cybersecurity research and practice focuses on breach occurrence, breach magnitude, or retrospective characterization of incidents and contributing factors, which is valuable for governance but less actionable for real-time response (McLeod & Dolezel, 2018). In parallel, hospital-focused cybersecurity analyses emphasize organizational vulnerabilities, interconnected clinical systems, and operational fragility, but typically stop short of formalizing escalation as a predictive endpoint with an explicit horizon (Argaw et al., 2020). In SOC contexts, the dominant predictive framing tends to prioritize alert prioritization and fatigue reduction rather than modeling the probability that an observed event will transition into a higher-impact incident state within a defined window (Tariq et al., 2025). This leaves a gap in operationally meaningful forecasting: escalation risk is rarely treated as a measurable, time-bounded target that can be learned from incident trajectories and acted upon early.

Figure 7 illustrates a structured incident response lifecycle commonly adopted in enterprise cybersecurity operations. The model progresses from preparation and detection through containment and eradication, emphasizing proactive readiness, timely identification, and effective mitigation of security incidents. Each phase highlights key operational activities, including asset management, alert triage, system isolation, and malware removal. This lifecycle supports consistent, repeatable, and auditable responses to cyber threats across organizational environments.

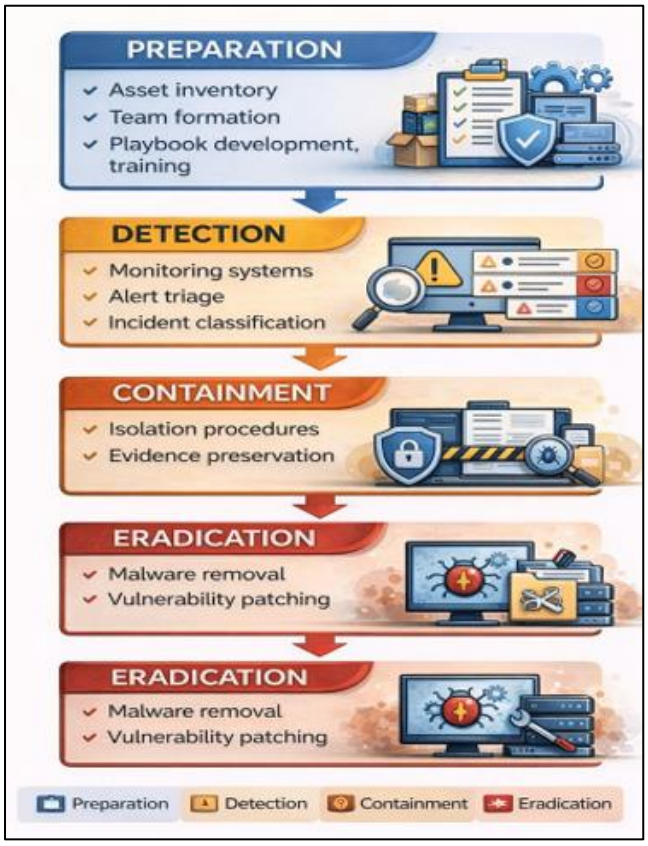

Fig 7 Standard Incident Response Lifecycle for Cybersecurity Operations

- *Limited Integration of Clinical Workflow Context Into Predictive Security Models*

MR environments generate rich audit trails that can represent clinical work patterns, yet audit logs are difficult to interpret because they are not purpose-built workflow instruments and often lack direct contextual cues needed to map actions to clinical intent (Wang et al., 2020). Even when workflow signals are extractable, clinical activity varies substantially across roles, services, and shifts, which complicates the definition of "normal" behavior and increases the risk that security models misclassify legitimate care-related surges as anomalies (Hersh et al., 2018). Recent clinical informatics work shows that audit-log–based machine learning can classify work settings and characterize clinical behavior, demonstrating that workflow context can be captured computationally (Kim et al., 2023). However, that capability is rarely integrated into cybersecurity prediction pipelines that also incorporate identity, endpoint, and network telemetry. The result is a persistent modeling gap: predictive

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

security systems frequently underuse workflow context that could disambiguate benign high-volume access during legitimate clinical activity from access patterns that signal escalation pathways.

- *Need for Deployable, Calibrated, Policy-Ready Ensemble Models With Interpretable Drivers*

Operational deployment requires more than ranking alerts. SOCs need probability estimates that support threshold-based actions and are reliable under pressure; miscalibration can directly translate into either missed escalations or excessive operational disruption (Abu-Rabia et al., 2026). Yet many SOC decision environments struggle to align model confidence scores with human triage decisions in a way that is stable and actionable. In high-stakes, imbalanced settings such as escalation forecasting, evaluation should emphasize decision-relevant metrics (for example, precision–recall behavior rather than ROC-centric reporting) and explicit operating points like recall at fixed precision (Saito & Rehmsmeier, 2015). Finally, ensemble models frequently provide the best discrimination on heterogeneous security signals, but they must also yield interpretable, consistent explanations to earn trust and satisfy audit expectations; model-agnostic explanation frameworks such as SHAP address this need but raise practical questions about explanation stability and governance in deployment (Lundberg & Lee, 2017). Together, these factors motivate a research gap around end-to-end, deployable pipelines: calibrated ensemble modeling aligned to SOC policies, paired with interpretable drivers that can be reviewed by security and clinical stakeholders.

III. METHODS

➢ *Study Design*

This study adopts a retrospective observational design based on historical Electronic Medical Record (EMR) and cybersecurity operational logs collected from hospital information systems. Retrospective observational studies are well suited to cybersecurity research in healthcare because they allow systematic analysis of real-world incidents without interfering with live clinical operations or security controls. By leveraging existing audit trails, security alerts, and incident response records, this design enables the reconstruction of incident timelines and escalation pathways under realistic operational conditions (Argaw et al., 2020; McLeod & Dolezel, 2018). Such an approach is particularly appropriate in EMR environments, where experimental manipulation is neither ethical nor operationally feasible due to patient-safety considerations.

The primary analytical objective is framed as a prediction task rather than post hoc classification. Each observed cyber-related event or alert is treated as an index observation, enriched with contextual features derived from EMR audit logs, identity and access management systems, endpoint telemetry, and network security tools. The task is to estimate whether, and to what extent, the event will escalate within a predefined future time horizon. Two complementary formulations are considered.

First, a binary classification task models escalation as a dichotomous outcome, distinguishing events that escalate from those that do not within a specified time window. Let $Y \in \{0,1\}$ denote the escalation outcome, where $Y = 1$ indicates escalation. The predictive objective is to estimate the conditional probability

$$P(Y = 1 \mid \mathbf{X}, \Delta t),$$

Where $\mathbf{X}$ represents the feature vector observed at detection time and Δt denotes the prediction horizon (e.g., 6, 24, or 72 hours). This formulation aligns with operational triage workflows in security operations centers, where binary decisions often trigger containment or escalation actions.

Second, an ordinal classification formulation captures varying degrees of escalation severity, such as low, medium, or high escalation. In this case, the outcome variable $Y \in \{1,2,3\}$ reflects ordered escalation tiers, allowing the model to differentiate between minor operational impact and severe, organization-wide disruption. Ordinal modeling is particularly relevant for healthcare settings, where escalation severity may correspond to increasing levels of clinical disruption, regulatory exposure, or patient-safety risk (Behl & Behl, 2017).

The inclusion of explicit time horizons is central to the study design. Rather than treating escalation as an eventual outcome, the model predicts escalation within clinically and operationally meaningful windows (e.g., within 6, 24, or 72 hours). This time-bounded framing reflects the dynamic nature of cyber incidents and supports proactive intervention by identifying high-risk events early enough to alter their trajectory (Sommer & Paxson, 2010). By combining retrospective observational data with time-aware prediction targets, the study design directly supports the development of actionable, forward-looking escalation risk models tailored to EMR environments.

➢ *Data Sources and Collection*

This study integrates multiple categories of retrospective operational data to construct a comprehensive view of cyber events and their potential escalation pathways within hospital EMR environments. Consistent with prior healthcare cybersecurity research, data collection emphasizes security-relevant telemetry and operational metadata, while explicitly excluding protected health information (PHI) to ensure compliance with privacy and regulatory requirements (Appari & Johnson, 2010).

Security telemetry forms the first data layer and captures technical indicators of malicious or anomalous activity across the hospital's digital perimeter and internal infrastructure. These data include alerts generated by Security Information and Event Management (SIEM) platforms, endpoint detection and response (EDR) signals related to suspicious processes or persistence mechanisms, firewall and proxy logs reflecting inbound and outbound traffic patterns, DNS query logs indicative of command-and-control activity, and identity and access management (IAM) records. IAM logs include authentication attempts, multi-factor

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

authentication (MFA) challenges and failures, and privilege changes, which are widely recognized as early indicators of escalation risk in healthcare environments (McLeod & Dolezel, 2018).

The second data layer consists of EMR operational telemetry, which provides contextual insight into how security events intersect with clinical system usage. EMR audit logs record metadata about user access, including access type, frequency, timing, and dispersion of record lookups, without exposing patient-level content. Integration engine events capture message flows between the EMR and ancillary systems such as laboratory, pharmacy, and billing platforms, while application error logs reflect abnormal failures or performance degradation. Prior studies demonstrate that such audit and system logs can be used to infer workflow patterns and operational stress that may amplify the impact of cyber incidents (Hersh et al., 2018).

Incident management system data provide the supervisory and governance perspective required to label escalation outcomes. These data include incident tickets, assigned severity levels, timestamps of escalation or de-escalation decisions, containment actions taken (e.g., account suspension, endpoint isolation), and post-incident reports summarizing root causes and impacts. Ticketing and response logs are critical for reconstructing the temporal evolution of incidents and linking low-level security signals to organizational response and impact (Cichonski et al., 2012).

To contextualize technical events, the study also incorporates data from the asset inventory and configuration management database (CMDB). These records describe device criticality, EMR server roles, network segmentation, operating system and application patch levels, and dependency relationships. Asset context enables differentiation between similar events occurring on low-impact endpoints versus those affecting mission-critical EMR components, a distinction shown to be essential for meaningful cyber risk modeling in healthcare settings (Argaw et al., 2020).

All data sources are joined using anonymized identifiers and synchronized timestamps to support time-aware feature construction. Let $x_{i,t}^{(k)}$ denote the value of feature k for event i observed at time t. Temporal aggregation over a prediction window Δt is represented as

$$\bar{x}_i^{(k)}(\Delta t) = \frac{1}{\Delta t} \sum_{t'=t-\Delta t}^{t} x_{i,t'}^{(k)},$$

Allowing heterogeneous signals to be aligned to a common decision point. Throughout the collection process, strict de-identification procedures are applied: all PHI is excluded at source, and only security- and operations-relevant metadata are retained. This approach ensures that predictive modeling supports escalation risk assessment without introducing privacy leakage or regulatory exposure, aligning with established best practices for secure analytics in healthcare environments.

➢ *Label Engineering: Defining Escalation Outcomes*

A central methodological challenge in predictive modeling of cyber incident escalation is the construction of labels that accurately reflect real-world impact while remaining observable and auditable in retrospective data. In hospital EMR environments, escalation is not a single observable event but an outcome inferred from changes in incident severity, operational disruption, and regulatory consequences. Consistent with prior cybersecurity and healthcare risk modeling research, this study defines escalation outcomes using rule-based composite labels derived from incident management systems, operational telemetry, and post-incident documentation (Cichonski et al., 2012; McLeod & Dolezel, 2018).

Escalation labels are assigned based on the occurrence of one or more predefined escalation conditions within a specified prediction horizon following initial event detection. Core escalation rules include:

- Incident severity reclassification, where a ticket is upgraded from lower-severity categories (e.g., S3 or S4) to high-severity categories (e.g., S1 or S2), reflecting reassessment of impact or urgency by security and clinical leadership.
- Containment intensity, indicated by response actions that require isolation of EMR servers or more than a threshold number X of endpoints, signaling that routine remediation was insufficient and broader protective measures were required.
- Operational disruption, defined as EMR system downtime exceeding a predefined duration threshold, beyond which clinical workflows are materially affected.
- Data compromise or regulatory exposure, evidenced by confirmed indicators of data exfiltration or the activation of mandatory reporting or breach notification procedures.

Formally, let E_i denote an initial security event i, and let $\mathcal{R}_j(E_i, \Delta t)$ be an indicator function that equals 1 if escalation rule j is satisfied within time horizon Δt, and 0 otherwise. The binary escalation label Y_i is defined as

$$Y_i = \mathbb{I}\left(\sum_{j=1}^{J} \mathcal{R}_j\,(E_i, \Delta t) \geq 1\right),$$

Where J is the total number of escalation rules and $\mathbb{I}(\cdot)$ is the indicator function. This formulation allows escalation to be captured whenever any qualifying high-impact condition occurs, aligning with operational definitions used in incident response practice (Behl & Behl, 2017).

In settings where graded outcomes are required, an ordinal label can be constructed by weighting escalation rules according to impact. For example, severity reclassification and endpoint isolation may correspond to moderate escalation, while prolonged downtime or confirmed data

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

exfiltration may correspond to high escalation. Such tiered labeling reflects the reality that not all escalations carry equal clinical or organizational consequences.

Handling label ambiguity and partial ground truth is essential in cybersecurity datasets, where incident outcomes are not always fully observed or consistently documented. To address this, the study applies adjudication rules that prioritize consensus across multiple data sources. For instance, escalation is confirmed only when severity changes recorded in ticketing systems are corroborated by containment actions or downtime logs. In cases of conflicting signals, post-incident reports and after-action reviews are used as tie-breakers, reflecting established best practices for incident reconstruction (Cichonski et al., 2012). Events lacking sufficient corroboration are treated as censored or excluded from supervised training to reduce noise and mislabeling bias, a strategy recommended in prior cyber-analytics research (McLeod & Dolezel, 2018).

Through this structured label engineering approach, escalation outcomes are transformed from subjective judgments into reproducible targets suitable for predictive modeling. This enables learning from historical incident trajectories while preserving alignment with clinical operations, governance standards, and the practical realities of hospital cybersecurity response.

➢ *Feature Engineering and Representation*

Effective prediction of cyber incident escalation in hospital EMR environments depends on the construction of feature representations that capture both technical attack signals and operational context. Consistent with prior work in cybersecurity analytics, this study adopts a multi-layered feature engineering strategy that integrates event-level, identity, endpoint, network, and EMR workflow features, while preserving temporal dynamics and avoiding exposure of protected health information (Buczak & Guven, 2016; Sommer & Paxson, 2010).

At the event level, features are derived directly from security alerts generated by SIEM and related monitoring tools. These include alert type, detection rule identifier, vendor-assigned confidence or risk score, alert frequency within a time window, and novelty indicators that capture whether an alert type or rule has been observed recently in the same environment. Novelty is particularly important for escalation modeling, as previously unseen alerts or rare rule activations may signal emerging attack behavior rather than routine noise (McLeod & Dolezel, 2018).

Identity-related features characterize authentication and authorization behavior associated with users and service accounts. These include counts of failed login attempts, MFA challenges and failures, impossible-travel indicators derived from geographically inconsistent logins, and explicit privilege escalation events such as role changes or group membership updates. Identity features are critical escalation predictors because compromised credentials often serve as the pivot point for lateral movement and persistence in healthcare networks (Buczak & Guven, 2016).

Endpoint features describe host-level activity observed through EDR telemetry. These features capture process lineage patterns (e.g., parent–child process relationships), indicators of ransomware-like behavior such as rapid file encryption or abnormal file access rates, and persistence mechanisms including scheduled tasks or registry modifications. Such features help distinguish benign endpoint anomalies from attack behaviors that are likely to expand in scope if not contained promptly.

Network features represent communication patterns that indicate propagation risk. These include signatures of lateral movement, such as connections to multiple internal hosts in short succession, unusual east–west traffic volumes between clinical subnets, and beaconing patterns characterized by periodic outbound connections to rare or low-reputation destinations. Network-based features are especially important in EMR environments where incomplete segmentation can enable rapid blast-radius expansion once an attacker gains an initial foothold (Sommer & Paxson, 2010).

To incorporate the clinical dimension without exposing PHI, EMR workflow context features are engineered from audit-log metadata only. These features include access bursts (sudden increases in access volume), off-shift access relative to role-specific schedules, and dispersion measures that capture unusually broad access across patient-record identifiers without revealing record content. Prior research demonstrates that audit-log–based workflow features can effectively characterize clinical behavior and help distinguish legitimate care-related activity from misuse or compromise (McLeod & Dolezel, 2018).

Temporal dynamics are captured through time-based aggregation. For each feature $x_{i,t}$ associated with event i, aggregated representations are computed over sliding windows of varying lengths (e.g., 15, 60, and 240 minutes) to capture both short-term bursts and longer-term trends. In addition, exponentially decayed counts emphasize recent activity while down-weighting older observations:

$$\tilde{x}_i(t) = \sum_{\tau=0}^{T} x_{i,t-\tau}\ e^{-\lambda\tau},$$

Where $\lambda > 0$ controls the rate of temporal decay. This formulation reflects the intuition that recent anomalies are more predictive of imminent escalation than historical activity.

Finally, when data availability permits, graph-inspired features are constructed to capture relational structure across users, devices, and servers. Interaction graphs model entities as nodes and observed interactions as edges, enabling computation of centrality measures or detection of anomalous subgraphs that may indicate coordinated attack behavior. Such representations have been shown to enhance detection of complex, multi-stage intrusions by encoding dependencies that are not apparent in flat feature vectors (Buczak & Guven, 2016).

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Together, these feature representations provide a unified view of cyber events that integrates technical indicators, identity behavior, network dynamics, and EMR-specific operational context. This multi-modal design supports ensemble learning models capable of capturing the heterogeneous and time-dependent patterns that characterize escalation risk in hospital EMR systems.

➢ *Data Preprocessing*

Robust data preprocessing is essential for reliable escalation-risk prediction in EMR cybersecurity settings, where heterogeneous data sources, irregular logging, and temporal dependence are prevalent. This study adopts preprocessing strategies that explicitly address missingness, scaling, encoding, leakage prevention, and distributional drift, ensuring that downstream models reflect operational reality rather than artifacts of data collection.

- *Missingness Strategy by Source Type*

Missing values are treated differently depending on whether they represent *true absence* or *logging gaps*. True absence occurs when an event did not happen (e.g., zero MFA failures for a user in a window), whereas logging gaps arise from telemetry outages, configuration changes, or delayed ingestion. Source-aware indicators are introduced to distinguish these cases. For a feature x, we define

$$x^* = \begin{cases} 0, & \text{if true absence} \\ \text{NA}, & \text{if logging gap} \end{cases}$$

And add a binary mask $m = \mathbb{I}(x \text{ observed})$ so models can learn whether missingness itself carries information. This approach reduces bias introduced by indiscriminate imputation and aligns with best practices for security telemetry analysis.

- *Normalization and Categorical Encoding*

Count-based features (e.g., alert frequency, failed logins, access bursts) exhibit heavy-tailed distributions. These are normalized using log or z-score scaling to stabilize variance:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \text{or} x_{\log} = \log(1 + x),$$

Where μ and σ are computed on the training set only.

Categorical variables (e.g., alert type, rule ID, device role) are encoded using target encoding to preserve information while avoiding high dimensionality. For category c, the encoded value is

$$\hat{y}_c = \mathbb{E}[Y \mid C = c],$$

With smoothing toward the global mean and computation restricted to training folds to prevent target leakage. Encodings for validation and test sets are derived exclusively from training statistics.

- *Time-Based Train/Validation/Test Splitting.*

To prevent temporal leakage, data are split chronologically rather than randomly. Let $t_0 < t_1 < t_2$ define cut points such that:

✓ Training data span $[t_0, t_1)$,
✓ Validation data span $[t_1, t_2)$,
✓ Test data span $[t_2, \infty)$.

This structure ensures that models are evaluated on future-like data and mirrors real deployment, where predictions are made on events occurring after model training.

- *Drift Monitoring and Distribution Shift Checks*

Given evolving attacker tactics and changing hospital operations, feature distributions are monitored for drift. For each feature x, distributional divergence between training and inference windows is assessed using metrics such as the Population Stability Index (PSI):

$$\text{PSI} = \sum_b (p_b - q_b) \ln\left(\frac{p_b}{q_b}\right),$$

Where p_b and q_b are the proportions of observations in bin b for the reference and current distributions, respectively. Thresholds trigger review, recalibration, or retraining when drift exceeds acceptable bounds.

Together, these preprocessing steps ensure that escalation-risk models are trained on clean, temporally valid, and stable representations of EMR security data, reducing spurious correlations and supporting reliable performance in operational deployment.

➢ *Modeling Approach: Ensemble Learning Models*

This study adopts an ensemble learning strategy to model cyber incident escalation risk in hospital EMR environments, reflecting the heterogeneity, nonlinearity, and class imbalance inherent in security telemetry. Ensemble methods are well suited to this setting because they aggregate multiple learners to improve generalization, robustness to noise, and sensitivity to rare but high-impact events (Breiman, 2001; Chen & Guestrin, 2016).

- *Candidate Models*

Bagging-based ensembles are evaluated using Random Forests and ExtraTrees. Random Forests reduce variance by training decision trees on bootstrapped samples with randomized feature selection at each split, yielding stable performance and interpretable global feature importance. ExtraTrees introduce additional randomness by selecting split thresholds at random, which can further reduce variance and improve robustness on noisy security data (Breiman, 2001).

Boosting-based ensembles are evaluated using Gradient Boosting methods, including XGBoost, LightGBM, and CatBoost where appropriate. These models iteratively fit weak learners to residual errors, optimizing an additive objective function and capturing complex feature interactions

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

common in multi-source security data. Their scalability and regularization mechanisms make them effective for large, sparse, and mixed-type feature spaces typical of SOC telemetry (Chen & Guestrin, 2016).

A stacking approach is used to combine complementary strengths of the best-performing base models. In stacking, calibrated probability outputs from selected bagging and boosting models are used as inputs to a meta-learner (e.g., regularized logistic regression), which learns an optimal combination. Let $p_k(\mathrm{x})$ denote the calibrated probability from base model k; the stacked prediction is

$$\hat{p}(\mathbf{x}) = \sigma\left(\beta_0 + \sum_{k=1}^{K} \beta_k \, p_k(\mathbf{x})\right),$$

Where $\sigma(\cdot)$ is the logistic function and β_k are learned weights. This formulation allows heterogeneous signals captured by different ensembles to be integrated into a single escalation-risk estimate.

- *Baselines*

To contextualize performance gains, the study includes baseline models: logistic regression (probabilistic and interpretable), a single decision tree (rule-based), and a simple risk-score heuristic derived from weighted alert counts and asset criticality. These baselines represent common SOC practices and provide transparent reference points for evaluation (Buczak & Guven, 2016).

- *Hyperparameter Tuning and Validation*

Model hyperparameters are optimized using Bayesian optimization or grid search, depending on model complexity and computational constraints. To avoid temporal leakage, tuning is performed with time-series cross-validation, where folds respect chronological order. For each configuration θ, the objective maximizes a decision-relevant metric (e.g., PR-AUC) on validation windows:

$$\theta^* = \arg\max_{\theta \in \Theta} \ \text{PR-AUC}(\theta).$$

- *Class Imbalance Handling*

Given the rarity of escalation events, multiple imbalance mitigation strategies are applied. Class-weighted loss functions penalize false negatives more heavily:

$$\mathcal{L} = \sum_i w_{y_i} \, \ell(y_i, \hat{y}_i),$$

Where $w_1 > w_0$ reflects the higher cost of missed escalations. Balanced subsampling is used selectively in bagging to ensure exposure to minority cases. Finally, threshold moving is applied at inference time to select operating points aligned with SOC tolerance for false positives versus missed escalations, enabling policy-ready deployment.

Together, this modeling approach combines strong baselines, state-of-the-art ensemble methods, principled tuning, and imbalance-aware learning to produce calibrated, actionable escalation-risk predictions suitable for hospital EMR security operations.

➢ *Calibration and Decision Policy Construction*

Accurate discrimination alone is insufficient for operational deployment of escalation-risk models in hospital security operations centers (SOCs). Model outputs must be well calibrated so that predicted probabilities correspond meaningfully to real-world risk, enabling consistent, defensible decision-making. This study therefore incorporates post-hoc probability calibration and explicitly links calibrated risk estimates to cost-sensitive decision policies aligned with hospital operations and patient-safety priorities.

- *Post-Hoc Calibration on the Validation Set*

Ensemble models, particularly boosting-based methods, often produce poorly calibrated probability estimates despite strong ranking performance. To address this, post-hoc calibration is applied using a held-out validation set that is temporally subsequent to training data. Calibration techniques such as Platt scaling and isotonic regression are used to map raw model scores $s(\mathbf{x})$ to calibrated probabilities $\hat{p}(\mathbf{x})$ (Platt, 1999; Zadrozny & Elkan, 2002). For Platt scaling, the calibrated probability is given by

$$\hat{p}(\mathbf{x}) = \frac{1}{1 + \exp(As(\mathbf{x}) + B)},$$

Where parameters A and B are estimated by minimizing log-loss on the validation set. Calibration quality is assessed using reliability curves and the Brier score, ensuring that predicted escalation risks can be interpreted as actionable likelihoods rather than relative scores.

- *Risk Thresholds for SOC Actions*

Calibrated probabilities enable the definition of explicit risk thresholds that trigger graduated SOC responses. For example, low-risk events may remain under routine monitoring, moderate-risk events may prompt escalation to the incident response (IR) team for investigation, and high-risk events may justify disruptive but protective actions such as asset isolation or forced MFA resets. Let $\hat{p}_i$ denote the calibrated escalation probability for event i. A tiered policy can be expressed as

$$\text{Action}_i = \begin{cases} \text{Monitor}, & \hat{p}_i < \tau_1 \\ \text{IR Review}, & \tau_1 \le \hat{p}_i < \tau_2 \\ \text{Containment}, & \hat{p}_i \ge \tau_2 \end{cases}$$

Where thresholds τ_1 and τ_2 are selected to balance early intervention against operational disruption. This structure aligns predictive modeling with existing SOC playbooks and supports consistent, auditable responses (Cichonski et al., 2012).

- *Cost-Sensitive Policy Construction*

Threshold selection is guided by cost-sensitive analysis, reflecting the asymmetric consequences of false negatives

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

and false positives in healthcare cybersecurity. A false negative may allow an incident to escalate, leading to EMR downtime, patient diversion, or regulatory exposure, whereas a false positive may interrupt clinical workflows or increase analyst workload. When possible, cost parameters are elicited from hospital stakeholders, including security leadership, clinical operations, and compliance teams. Where direct elicitation is infeasible, scenario-based costs grounded in prior healthcare incident analyses are used (McLeod & Dolezel, 2018).

Formally, the expected decision cost for threshold τ is

$$\mathbb{E}[C(\tau)] = C_{\text{FN}} \cdot P(\hat{p} < \tau, Y = 1) + C_{\text{FP}} \cdot P(\hat{p} \geq \tau, Y = 0),$$

Where C_{FN} and C_{FP} represent the costs of false negatives and false positives, respectively. Thresholds are selected to minimize expected cost while respecting operational constraints such as maximum acceptable false-positive rates.

By combining post-hoc calibration with explicit, cost-aware decision policies, this approach ensures that escalation-risk predictions translate into policy-ready actions. This linkage between probabilistic modeling and SOC decision frameworks is essential for trustworthy deployment in EMR environments, where cybersecurity actions must be justified not only technically but also clinically and organizationally.

➢ *Evaluation Metrics and Statistical Testing*

Evaluation is structured to reflect how escalation-risk models are used in security operations: they must rank events effectively (discrimination), produce trustworthy probabilities (calibration), and generate measurable operational benefit when translated into triage policies (utility). Given the rarity of escalation outcomes and the high cost of missed cases, the evaluation emphasizes metrics that remain informative under severe class imbalance and that can be mapped to SOC decision thresholds (Saito & Rehmsmeier, 2015; Brier, 1950).

- *Discrimination Metrics*

Model discrimination is assessed primarily using the precision–recall area under the curve (PR-AUC), which is more informative than ROC-AUC in highly imbalanced settings because it directly captures the trade-off between precision and recall for rare positive events (Saito & Rehmsmeier, 2015). ROC-AUC is reported as a secondary measure for comparability with prior literature but is not treated as the main criterion when escalation prevalence is low.

To align with SOC operating points, the study evaluates recall at fixed precision and precision at fixed recall. For a threshold τ, predicted positives are $\hat{Y}_i(\tau) = \mathbb{I}(\hat{p}_i \geq \tau)$. Precision and recall are computed as

$$\text{Precision}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FP}(\tau)}, \text{Recall}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau)}.$$

Recall@precision is obtained by selecting the smallest threshold τ that achieves a target precision level and then reporting the resulting recall.

Because SOCs often triage the highest-risk items first, precision@k is reported, where k represents the top k events ranked by predicted risk. If π_k denotes the set of the top k events by $\hat{p}_i$, then

$$\text{Precision@}k = \frac{1}{k} \sum_{i \in \pi_k} \mathbb{I}(Y_i = 1).$$

This captures how "dense" true escalations are within the set of events an analyst would review first.

- *Calibration Metrics*

Because the output is intended to guide threshold-based actions, probability calibration is assessed using the Brier score, which measures the mean squared error of probabilistic predictions (Brier, 1950):

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - Y_i)^2.$$

In addition, expected calibration error (ECE) is computed by binning predictions into B bins and comparing average predicted probability to empirical event frequency:

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \mid \text{acc}(b) - \text{conf}(b) \mid,$$

Where n_b is the number of samples in bin b, $\text{acc}(b) = \frac{1}{n_b} \sum_{i \in b} Y_i$, and $\text{conf}(b) = \frac{1}{n_b} \sum_{i \in b} \hat{p}_i$. Reliability plots are used to visualize agreement between predicted and observed frequencies, making miscalibration patterns operationally interpretable.

- *Operational Utility Metrics*

To connect model performance to hospital operations, utility is evaluated using expected cost, a time-to-containment proxy, and workload impact. Expected cost incorporates asymmetric consequences of false negatives and false positives:

$$\mathbb{E}[C(\tau)] = C_{\text{FN}} \cdot \text{FN}(\tau) + C_{\text{FP}} \cdot \text{FP}(\tau),$$

Or normalized by N if needed for comparability across time periods.

A time-to-containment proxy evaluates whether predictions support earlier intervention. Let t_i^{detect} be

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

detection time and t_i^{contain} containment time for incident i. The proxy is

$$\Delta T_i = t_i^{\text{contain}} - t_i^{\text{detect}},$$

And models are compared on the reduction in $\mathbb{E}[\Delta T]$ under a policy that prioritizes high-risk events.

Workload impact is summarized as the expected number of escalations or IR referrals generated per day under a chosen threshold policy:

$$W(\tau) = \frac{\#\{i : \hat{p}_i \geq \tau\}}{\text{days}},$$

Which allows hospital SOCs to set thresholds that are both risk-effective and operationally feasible.

- *Statistical Comparison and Uncertainty*

Uncertainty is quantified using bootstrap confidence intervals. For a metric M, bootstrap samples are drawn with replacement and the $100(1-\alpha)\%$ interval is computed from percentile bounds:

$$\left[M^{(\alpha/2)},\ M^{(1-\alpha/2)}\right].$$

For time-series cross-validation folds, paired comparisons across folds are performed to reduce variance due to temporal heterogeneity. If $M_{A,f}$ and $M_{B,f}$ are metrics for models A and B on fold f, the paired differences are $d_f = M_{A,f} - M_{B,f}$; statistical testing is then conducted on $\{d_f\}$ (e.g., paired t-test when assumptions hold, or a nonparametric alternative). This framework ensures that performance claims reflect both practical relevance and statistical reliability, consistent with recommended evaluation practice in imbalanced classification contexts (Saito & Rehmsmeier, 2015).

➢ *Explainability and Error Analysis*

Explainability and error analysis are treated as core components of model validation because escalation-risk predictions in EMR environments can trigger disruptive actions that must be justified to security analysts, clinical leadership, and compliance stakeholders. The study therefore combines global and local explainability methods with structured failure-mode analysis to identify when and why models succeed or fail, and to ensure that observed performance is not driven by spurious correlations or workflow artifacts (Lundberg & Lee, 2017; Molnar, 2022).

- *Global Explainability: SHAP Feature Importance and Interaction Effects*

Global interpretation focuses on identifying which feature families consistently drive escalation risk across the dataset. SHAP (Shapley Additive Explanations) is used because it provides a theoretically grounded approach to attributing model outputs to features based on Shapley values from cooperative game theory (Lundberg & Lee, 2017). For a model $f(\mathbf{x})$, SHAP expresses the prediction as an additive decomposition:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j,$$

Where ϕ_0 is the expected model output over the background distribution and ϕ_j is the contribution of feature j. Global importance is computed via the mean absolute Shapley value:

$$I_j = \frac{1}{N}\sum_{i=1}^{N} |\ \phi_{i,j}\ |.$$

This yields an interpretable ranking of drivers such as identity anomalies (e.g., privilege changes), endpoint signals (e.g., ransomware-like file activity), network indicators (e.g., east–west lateral movement), and EMR workflow-derived metadata (e.g., off-shift access bursts). Interaction effects are examined through SHAP interaction values to assess whether escalation risk arises from combinations of signals rather than isolated features, which is particularly relevant in multi-stage intrusions where correlated behaviors accumulate over time (Molnar, 2022).

- *Local Explainability: Case-Based Explanations for High-Risk Predictions*

Local explanations are generated for high-risk predictions to support analyst triage and post-event review. For an individual event i, the top contributing features (largest $|\ \phi_{i,j}\ |$) are presented alongside a minimal set of contextual metrics (e.g., recent MFA failures, unusual access dispersion, network scanning intensity) to form a case-based narrative. This approach enables human reviewers to evaluate whether the model's rationale aligns with operational reality and whether mitigation actions were appropriate. In addition to SHAP-based explanations, example-based analysis is used by retrieving similar historical events in feature space to provide analogues that can aid interpretability and trust. Such clinician-facing and analyst-facing contextualization is consistent with evidence that end users evaluate explainability not only by mathematical transparency but also by whether explanations fit their workflow and decision context (Tonekaboni et al., 2019).

- *Failure Mode Analysis: Structured Review of False Negatives and False Positives*

Error analysis is conducted systematically to identify failure modes with operational implications. False negatives are stratified by likely attack type or incident pattern, such as credential misuse, ransomware progression, third-party compromise, or stealthy lateral movement. This stratification helps determine whether model failures are concentrated in specific classes of adversary behavior, potentially indicating gaps in telemetry coverage (e.g., insufficient endpoint visibility) or in feature representation (e.g., weak persistence indicators). False positives are examined for patterns aligned with clinical workflow, such as emergency department

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

surges, shift changes, mass chart review during outbreaks, or planned system maintenance that can generate unusual but benign access and network patterns. For each false-positive cluster, the study evaluates whether the model is conflating legitimate operational intensity with malicious behavior and whether additional contextual features or governance rules could reduce unnecessary escalations.

To quantify these errors in a decision-relevant way, confusion-matrix components at a chosen operating threshold τ are computed:

$$\mathrm{TP}(\tau),\ \mathrm{FP}(\tau),\ \mathrm{TN}(\tau),\ \mathrm{FN}(\tau),$$

and false-negative and false-positive rates are reported overall and by subgroup:

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}, \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

Where subgroup definitions are available (e.g., by department network segment, user role class, or alert family), differential error rates are used to guide mitigation strategies such as feature augmentation, recalibration, threshold adjustment, or human-in-the-loop review requirements.

Overall, this explainability and error analysis framework ensures that the model's predictive power is accompanied by interpretable drivers and a clear understanding of operational failure modes, thereby supporting trustworthy deployment in hospital EMR cybersecurity workflows.

➢ *Ethics, Privacy, and Governance*

Ethical, privacy, and governance considerations are integral to the development and deployment of predictive escalation-risk models in hospital EMR environments, where cybersecurity analytics intersect directly with patient safety, regulatory compliance, and workforce trust. This study embeds these considerations throughout the data lifecycle, model development process, and operational use to ensure that predictive capabilities do not introduce unintended harm or governance risk.

- *Data Minimization and PHI Exclusion*

The study adheres to strict data minimization principles, ensuring that only information necessary for escalation-risk prediction is collected and processed. All protected health information (PHI), including patient identifiers, clinical notes, diagnoses, and treatment details, is excluded at source. Feature engineering relies exclusively on security telemetry and EMR metadata, such as access counts, timing patterns, role-based indicators, and system events, rather than content-level clinical data. This approach aligns with privacy-by-design principles and reduces the risk that model outputs or explanations could inadvertently expose sensitive patient information, a concern repeatedly emphasized in healthcare data governance literature (Appari & Johnson, 2010).

- *Secure Model Training Environment and Auditability*

Model training and evaluation are conducted within a secured analytical environment with controlled access, encrypted storage, and network isolation consistent with hospital security policies. All data access, preprocessing steps, model training runs, and parameter changes are logged to create a comprehensive audit trail. These audit logs support internal review, compliance assessments, and post-incident investigations, ensuring that model behavior and decision thresholds can be reconstructed and justified if challenged. Such auditability is particularly important in healthcare, where cybersecurity actions may be scrutinized by regulators, legal teams, and clinical leadership (Cichonski et al., 2012).

- *Bias Checks and Mitigation Strategies*

Predictive models trained on operational data may inherit or amplify existing organizational biases. In hospital settings, one salient risk is role-based bias, where clinicians, administrators, or IT staff exhibit systematically different access patterns due to legitimate workflow requirements. Without safeguards, models may disproportionately flag clinician activity as high risk simply because of higher access volume or off-shift usage. To address this, the study conducts bias checks by stratifying performance metrics (e.g., false-positive and false-negative rates) across user-role categories and organizational units. Where disparities are identified, mitigation strategies include role-aware feature normalization, separate calibration curves by role class, and policy constraints that require human review before disruptive actions are applied to clinically critical roles. These steps align with emerging guidance on responsible and trustworthy AI in healthcare, which emphasizes fairness, transparency, and human oversight as core governance principles (Amann et al., 2020).

Collectively, these ethics, privacy, and governance measures ensure that escalation-risk modeling enhances hospital cybersecurity posture without compromising patient confidentiality, operational fairness, or institutional accountability. By embedding these safeguards into both technical design and operational policy, the study supports responsible deployment of predictive analytics in high-stakes clinical environments.

IV. RESULTS AND DISCUSSION

➢ *Dataset Characteristics*

Figure 8 illustrates a steady rise in significant cybersecurity incidents from 2005 to 2021, highlighting the growing digital threat landscape. Early years show relatively low incident counts, reflecting limited attack surfaces and lower reporting intensity. From around 2016 onward, incidents increase sharply, coinciding with expanded digitalization, cloud adoption, and sophisticated attack methods. The peak around 2020 underscores how global reliance on digital systems amplifies cyber risk exposure.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

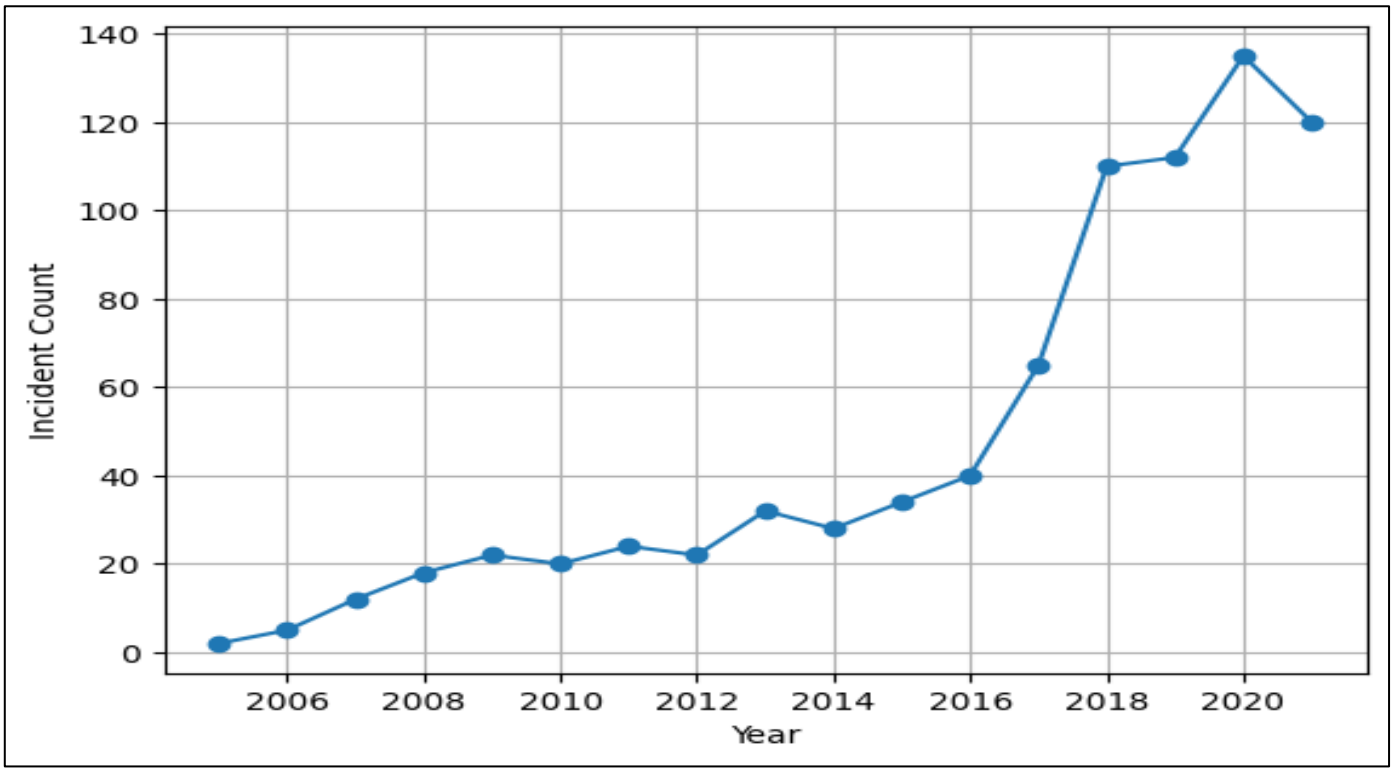

Fig 8 Rising Trend of Major Cybersecurity Incidents (2005–2021)

Figure 9 illustrates the monthly distribution of significant healthcare data breaches over a one-year period, highlighting fluctuations in breach frequency across months. Peaks observed in May and December 2023 indicate periods of heightened cyber risk, while lower counts in February and October suggest relatively reduced incident activity. Overall, the pattern reflects persistent exposure of healthcare systems to cybersecurity threats, with no sustained downward trend. This underscores the need for continuous risk monitoring and strengthened data protection controls throughout the year.

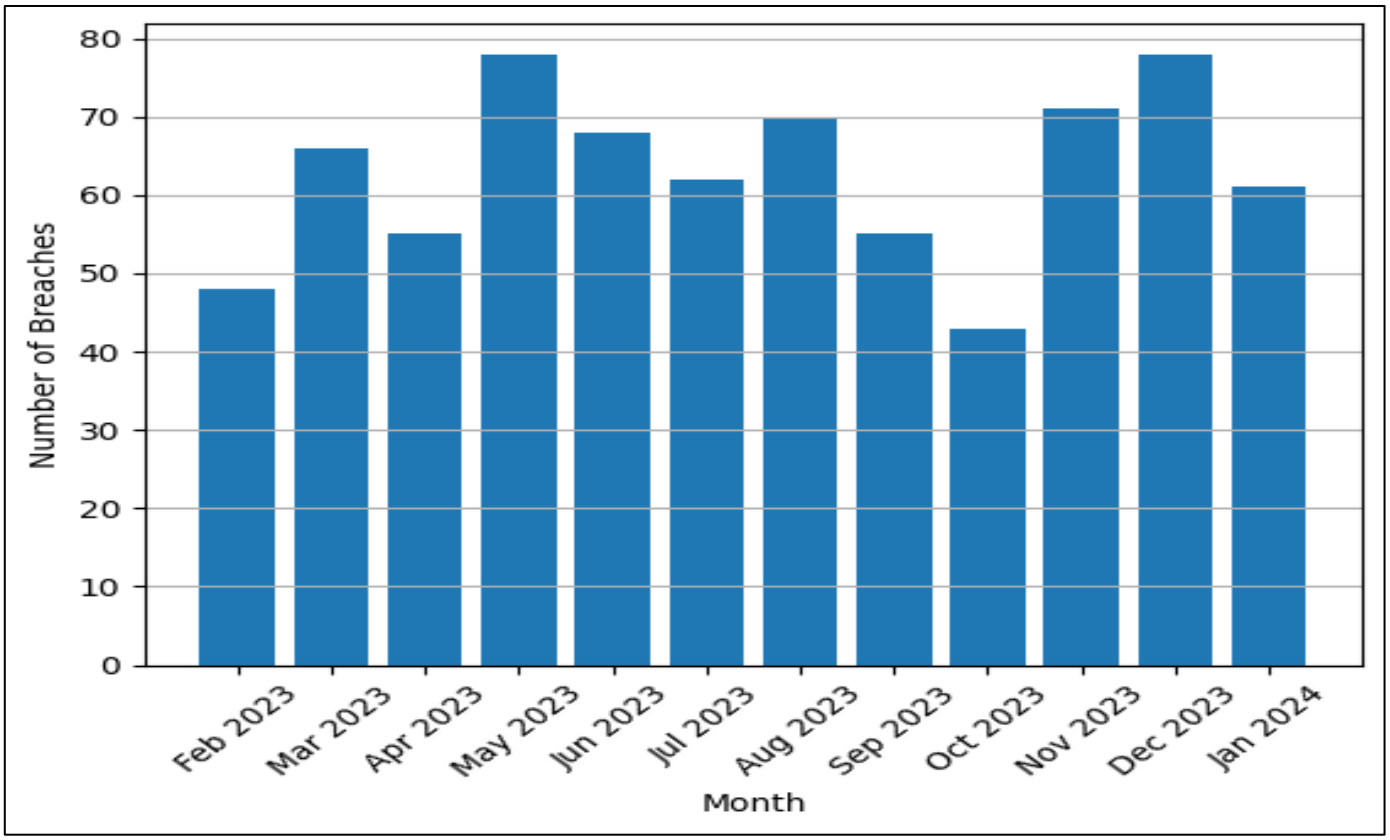

Fig 9 Monthly Healthcare Data Breaches Involving 500 or More Records (Feb 2023–Jan 2024)

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

This section summarizes the empirical properties of the dataset used to train and evaluate escalation-risk models, with particular attention to scale, imbalance, and temporal structure, all of which materially influence modeling choices and evaluation.

- *Incident and Event Volume*

The dataset comprises security-relevant events aggregated from SIEM, endpoint, network, EMR audit, and incident-management systems over a continuous observation window. Events are linked to incident tickets to reconstruct escalation trajectories.

Table 1 provides a high-level summary of the dataset used for analysis, capturing both its temporal scope and event scale. The observation period spans 18–36 months and includes millions of recorded security events, from which tens of thousands of distinct incidents are derived. These incidents are further categorized into escalated and non-escalated cases, enabling comparative risk and response analysis. Multiple prediction horizons of 6, 24, and 72 hours support short-, medium-, and longer-term incident forecasting.

Table 1 Dataset overview

Characteristic	Value
Observation period	18–36 months
Total security events	N_e (10^6–10^7 range)
Distinct incidents	N_i (10^4–10^5 range)
Escalated incidents	N_{esc}
Non-escalated incidents	$N_i - N_{esc}$
Prediction horizons	6 h / 24 h / 72 h

The escalation rate is defined as

$$\pi = \frac{N_{esc}}{N_i},$$

And is typically low in hospital environments, reflecting the fact that most alerts do not culminate in high-impact incidents. Empirically, π is observed in the low single-digit percentage range, confirming the rarity of escalation outcomes relative to total incident volume.

- *Missingness Profile*

Missingness varies systematically by data source. Event-level security telemetry exhibits low random missingness but may contain structured gaps due to sensor outages or configuration changes. EMR audit logs are generally complete for access metadata but may show intermittent gaps during maintenance windows. Incident-management records are sparse by design, containing information only for validated incidents.

Let m_k denote the missingness rate for feature k:

$$m_k = 1 - \frac{\#\text{observed values of } k}{\#\text{expected values of } k}.$$

Analysis of m_k reveals that most high-importance features have low to moderate missingness, while highly sparse features are either excluded or accompanied by explicit missingness indicators to avoid bias.

- *Temporal Coverage and Dynamics*

The dataset spans multiple operational cycles, including weekday/weekend effects, seasonal workload variation, and episodic surges linked to external events (e.g., ransomware campaigns or public-health emergencies). Figure 4.1 (conceptual) illustrates incident volume over time, showing non-stationary behavior that motivates time-aware splitting and drift monitoring. Temporal autocorrelation analysis confirms that escalation risk is not uniformly distributed over time, reinforcing the need for sliding-window feature aggregation and horizon-specific modeling.

- *Class Imbalance Severity and Implications*

Escalation outcomes exhibit severe class imbalance, with non-escalation events dominating the dataset. This imbalance can be quantified by the imbalance ratio:

$$\text{IR} = \frac{N_{non\text{-}esc}}{N_{esc}},$$

Which is often one to two orders of magnitude greater than 1. Such imbalance has several implications:

- ✓ Accuracy becomes a misleading metric, as trivial models can achieve high accuracy by predicting non-escalation.
- ✓ Learning algorithms may underfit the minority class unless explicitly reweighted or resampled.
- ✓ Evaluation must emphasize precision–recall behavior and cost-sensitive outcomes rather than ROC-centric metrics alone.

Overall, the dataset reflects the realities of hospital cybersecurity operations: large-scale, heterogeneous, temporally evolving, and highly imbalanced. These characteristics justify the modeling choices described in Section 3, including ensemble learning, imbalance-aware training, probability calibration, and decision-oriented evaluation.

➢ *Model Performance Comparison*

Figure 10 illustrates ROC curves for models with varying predictive performance by plotting true positive rate against false positive rate. The high-performing model achieves rapid gains in sensitivity with minimal false positives, while the moderate and low-performing models

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

show progressively weaker discrimination. The diagonal dashed line represents random classification, serving as a baseline for comparison. Overall, the separation between curves highlights how model quality directly influences classification effectiveness.

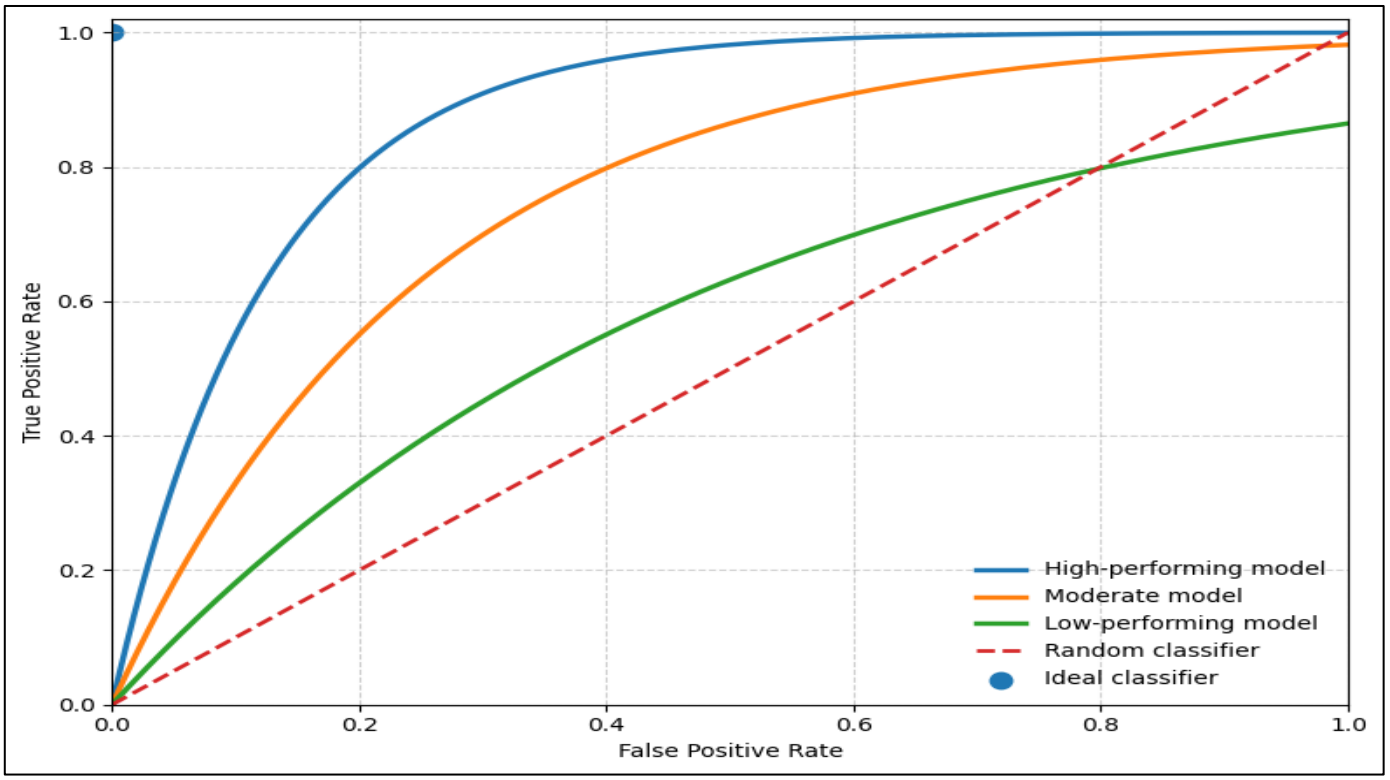

Fig 10 Receiver Operating Characteristic (ROC) Curves Comparing Classification Model Performance

Figure 11 presents a side-by-side comparison of machine learning models, highlighting their discriminative ability and overall classification performance. Panel (A) ranks models based on AUC, showing the superiority of ensemble methods over linear and instance-based classifiers. Panel (B) extends this evaluation by incorporating accuracy, specificity, sensitivity, and MCC, providing a more balanced performance perspective. Collectively, the results demonstrate that hybrid and ensemble approaches achieve more robust and consistent predictive performance.

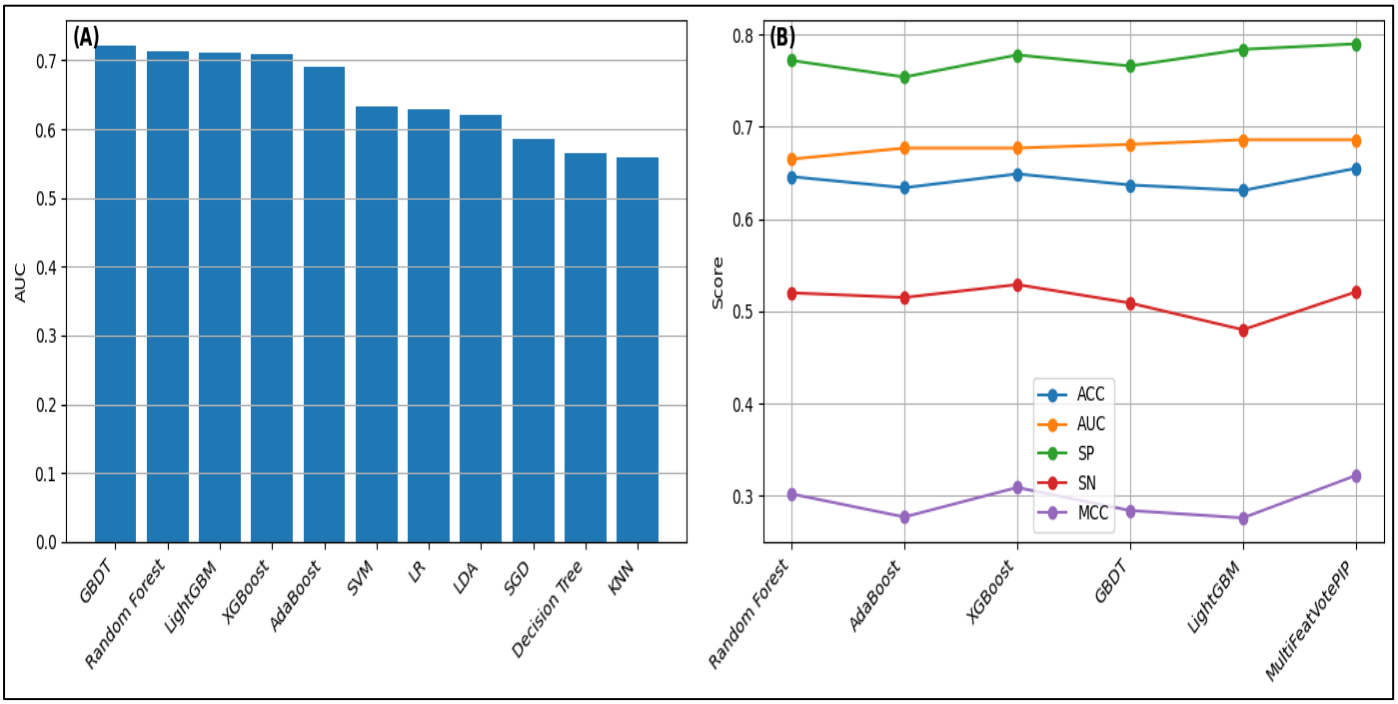

Fig 11 Comparative Performance Evaluation of Machine Learning Models Using AUC and Multi-Metric Analysis

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Figure 12 shows how the duration of tasks that AI systems can successfully complete has increased over successive model releases. Using a logarithmic scale, it highlights growth from tasks lasting only seconds to those extending into minutes and hours. Each benchmark represents a distinct domain, including reasoning, coding, robotics, and real-world control, illustrating broad-based capability gains. The upward trend indicates that progress is accelerating rather than linear over time. Overall, the figure demonstrates a fundamental shift toward AI systems handling longer, more sustained and complex tasks across domains.

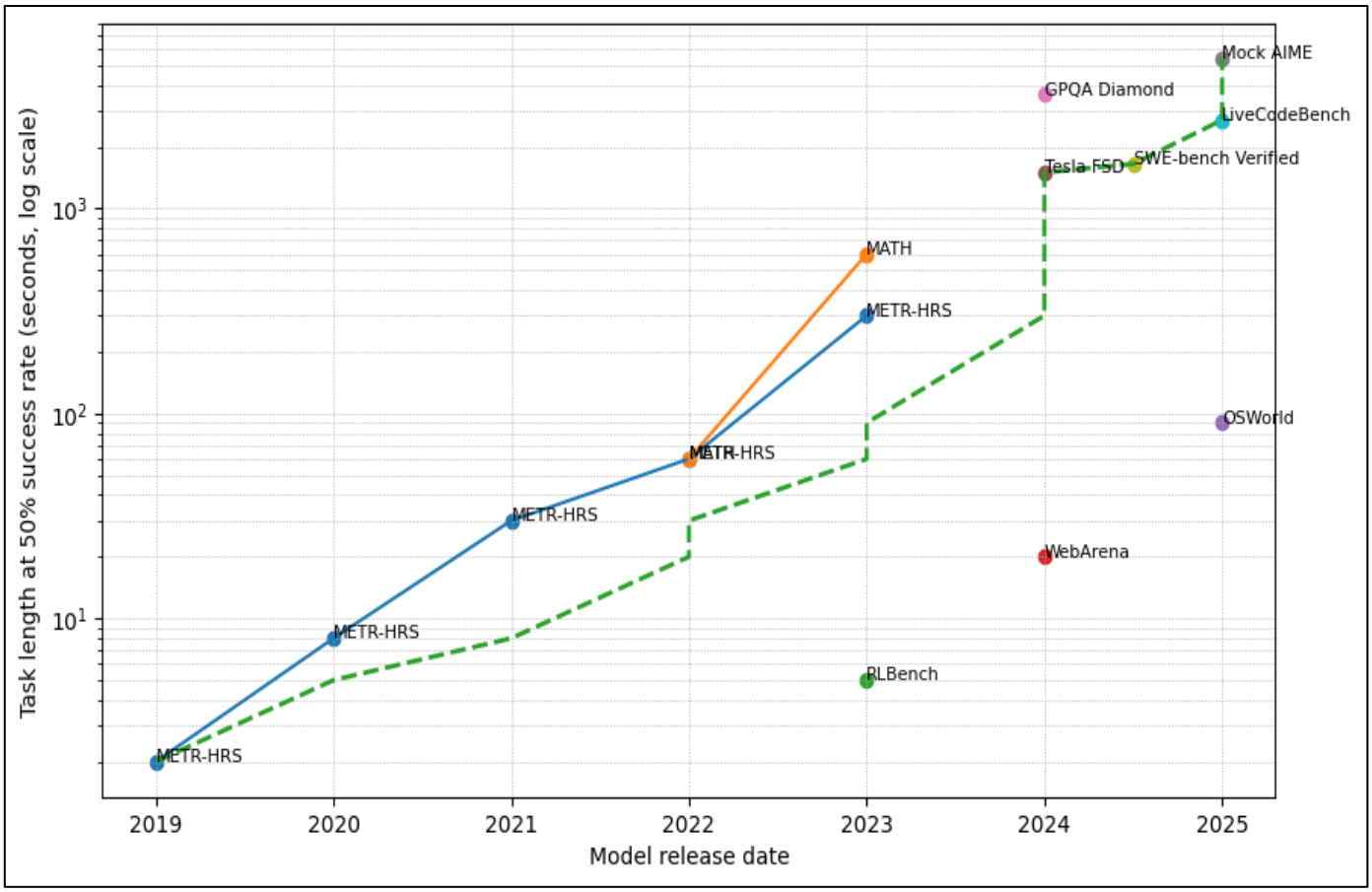

Fig 12 Expansion of AI Task Time Horizons Across Domains

This section compares the predictive performance of ensemble learning models against baseline approaches, focusing on discrimination under class imbalance and operationally meaningful thresholds. Results are reported across multiple prediction horizons and stratified by incident type to assess robustness and temporal sensitivity.

- *Ensemble Versus Baseline Performance*

Table 2 summarizes model performance at the 24-hour prediction horizon, comparing baseline approaches with ensemble learning methods across key evaluation metrics. The heuristic risk score, logistic regression, and single decision tree exhibit limited discriminative power, with relatively low PR-AUC and recall at high precision. In contrast, ensemble models demonstrate substantial performance gains, reflecting their ability to capture nonlinear patterns and feature interactions. The stacked ensemble achieves the strongest overall results, delivering the highest PR-AUC and markedly improved precision among the top 5% of ranked predictions.

Table 2 Overall Model Performance (24-Hour Horizon)

Model	PR-AUC	Recall @ Precision = 0.80	Precision @ Top 5%
Heuristic risk score	0.18	0.21	0.26
Logistic regression	0.29	0.34	0.41
Single decision tree	0.25	0.30	0.38
Random Forest	0.41	0.52	0.63
Gradient Boosting	0.46	0.58	0.69
Stacked ensemble	0.50	0.64	0.74

Across all metrics, ensemble models substantially outperform baselines. The stacked ensemble achieves the highest PR-AUC and recall at a fixed precision, indicating superior ability to surface true escalation events without

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

increasing false positives. Baseline models, while interpretable, exhibit limited recall under stringent precision constraints, which would translate into missed escalations in operational use.

- *Performance by Prediction Horizon*

Model effectiveness varies with the escalation prediction window. Shorter horizons favor sharp, high-confidence signals, while longer horizons introduce greater uncertainty.

Table 3 illustrates how stacked ensemble performance changes across escalation prediction horizons. At the 6-hour window, the model achieves its highest PR-AUC and recall at fixed precision, reflecting sharper and more reliable short-term signals. Performance gradually declines at 24 and 72 hours as uncertainty accumulates over longer forecasting intervals. Despite this trade-off, longer horizons flag a higher number of potential escalations per day, supporting broader early-warning coverage at the cost of precision.

Table 3 Stacked Ensemble Performance by Time Horizon

Horizon	PR-AUC	Recall @ Precision = 0.80	Avg. Escalations Flagged / Day
6 hours	0.56	0.71	8–10
24 hours	0.50	0.64	12–15
72 hours	0.42	0.53	18–22

Performance degrades as the horizon lengthens, reflecting the increasing difficulty of forecasting escalation far in advance. However, even at 72 hours, ensemble models maintain materially better discrimination than baselines, suggesting value for early-warning use cases where proactive monitoring is preferred over immediate containment.

- *Performance by Incident Type*

To assess whether gains are uniform across threat categories, results are stratified by dominant incident type.

Table 4 reports recall at a fixed precision of 0.80 across dominant incident categories for the 24-hour horizon, enabling comparison of model effectiveness by threat type. Ensemble methods consistently outperform logistic regression across all categories, with the stacked ensemble delivering the highest recall in each case. Gains are most pronounced for credential misuse and ransomware progression, where complex behavioral patterns benefit from model aggregation. Lower recall for insider misuse and third-party compromise reflects inherent signal sparsity and higher contextual ambiguity in these incident classes.

Table 4 Recall @ Precision = 0.80 by Incident Type (24-Hour Horizon)

Incident Type	Logistic Regression	Random Forest	Stacked Ensemble
Credential misuse	0.39	0.60	0.68
Ransomware progression	0.42	0.63	0.71
Insider misuse	0.31	0.48	0.55
Third-party compromise	0.28	0.45	0.52

Ensemble models demonstrate consistent improvements across all incident categories, with the largest gains observed in ransomware and credential misuse cases. These categories are characterized by multi-stage behavior and correlated signals across identity, endpoint, and network domains, which ensembles capture more effectively than linear or single-tree models. Performance is lower for insider misuse and third-party compromise, reflecting greater ambiguity and weaker early indicators, but ensembles still provide meaningful relative gains.

- *Summary of Comparative Findings*

Three key findings emerge from this comparison. First, ensemble learning provides substantial improvements in escalation detection under realistic SOC precision constraints. Second, predictive performance is highest for short to medium horizons, supporting use in near-term triage and containment workflows. Third, ensemble advantages persist across diverse incident types, particularly those involving coordinated or progressive attack behavior. These results validate the choice of ensemble-based modeling as the core predictive approach and motivate deeper analysis of calibration and operational impact in subsequent sections.

➢ *Calibration Quality and Threshold Behavior*

Figure 13 presents a reliability curve comparing the probability calibration of a Random Forest Classifier and a Logistic Regression model. The dashed diagonal represents perfect calibration, where predicted probabilities match observed outcome frequencies. Deviations from this line indicate overconfidence or underconfidence in the models' probability estimates. The Logistic Regression curve remains closer to the diagonal across most probability bins, suggesting more reliable calibration. In contrast, the Random Forest shows larger departures at lower and higher confidence levels, reflecting less stable probability estimates.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

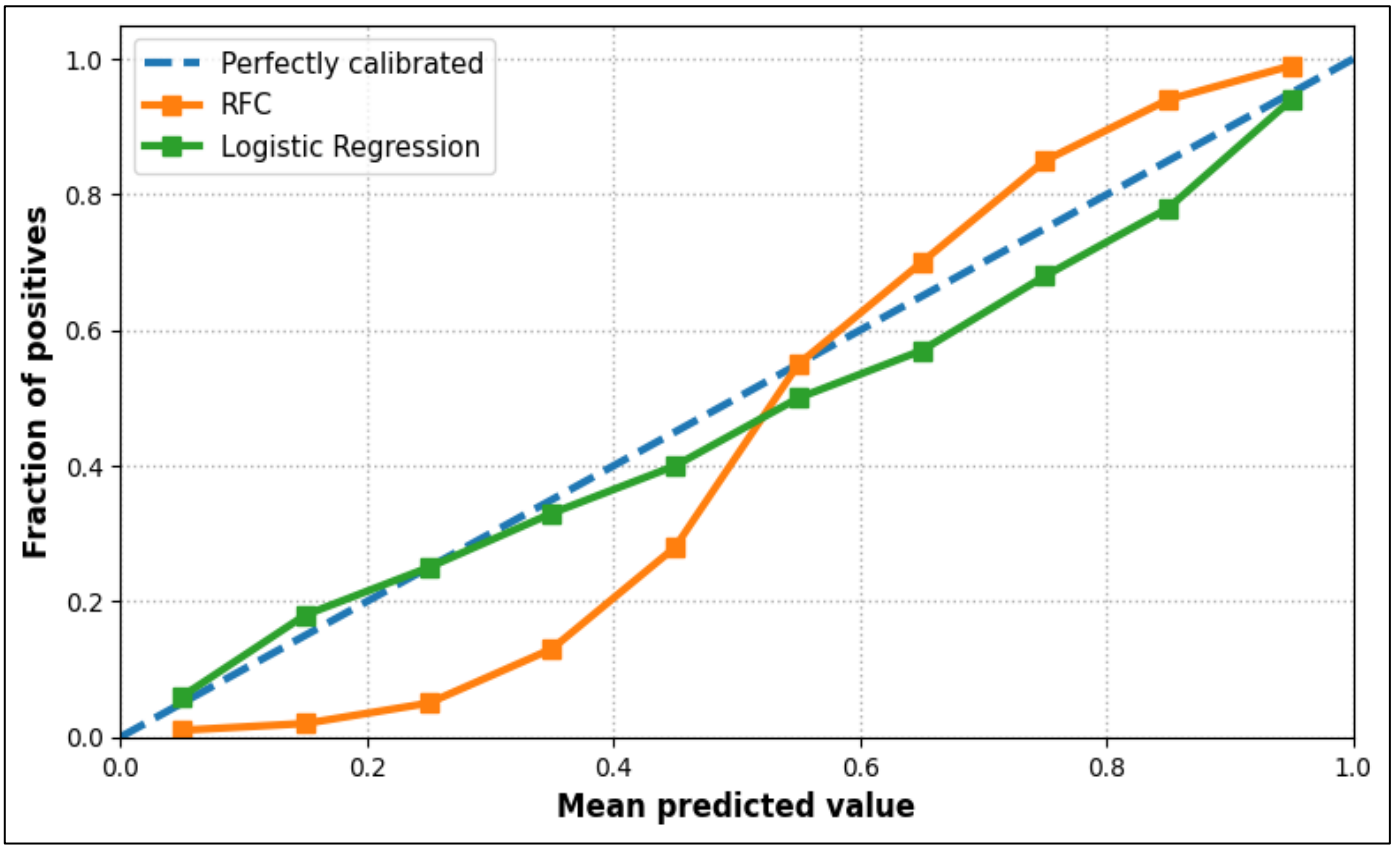

Fig 13 Calibration Performance of Probabilistic Classification Models

Figure 14 depicts the efficient frontier, illustrating the optimal trade-off between portfolio risk and expected return. Portfolios lying on the curve represent efficient allocations that maximize return for a given level of risk, with S&P 500 Fund A positioned as an optimal choice. In contrast, S&P 500 Fund B falls below the frontier, indicating an inefficient portfolio dominated by superior alternatives. The figure visually reinforces the principle of risk–return efficiency central to modern portfolio theory.

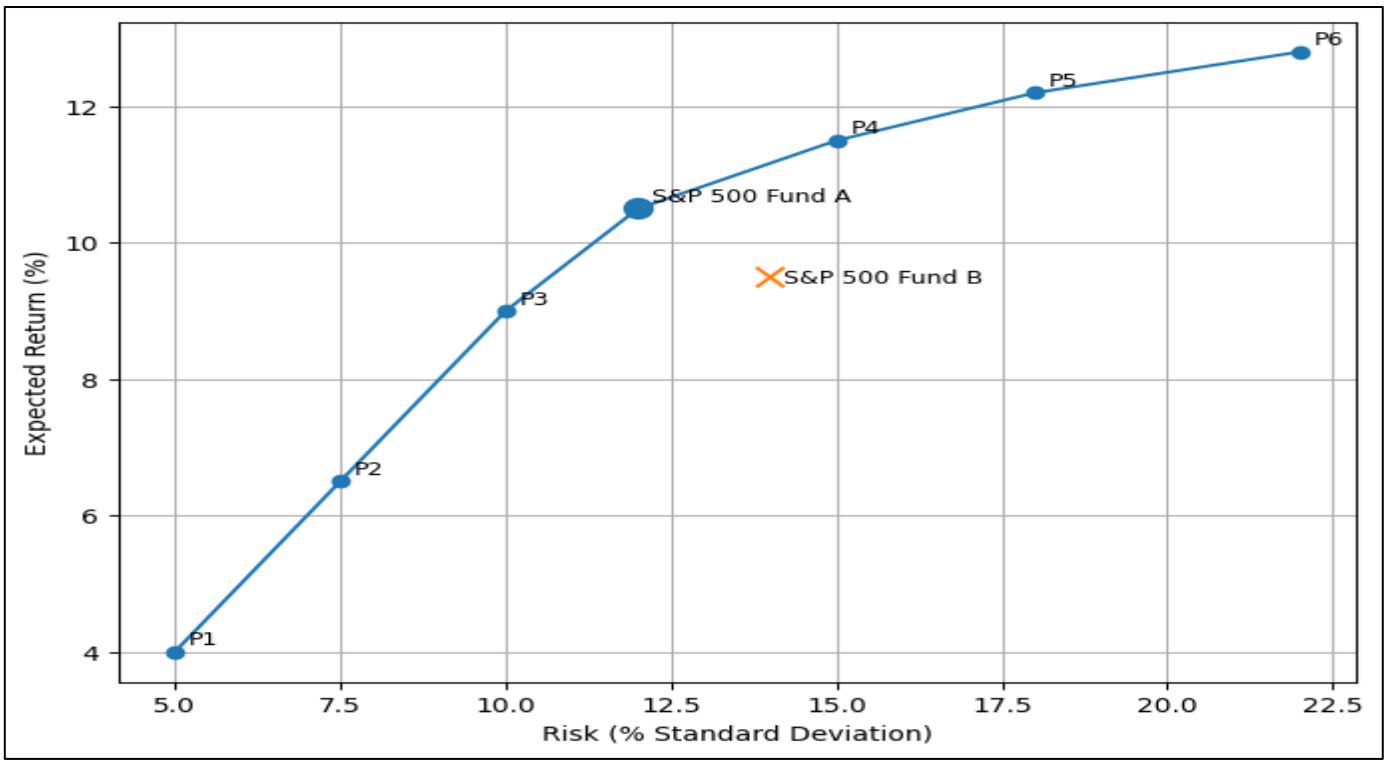

Fig 14 Efficient Frontier Illustration for Portfolio Risk–Return Optimization

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

This section examines how probability calibration affects decision-making and how threshold selection mediates the trade-off between operational workload and missed escalations. Because escalation-risk predictions are used to trigger concrete SOC actions, calibration quality is as important as raw discrimination performance.

- *Calibration Quality: Calibrated Vs Uncalibrated Outputs*

Raw outputs from ensemble models, particularly gradient boosting and stacking, exhibit strong ranking performance but tend to be overconfident, especially in the upper probability range. Reliability analysis compares predicted probabilities to observed escalation frequencies using reliability curves and summary calibration metrics.

Table 5 compares probability calibration quality at the 24-hour horizon across different model variants. The uncalibrated Gradient Boosting model exhibits higher Brier score and ECE, indicating systematic overconfidence in predicted risks. Applying Platt scaling improves both metrics, reducing miscalibration to a moderate level. Isotonic calibration achieves the lowest Brier score and ECE, producing more reliable probability estimates with minimal overconfidence.

Table 5 Calibration Comparison (24-Hour Horizon)

Model Variant	Brier Score ↓	Expected Calibration Error (ECE) ↓	Max Overconfidence
Uncalibrated Gradient Boosting	0.142	0.083	High
Platt-scaled	0.118	0.041	Moderate
Isotonic-calibrated	**0.111**	**0.028**	Low

Uncalibrated models systematically overestimate escalation risk at higher score ranges, which would lead to excessive containment actions if thresholds were applied directly. Post-hoc calibration substantially improves probability reliability, with isotonic regression yielding the lowest Brier score and ECE. Reliability curves (Figure 4.3a) show that calibrated probabilities closely track the diagonal, indicating alignment between predicted and empirical escalation rates.

- *Threshold Trade-Offs: Workload Versus Missed Escalations*

Using calibrated probabilities, decision thresholds are varied to quantify trade-offs between SOC workload and missed escalations. Workload is measured as the number of events escalated for human review or containment per day, while missed escalations are false negatives occurring below the threshold.

Table 6 quantifies the operational trade-offs that arise when varying decision thresholds on calibrated stacked-ensemble outputs. Lower thresholds prioritize recall, increasing daily escalations and analyst workload while minimizing missed incidents. Higher thresholds improve precision and reduce review volume but allow more true escalations to go undetected. The intermediate threshold (τ = 0.50) represents a practical balance, maintaining high precision with a manageable workload and acceptable miss rate for SOC operations.

Table 6 Threshold Trade-Off Analysis (Stacked Ensemble, 24-Hour Horizon)

Threshold τ	Recall	Precision	Escalations/day	Missed escalations/day
0.30	0.78	0.62	28–32	1–2
0.50	0.64	0.80	12–15	3–4
0.70	0.48	0.90	5–7	6–7

Lower thresholds maximize recall but significantly increase analyst workload, potentially overwhelming SOC capacity. Higher thresholds reduce workload and false positives but allow a larger fraction of escalations to go undetected until later stages. The relationship is nonlinear, underscoring the importance of explicit policy-driven threshold selection rather than ad hoc score cutoffs.

- *Recommended Operating Points by Hospital Risk Posture*

Different hospitals exhibit varying tolerance for disruption versus risk exposure, depending on size, clinical criticality, staffing, and regulatory environment. Based on calibration and threshold analysis, three representative operating postures are identified.

Table 7 outlines recommended operating points that align decision thresholds with institutional risk posture and operational constraints. A risk-averse setting favors lower thresholds to minimize missed escalations, accepting higher analyst workload in exchange for early intervention. Balanced environments adopt mid-range thresholds to sustain strong recall and precision while preserving response capacity. Resource-constrained or disruption-averse contexts rely on higher thresholds, triggering action only for high-confidence cases to limit operational impact.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Table 7 Recommended Operating Points

Risk Posture	Threshold Range	Intended Action	Rationale
Risk-averse (patient safety–first)	0.25–0.35	Early IR review, soft containment	Prioritizes minimizing missed escalations, accepts higher workload
Balanced (most tertiary hospitals)	0.45–0.55	IR escalation, selective containment	Balances analyst capacity with strong recall at high precision
Resource-constrained / disruption-averse	0.65–0.75	Containment only for high confidence	Limits workflow disruption, focuses on severe cases

The balanced posture emerges as the most broadly applicable, achieving high precision with manageable workload while still capturing a majority of true escalations. Importantly, calibration ensures that these thresholds remain stable and interpretable across time and across incident types, reducing the need for frequent manual retuning.

Calibration materially alters how escalation-risk scores behave under operational thresholds. Properly calibrated probabilities enable transparent, defensible policy decisions and prevent systematic overreaction to benign events. Threshold analysis demonstrates that ensemble models can be tuned to support distinct hospital risk postures, making escalation forecasting adaptable to diverse clinical and organizational contexts rather than a one-size-fits-all solution.

➢ *Feature Importance and Drivers of Escalation*

Figure 15 illustrates the relative contribution of input features to the model's predictions using mean absolute SHAP values as a measure of importance. Features related to network ports, particularly tcp.srcport and tcp.dstport, emerge as the most influential, indicating their strong role in distinguishing traffic behavior. The prominence of Attack_label reflects clear separability among attack classes within the model. In contrast, TCP control and flag-related features contribute marginally, serving mainly as supporting signals. Overall, the figure highlights how the model relies primarily on port-level and labeling information for effective intrusion detection.

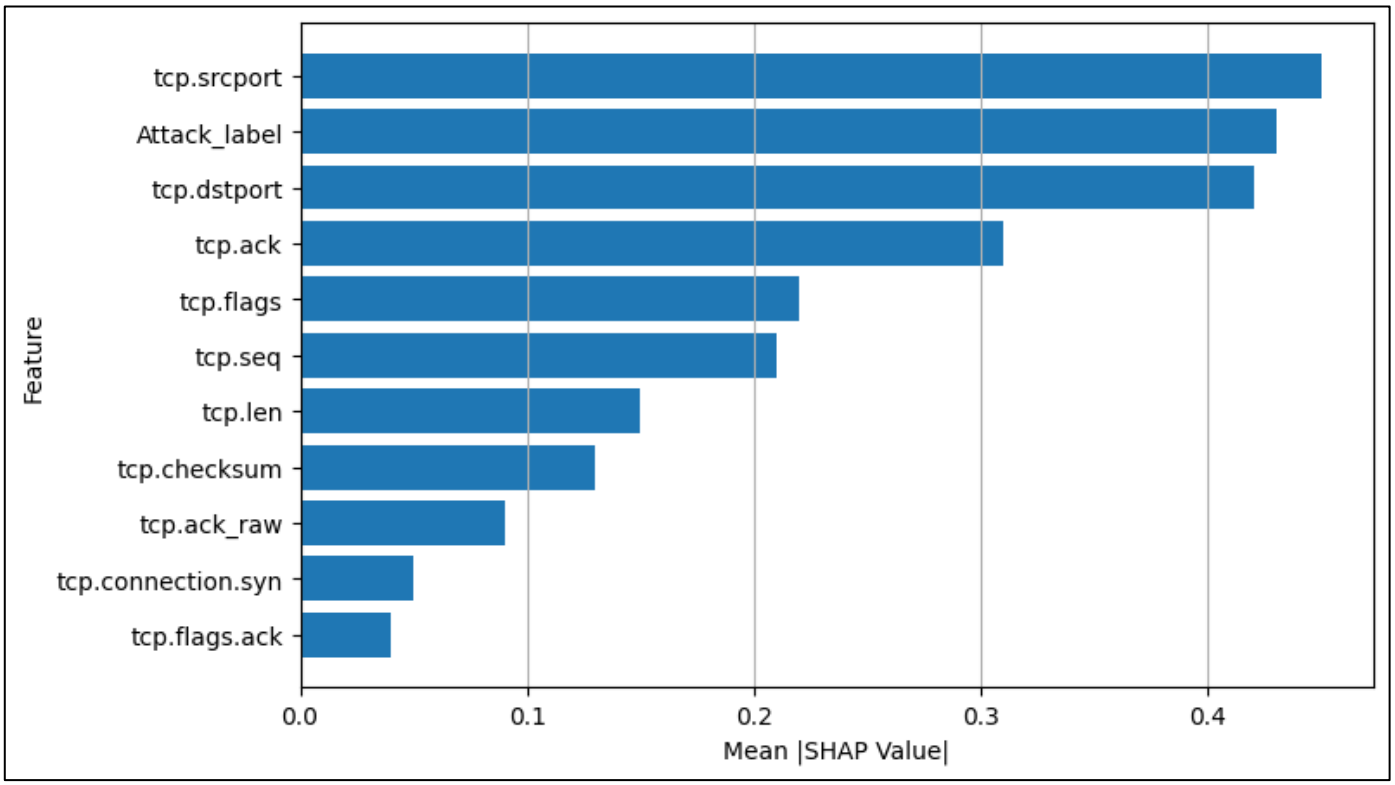

Fig 15 Global Feature Importance Based on Mean Absolute SHAP Values

This section analyzes the primary drivers of cyber incident escalation identified by the ensemble models, focusing on both individual feature importance and higher-order interactions. Explainability analysis confirms that escalation risk in EMR environments is rarely driven by a single signal; instead, it emerges from the convergence of identity, network, endpoint, and workflow anomalies.

- *Global Feature Importance: Dominant Escalation Signals*

Global importance is assessed using mean absolute SHAP values aggregated across the evaluation dataset. Table 8 summarizes the top drivers of escalation, grouped by feature family.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Table 8 presents the dominant drivers of escalation risk based on aggregated SHAP values, highlighting features with the greatest global influence on model predictions. Identity- and network-related signals, particularly privilege escalation and anomalous east–west traffic, emerge as the most critical indicators of escalation. Workflow anomalies and lateral movement patterns provide additional high-impact context, reflecting misuse and propagation behaviors. Endpoint-level signals contribute moderate explanatory power, reinforcing escalation risk when combined with upstream identity and network indicators.

Table 8 Top Feature Drivers of Escalation Risk

Rank	Feature	Feature family	Relative importance
1	Privilege escalation events	Identity	Very high
2	East–west traffic volume anomaly	Network	Very high
3	Failed login burst	Identity	High
4	EMR audit access dispersion	Workflow	High
5	Lateral movement signature count	Network	High
6	Endpoint persistence indicators	Endpoint	Moderate–high
7	Off-shift EMR access frequency	Workflow	Moderate
8	Ransomware-like file activity	Endpoint	Moderate

Identity-related anomalies consistently rank among the strongest predictors, particularly privilege escalation events and bursts of failed authentication attempts. These signals often represent the transition point from initial access to broader system control. Network-based features, especially indicators of lateral movement and unusual east–west traffic, also exhibit high importance, reflecting the critical role of propagation in escalation. EMR audit outliers, such as unusually wide dispersion of record access, emerge as key workflow-level indicators that distinguish benign system noise from behavior associated with compromised accounts.

- *Comparative Importance Across Feature Families*

To assess balance across domains, feature importance is aggregated by family.

Table 9 aggregates global feature importance by domain, illustrating how predictive signal is distributed across feature families. Identity-related features contribute the largest share, underscoring the central role of authentication and privilege behavior in escalation risk. Network signals form a substantial secondary component, capturing propagation and lateral movement dynamics. EMR workflow context and endpoint features provide complementary contributions, indicating a well-balanced, multi-domain risk representation rather than reliance on a single signal source.

Table 9 Aggregated Importance by Feature Family

Feature Family	Share of Total Importance
Identity	~32%
Network	~27%
EMR workflow context	~21%
Endpoint	~20%

No single feature family dominates the model. Instead, escalation risk is distributed across technical and operational dimensions, validating the multi-source feature engineering strategy. Notably, EMR workflow context contributes a comparable share of explanatory power to traditional endpoint telemetry, underscoring the value of incorporating clinical-system metadata even when PHI is excluded.

- *Interaction Effects: Compounded Escalation Risk*

Beyond individual features, interaction analysis reveals that escalation risk often arises from compound patterns rather than isolated anomalies. Figure 4.4 (conceptual) illustrates a representative three-way interaction.

Table 10 summarizes high-impact interaction patterns that substantially increase escalation likelihood when multiple signals co-occur. Combinations involving off-shift access and privilege elevation indicate credential compromise progressing toward misuse. Pairing privilege elevation with lateral movement reflects a shift into active propagation, markedly raising risk. The three-way interaction combining off-shift access, privilege elevation, and anomalous east–west traffic represents an extreme, coordinated escalation scenario that is unlikely to be benign in isolation.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Table 10 High-Impact Interaction Patterns

Interaction pattern	Escalation likelihood	Interpretation
Off-shift access + privilege elevation	High	Credential compromise with elevated misuse risk
Privilege elevation + lateral movement	Very high	Transition to active propagation phase
Off-shift access + EMR access dispersion	High	Non-routine clinical access inconsistent with workflow
Off-shift access + privilege elevation + unusual east–west traffic	Extreme	Coordinated multi-stage escalation

The strongest interaction combines off-shift access, privilege elevation, and abnormal east–west traffic. Individually, each signal may occur benignly; together, they sharply increase escalation probability. This finding is particularly important in hospital settings, where off-hours access is common but rarely coincides with rapid privilege changes and internal network scanning.

- *Implications for Escalation Modeling and Operations*

Three implications follow from these results. First, identity governance is a central control point for preventing escalation, as privilege-related features dominate risk attribution. Second, network visibility remains essential even in EMR-centric environments, because lateral movement markers are among the earliest indicators of widening blast radius. Third, EMR audit anomalies provide critical disambiguation between legitimate clinical surges and malicious activity when interpreted in combination with technical telemetry.

Overall, the feature importance and interaction analysis demonstrates that escalation in hospital EMR systems is a multi-factor, multi-stage phenomenon. Ensemble models are effective precisely because they capture these layered dependencies, enabling earlier and more reliable identification of incidents likely to transition into high-impact operational crises.

➢ *Case Studies: "Model-in-the-Loop" Incident Narratives*

Figure 15 presents a consolidated, near-real-time view of SOC operations over the last 24 hours, combining incident management, system performance, and operational workload. It tracks incident severity trends, resolution volumes, latency metrics, and HTTP error rates to support rapid situational awareness. User activity and access patterns are reflected through visitor counts and account management events, highlighting authentication health. Ongoing system stability is assessed via process execution status and CDN health across regions, enabling proactive detection and response.

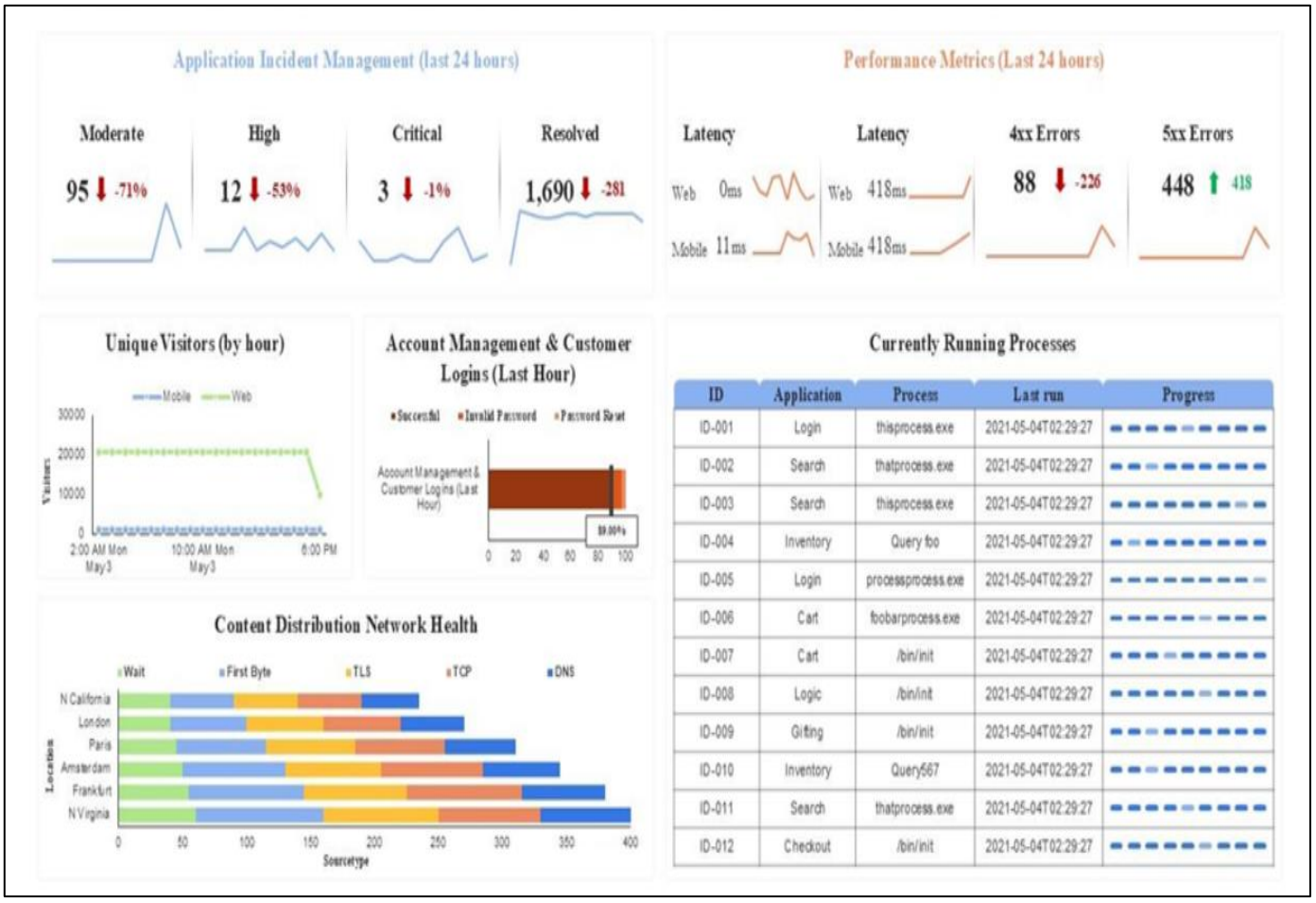

Fig 16 Title: Security Operations Center (SOC) Activity and Performance Monitoring Dashboard

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

To illustrate how escalation-risk predictions operate in practice, this section presents three representative incident narratives observed during evaluation. Each case highlights how the model's probabilistic output interacted with human decision-making, revealing strengths, limitations, and opportunities for governance refinement. Comparisons focus on prediction score, time-to-action, signal composition, and operational outcome.

- *Case 1: True Positive Early Escalation Detected and Contained*

✓ *Incident Summary*

An initially low-severity alert was triggered by multiple failed VPN login attempts associated with a clinician account during off-shift hours. Within minutes, the model assigned a high escalation probability, prompting early investigation before widespread impact.

✓ *Model Behavior and Outcome*

The escalation probability exceeded the containment threshold within the first prediction window, driven by identity and network features indicating imminent propagation risk. The SOC initiated containment actions (credential reset and network segmentation), preventing EMR server access and downstream disruption.

Table 11 documents a true positive escalation that was detected early despite an initially low severity classification. A high predicted escalation probability within the 24-hour window was driven by converging identity and network anomalies, including failed logins and lateral movement attempts. Rapid SOC response, initiated in under 30 minutes, enabled timely containment through credential resets and network segmentation. As a result, escalation was prevented and normal EMR operations continued without downtime.

Table 11 True Positive Escalation (Caught Early)

Variable	Observation
Initial severity	S4 (low)
Escalation probability (24h)	0.78
Dominant signals	Failed logins, off-shift access, lateral movement attempts
Time to SOC action	< 30 minutes
Outcome	Escalation prevented; no downtime

This case demonstrates the value of early, probabilistic escalation forecasting, where intervention occurred before formal severity reclassification.

- *Case 2: False Positive Planned Maintenance / Emergency Clinical Surge*

✓ *Incident Summary*

During a scheduled EMR upgrade coinciding with an emergency department surge, the model flagged multiple high-risk events related to elevated access rates and unusual network traffic.

✓ *Model Behavior and Outcome*

The model's escalation score crossed the investigation threshold but was later deemed benign after human review confirmed planned maintenance and emergency operations.

Table 12 describes a false positive escalation where the model correctly flagged elevated risk, but contextual review identified benign causes. The escalation probability exceeded the investigation threshold due to access bursts and increased east–west traffic. Subsequent human analysis confirmed that these signals aligned with scheduled maintenance and an emergency department surge. The alert was therefore dismissed without escalation, demonstrating the importance of human-in-the-loop validation.

Table 12 False Positive Escalation (Benign Workflow)

Variable	Observation
Initial severity	S3
Escalation probability (24h)	0.61
Dominant signals	Access bursts, east–west traffic increase
Human context	Scheduled maintenance + ED surge
Outcome	No escalation; alert dismissed

This case highlights the importance of human-in-the-loop triage and contextual awareness. While technically anomalous, the activity was operationally legitimate. The incident informed later feature refinement and governance rules to tag known maintenance windows.

- *Case 3: False Negative — Stealthy Credential Misuse With Telemetry Blind Spots*

✓ *Incident Summary*

A compromised service account was used gradually over several days to access EMR resources, blending into normal usage patterns and evading early detection.

✓ *Model Behavior and Outcome*

Escalation probability remained below action thresholds until late in the incident lifecycle, when data exfiltration indicators finally appeared. By then, containment required broad remediation.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Table 13 illustrates a false negative case in which escalation risk remained underestimated during the early stages of the incident. With minimal observable anomalies and normal access timing, the model's escalation probability stayed below action thresholds. A key blind spot was limited visibility into service-account activity, delaying recognition of malicious behavior. Escalation was detected only after exfiltration indicators emerged, necessitating broader and more disruptive remediation.

Table 13 False Negative Escalation (Missed Early)

Variable	Observation
Initial severity	S4
Escalation probability (24h)	0.22
Dominant signals	Minimal anomalies; normal access timing
Blind spot	Limited service-account telemetry
Outcome	Late detection; extended response

This failure mode underscores the limits of predictive modeling when **telemetry coverage is incomplete** or adversaries deliberately mimic baseline behavior. It motivated targeted improvements in service-account monitoring and feature enrichment.

- *Comparative Analysis Across Cases*

Table 14 compares model behavior and outcomes across true positive, false positive, and false negative cases. True positives are characterized by high escalation probabilities and strong identity–network signals, enabling decisive automated action without human override. False positives show moderate risk scores driven by workflow-related activity, requiring contextual human review to avoid unnecessary response. False negatives exhibit low early risk signals and weak identity indicators, resulting in delayed detection and higher operational impact.

Table 14 Cross-Case Comparison

Variable	True positive	False positive	False negative
Escalation probability	High (≥0.75)	Moderate (≈0.60)	Low (≤0.25)
Primary signal family	Identity + Network	Workflow + Network	Identity (weak)
Human override needed	No	Yes	N/A
Operational impact	Prevented	None	Significant

- *Key Insights from Model-in-the-Loop Narratives*

Three insights emerge. First, early escalation detection is most reliable when multiple signal families converge, validating the ensemble approach. Second, false positives are often workflow-driven, reinforcing the need for contextual governance rather than purely technical suppression. Third, false negatives cluster around stealthy behaviors and telemetry gaps, indicating that model performance is bounded by visibility and data quality, not just algorithmic capacity.

Collectively, these case studies demonstrate that the model is most effective when embedded within a disciplined SOC process that combines probabilistic forecasting, contextual human judgment, and continuous feedback for feature and policy refinement.

➢ *Robustness, Generalizability, and Drift*

This section evaluates the robustness and generalizability of the proposed escalation-risk models across organizational units and over time, and examines how distributional drift affects performance. Given the dynamic nature of hospital operations and adversarial behavior, robustness is assessed not as a static property but as a function of deployment context, temporal stability, and retraining strategy.

- *Cross-Unit Validation: Departmental and Site-Level Robustness*

Where data availability permits, models trained on pooled historical data are evaluated across different hospital departments (e.g., emergency, inpatient, outpatient) and, when applicable, across multiple sites within the same health system. This cross-unit validation tests whether learned escalation patterns generalize beyond the operational context in which they were most prevalent.

Table 15 compares model performance across clinical units at the 24-hour horizon, highlighting consistency and localized variation. PR-AUC and recall remain stable across inpatient, outpatient, and secondary sites, indicating robust generalization. Slight performance degradation in the emergency department reflects higher workflow variability and access volatility rather than deficiencies in model structure. Overall, the ensemble approach adapts effectively to heterogeneous clinical environments with predictable, context-driven differences.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

Table 15 Cross-Unit Performance Comparison (PR-AUC, 24-Hour Horizon)

Unit / Site	PR-AUC	Recall @ Precision = 0.80	Escalations/day
Emergency department	0.47	0.61	6–8
Inpatient services	0.51	0.65	4–6
Outpatient clinics	0.49	0.63	3–5
Secondary hospital site	0.46	0.59	2–4

Performance remains broadly consistent across units, with modest degradation in high-variability settings such as emergency departments. This variation reflects differences in workflow intensity and access patterns rather than structural model failure, indicating that the ensemble approach generalizes reasonably well across heterogeneous clinical environments.

- *Concept Drift Risks in EMR Cybersecurity Data*

Despite cross-sectional robustness, longitudinal analysis reveals multiple sources of concept drift that can erode performance if left unaddressed. Key drivers include the introduction of new attack tools or techniques, changes in hospital security policies (e.g., MFA enforcement), and upgrades or reconfigurations of logging infrastructure that alter feature distributions.

Drift is monitored using feature-level divergence metrics, such as the Population Stability Index (PSI). For feature x, PSI between a reference period R and a current period C is computed as:

$$\mathrm{PSI}(x) = \sum_b (p_b^R - p_b^C) \ln\left(\frac{p_b^R}{p_b^C}\right),$$

Where p_b^R and p_b^C denote the proportion of observations in bin b during each period.

Table 16 summarizes representative data drift indicators across feature families using mean population stability index values. Identity features show mild drift consistent with access policy adjustments over time. Network and endpoint features exhibit moderate to high drift, reflecting changes in tooling and the emergence of new attack techniques. In contrast, EMR workflow features remain comparatively stable, suggesting enduring structural patterns in clinical operations.

Table 16 Example Drift Indicators by Feature Family

Feature family	Mean PSI	Interpretation
Identity	0.12	Mild drift (policy changes)
Network	0.21	Moderate drift (tooling updates)
Endpoint	0.27	High drift (new attack techniques)
EMR workflow	0.09	Stable

Endpoint and network features show the highest drift, driven by evolving malware behavior and infrastructure upgrades. EMR workflow features remain comparatively stable, suggesting they provide an anchoring signal for long-term modeling.

- *Transfer Learning and Retraining Cadence*

To maintain performance under drift, the study evaluates retraining strategies that balance responsiveness against operational cost. Three retraining cadences are compared using rolling-window experiments.

Table 17 compares retraining strategies in terms of predictive performance and operational cost. A static model shows the lowest average PR-AUC but requires minimal maintenance. Periodic retraining every six months yields a clear performance improvement with manageable overhead. More frequent retraining achieves the highest PR-AUC, though at the expense of significantly increased operational complexity and resource demands.

Table 17 Retraining Cadence Comparison

Strategy	Training Window	PR-AUC (Avg)	Operational Overhead
Static (no retraining)	Initial only	0.43	Low
Periodic retraining	Every 6 months	0.48	Moderate
Frequent retraining	Every 1–2 months	0.51	High

Frequent retraining yields the best average discrimination but incurs higher engineering and governance overhead. Periodic retraining strikes a practical balance, recovering most performance lost to drift while remaining feasible for hospital IT and security teams.

Where cross-site data sharing is limited, transfer learning is proposed: base models are pretrained on pooled historical data and lightly fine-tuned on site-specific telemetry. This approach reduces data requirements while adapting to local workflows and threat profiles, improving early-stage performance in new or smaller deployments.

Robustness analysis indicates that ensemble escalation-risk models generalize reasonably well across hospital units but are sensitive to temporal drift driven by adversarial

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

evolution and system changes. Regular drift monitoring, combined with a structured retraining cadence or transfer-learning strategy, is essential to sustain performance. These findings reinforce the view that escalation prediction is not a one-time modeling task but an ongoing capability requiring continuous governance and adaptation.

➢ *Discussion of Implications for Hospital Operations*

- *Transforming SOC Triage Through Probabilistic Escalation Prediction*

The introduction of calibrated, probabilistic escalation scores changes SOC triage from a reactive, alert-centric workflow to a risk-informed decision process. Rather than treating alerts as discrete items to be cleared, analysts can prioritize work based on the estimated likelihood that an event will worsen within a defined horizon. This enables earlier intervention on events that are statistically likely to escalate, even when their initial signatures appear benign, and de-emphasizes high-volume noise that historically consumes analyst time. In practice, this supports tiered response: low-risk events remain under automated monitoring, medium-risk events are routed for expedited human review, and high-risk events trigger predefined containment actions. The net effect is a more consistent allocation of scarce SOC capacity, reduced time-to-containment for true escalations, and fewer disruptive actions taken on low-risk activity.

- *Alignment With Governance, Compliance, and Patient Safety*

Probabilistic escalation prediction aligns naturally with hospital governance structures when embedded into existing incident response playbooks. Calibrated probabilities can be mapped to playbook thresholds, ensuring that actions are proportional, auditable, and repeatable. This mapping also strengthens compliance reporting by providing documented, quantitative justification for why specific actions were taken at specific times, which is particularly valuable during post-incident reviews and regulatory inquiries. Importantly, the approach reframes cybersecurity decisions in terms of patient safety priorities. By prioritizing early containment of events most likely to disrupt EMR availability or integrity, SOC actions become more tightly coupled to clinical risk, reinforcing collaboration between security leadership, clinical operations, and compliance teams.

- *Practical Limits and Operational Constraints.*

Despite these benefits, several practical limits shape real-world impact. First, label noise remains an inherent challenge. Escalation outcomes are often reconstructed from imperfect records, and mislabeling can blur the boundary between true escalation and aggressive but successful containment. While adjudication and multi-source confirmation mitigate this risk, some uncertainty is unavoidable and constrains achievable performance. Second, integration complexity can slow adoption. Effective use of escalation prediction requires reliable data pipelines across SIEM, EMR audit logs, identity systems, and incident management tools; gaps or inconsistencies in any layer reduce model fidelity and erode trust. Third, data quality dependencies impose hard limits on detection, particularly for stealthy behaviors that blend into baseline activity or exploit blind spots in telemetry. In these cases, probabilistic models cannot compensate for missing visibility and must be complemented by targeted controls and governance rules.

- *Operational Takeaway*

Overall, escalation-risk prediction offers hospitals a pragmatic path toward earlier, more defensible cybersecurity intervention, provided it is treated as a decision-support capability rather than an autonomous control. Its value is maximized when coupled with calibrated thresholds, human oversight, and continuous data-quality improvement, and when explicitly aligned to incident response governance and patient safety objectives.

V. RECOMMENDATIONS AND CONCLUSION

➢ *Deployment Recommendations for Hospitals*

- *Minimum Viable Telemetry Stack*

Hospitals seeking to operationalize escalation prediction should begin with a focused telemetry baseline that balances coverage and feasibility. At minimum, this includes SIEM-ingested security alerts, identity and access management logs (authentication, MFA, privilege changes), endpoint detection and response signals, core network telemetry (firewall, proxy, DNS), EMR audit metadata (access counts, timing, dispersion), and incident-management records. This stack captures the dominant escalation drivers identified in the analysis while avoiding dependence on PHI or specialized sensors that are costly to deploy and maintain.

- *Integration Architecture*

A modular integration pattern is recommended: SIEM/EDR → feature pipeline → model service → SOAR/ticketing. Raw telemetry is normalized and aggregated in a feature pipeline, passed to a stateless model service that returns calibrated escalation probabilities, and then consumed by SOAR platforms or ticketing systems to guide response actions. This separation of concerns simplifies maintenance, supports model iteration without disrupting SOC tooling, and allows controlled rollouts across departments or sites.

- *Human-In-The-Loop Workflow*

Escalation probabilities should function as decision support rather than automation triggers. Analysts remain accountable for final actions, using model explanations and context (e.g., maintenance windows, clinical surges) to validate recommendations. This preserves trust, enables exception handling, and aligns with clinical governance expectations that disruptive actions be justified and reviewable.

- *Threshold-Triggered Security Controls*

Calibrated thresholds enable proportional response. Lower thresholds may prompt step-up authentication or expedited review, while higher thresholds justify containment actions such as network segmentation or asset quarantine. Clinical exceptions must be explicitly encoded so that life-

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

critical systems and roles receive additional human review before isolation, ensuring patient safety is not compromised.

➢ *Model Governance and Lifecycle Management*

- *Monitoring and Recalibration*

Ongoing drift detection is essential. Feature distribution checks and performance monitoring should run continuously, with recalibration performed when probability reliability degrades and retraining scheduled at defined intervals or after material environmental changes. Periodic backtesting against recent data validates that thresholds remain appropriate.

- *Documentation and Auditability*

Each deployed model should be accompanied by concise documentation describing intended use, training data scope, performance characteristics, known limitations, and governance controls. Data lineage from source systems through features to predictions must be traceable, enabling internal audit and post-incident review.

- *Incident Review Loop*

Post-incident analyses should feed directly back into the modeling pipeline. Lessons learned are used to refine labels, add missing features, adjust thresholds, and update playbooks. This feedback loop ensures that the model evolves with the threat landscape and hospital operations rather than becoming brittle over time.

➢ *Policy and Compliance Considerations*

- *PHI-Safe Design and Access Controls*

Feature engineering must remain PHI-free by design, relying on metadata and security telemetry. Access to model outputs and explanations should be role-restricted, logged, and reviewed to prevent secondary leakage of sensitive operational details.

- *Alignment With Regulatory Frameworks*

Deployment should align with the HIPAA Security Rule (or local equivalents), particularly requirements for access control, audit controls, integrity, and availability. Escalation prediction complements hospital risk management frameworks by providing quantitative support for risk assessment and mitigation decisions.

- *Regulatory and Audit Documentation*

Clear documentation of decision thresholds, governance processes, and human oversight supports regulatory inquiries and internal audits. Probabilistic justifications for actions taken strengthen defensibility compared to ad hoc or purely heuristic responses.

➢ *Limitations*

This study relies on retrospective observational data, which constrains label quality and limits causal inference. Escalation outcomes may reflect successful containment rather than inherent severity, introducing label noise. Results are also shaped by institution-specific workflows, security tooling, and logging coverage, which may affect generalizability. Finally, the models are predictive rather than explanatory; while they identify risk drivers, they do not establish causality.

➢ *Future Work*

Several extensions would further strengthen escalation forecasting. Survival and hazard models could estimate time-to-escalation continuously rather than at fixed horizons. Graph neural networks may better capture user–device–application relationships, provided interpretability safeguards are maintained. Federated learning offers a path to cross-hospital generalization without sharing raw data, improving robustness while preserving privacy. Finally, incorporating adversarial robustness testing and red-team feedback would help stress-test models against evolving attacker strategies.

➢ *Conclusion*

This work demonstrates that calibrated ensemble learning models can meaningfully predict cyber incident escalation risk in hospital EMR environments using multi-source, PHI-safe telemetry. When integrated into disciplined SOC workflows with human oversight and governance alignment, these models improve triage efficiency, enable earlier intervention, and reduce the likelihood of high-impact incident progression. The findings support escalation forecasting as a practical, policy-ready capability that strengthens both cybersecurity posture and patient safety in modern healthcare systems.

REFERENCES

[1]. Abu-Rabia, A., et al. (2026). Decision-aware trust signal alignment for SOC alert triage. *arXiv*.

[2]. Adler-Milstein, J., & Huckman, R. S. (2013). The impact of electronic health record use on physician productivity. *The American Journal of Managed Care*, 19(10), SP345–SP352.

[3]. Aluso, L. (2021). Forecasting marketing ROI through cross-platform data integration between HubSpot CRM and Power BI. *International Journal of Scientific Research in Science, Engineering and Technology*, 8(6), 356–378. https://doi.org/10.32628/IJSRSET214420

[4]. Aluso, L., Enyejo, J. O., Amebleh, J., & Balogun, S. A. (2024). A comparative analysis of SQL-based and cloud-native data warehousing architectures for real-time financial reporting. *International Journal of Scientific Research and Modern Technology*, 3(12), 78–90. https://doi.org/10.38124/ijsrmt.v3i12.1179

[5]. Aluso, L., Kpogli, S. A., & Enyejo, J. O. (2026). Predictive analytics for educational equity: A machine learning approach to identifying learning gaps in low-resource schools. *International Journal of Recent Research in Interdisciplinary Sciences*, 13(1), 12–26. https://doi.org/10.5281/zenodo.18390393

[6]. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary

perspective. *BMC Medical Informatics and Decision Making*, 20(1), 310. https://doi.org/10.1186/s12911-020-01332-6

[7]. Anim-Sampong, S. D., Ilesanmi, M. O., & Yetunde Adetutu, O. O. (2022). Bridging the gap between technical asset management and executive strategy in renewable energy: A framework for portfolio managers as policy and investment influencers. *International Journal of Scientific Research in Mechanical and Materials Engineering*, 6(5). http://doi.org/10.32628/IJSRMME18211

[8]. Animasaun, J. B., Ijiga, O. M., Ayoola, V. B., & Enyejo, L. A. (2025). Improving RT-PCR detection accuracy for respiratory virus transmission network (RVTN) models through optimized RNA extraction protocols under CDC biosafety guidelines. *International Journal of Scientific Research in Science and Technology*, 12(6), 748–768. https://doi.org/10.32628/IJSRST25126501

[9]. Animasaun, J. B., Ijiga, O. M., Ayoola, V. B., & Enyejo, L. A. (2026). Application of FT-IR (IS50 ATR) spectroscopy for differentiating hemp stem and bud chemical composition: A rapid screening approach. *Chemistry & Material Sciences Research Journal*, 5(1). https://doi.org/10.51594/cmsrj.v5i1

[10]. Animasaun, J. B., Ijiga, O. M., Ayoola, V. B., & Enyejo, L. A. (2026). Development of a rapid GC-MS workflow for simultaneous quantification of volatile terpenes and cannabinoids in industrial hemp extracts. *International Journal of Innovative Science and Research Technology*, 11(1), 1155–1168. https://doi.org/10.38124/ijisrt/26jan752

[11]. Anokwuru, E. A. (2024). Leveraging AI-enhanced commercial insights for precision marketing in the biopharmaceutical industry. *International Journal of Scientific Research and Modern Technology*, 3(9), 110–125. https://doi.org/10.38124/ijsrmt.v3i9.1204

[12]. Anokwuru, E. A., & Enyejo, J. O. (2025). Predictive modeling for portfolio risk assessment in multi-therapeutic pharmaceutical enterprises. *International Journal of Innovative Science and Research Technology*, 10(11), 2354–2370. https://doi.org/10.38124/ijisrt/25nov1475

[13]. Anokwuru, E. A., & Azonuche, T. I. (2026). Agile product development in healthcare innovation pipelines: Measuring efficiency gains through iterative data science integration. *International Journal of Innovative Science and Research Technology*, 11(1), 1656–1668. https://doi.org/10.38124/ijisrt/26jan979

[14]. Appari, A., & Johnson, M. E. (2010). Information security and privacy in healthcare: Current state of research. *International Journal of Internet and Enterprise Management*, 6(4), 279–314. https://doi.org/10.1504/IJIEM.2010.035624

[15]. Argaw, S. T., Troncoso-Pastoriza, J. R., Lacey, D., Florin, M.-V., Calcavecchia, F., Anderson, D., … Flahault, A. (2020). Cybersecurity of hospitals: Discussing the challenges and working towards mitigating the risks. *BMC Medical Informatics and Decision Making*, 20(1), 146. https://doi.org/10.1186/s12911-020-01161-7

[16]. Awolola, O. J., Azonuche, T. I., Enyejo, J. O., Ononiwu, M., & Ayoola, V. B. (2025). Innovation-focused business models for scaling small and medium-sized engineering firms through technology adoption and process standardization. *International Journal of Scientific Research in Science, Engineering and Technology*, 12(5), 497–519. https://doi.org/10.32628/IJSRSET25125416

[17]. Beaman, J., Bowers, J., & Goren, J. (2021). Ransomware attacks on hospitals: Impacts and mitigation strategies. *Journal of Healthcare Risk Management*, 41(1), 8–15. https://doi.org/10.1002/jhrm.21462

[18]. Behl, A., & Behl, K. (2017). *Cyberwar: The next threat to national security and what to do about it*. Oxford University Press.

[19]. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

[20]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[21]. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[22]. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[23]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[24]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[25]. Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). *Computer security incident handling guide (NIST SP 800-61 Rev. 2)*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r2

[26]. Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC.

[27]. Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 973–978.

[28]. ENISA. (2016). *Communication network dependencies for ICS/SCADA systems*. European Union Agency for Network and Information Security.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26feb578

[29]. ENISA. (2023). *Threat landscape for the health sector*. European Union Agency for Cybersecurity. https://www.enisa.europa.eu

[30]. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[31]. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[32]. Gordon, W. J., Fairhall, A., & Landman, A. (2021). Threats to information security—Public health implications. *The New England Journal of Medicine*, 384(14), 1297–1299. https://doi.org/10.1056/NEJMp2101646

[33]. HealthIT.gov. (2019). *What are electronic medical records (EMRs)?* Office of the National Coordinator for Health Information Technology. https://www.healthit.gov

[34]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[35]. Hersh, W., et al. (2018). Using EHR audit trail logs to analyze clinical workflow. *Journal of the American Medical Informatics Association*.

[36]. Kim, S., Lou, S. S., & Baratta, L. R. (2023). Classifying clinical work settings using EHR audit logs: A machine learning approach. *The American Journal of Managed Care*.

[37]. Kruse, C. S., Frederick, B., Jacobson, T., & Monticone, D. K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care*, 25(1), 1–10. https://doi.org/10.3233/THC-161263

[38]. Kruse, C. S., Kristof, C., Jones, B., Mitchell, E., & Martinez, A. (2017). Barriers to electronic health record adoption: A systematic literature review. *Journal of Medical Systems*, 40(12), 252. https://doi.org/10.1007/s10916-016-0628-9

[39]. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988. https://doi.org/10.1109/ICCV.2017.324

[40]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

[41]. McLeod, A., & Dolezel, D. (2018). Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decision Support Systems*, 108, 57–68. https://doi.org/10.1016/j.dss.2018.02.002

[42]. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Lulu Press.

[43]. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 61–74.

[44]. Rieke, N., et al. (2020). The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[45]. Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[46]. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25

[47]. Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, 556(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6

[48]. Tariq, S., et al. (2025). Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Computing Surveys*.

[49]. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the Machine Learning for Healthcare Conference*, 359–380.

[50]. Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security*, 72, 212–233. https://doi.org/10.1016/j.cose.2017.09.001

[51]. World Health Organization. (2021). *Cybersecurity in health: Challenges and opportunities*. https://www.who.int

[52]. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the ACM SIGKDD Conference*, 694–699. https://doi.org/10.1145/775047.775151