IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

Explainable AI Models for Secure and Transparent Decision-Making in Smart Computing Environments

Shallu Yadav
Assistant Professor, Skill Department of Computer Science / IT
Shri Vishwakarma Skill University, Dudhola, Palwal, India
shalluyadav1991@gmail.com

Abstract: *The rapid proliferation of smart computing environments—encompassing the Internet of Things (IoT), edge intelligence, cyber-physical systems, and autonomous control loops—has made automated decision-making indispensable for security, safety, and operational efficiency. However, the high-performing machine learning and deep learning models that drive these decisions typically behave as opaque "black boxes," offering little insight into why a particular outcome was produced. In security-critical settings such as intrusion detection, access control, and anomaly response, this opacity erodes trust, complicates auditing, and impedes regulatory compliance. This paper proposes an Explainable Artificial Intelligence (XAI) framework that couples a hybrid CNN-LSTM classifier with a post-hoc explanation layer based on SHAP, LIME, and attention weighting to deliver secure and transparent decisions. The framework was evaluated on a benchmark network-security dataset comprising normal and attack traffic across five categories. The proposed model attained an accuracy of 97.8%, an F1-score of 97.4%, and an area under the ROC curve (AUC) of 0.991, outperforming Random Forest, Support Vector Machine, XGBoost, and Multilayer Perceptron baselines. Beyond predictive accuracy, the explanation layer achieved a fidelity of 0.94 and a trust score of 0.89, while exposing the dominant decision drivers (e.g., packet inter-arrival time, flow byte rate, and failed authentication counts) in human-readable form. The results demonstrate that explainability and security are mutually reinforcing rather than competing objectives, and that transparent rationale generation can be embedded into smart computing pipelines with negligible loss of predictive power. The study contributes a reproducible methodology, a layered reference architecture, and a multi-dimensional evaluation protocol that jointly measures accuracy, robustness, and explanation quality.*

Keywords: Explainable AI (XAI); Smart Computing; Secure Decision-Making; Transparency; SHAP; LIME; Intrusion Detection; Deep Learning; Trustworthy AI; Edge Intelligence.

I. INTRODUCTION

Smart computing environments have evolved from isolated automation systems into deeply interconnected ecosystems in which sensors, actuators, edge nodes, gateways, and cloud services continuously exchange data and act with minimal human intervention. Smart homes regulate energy and security, smart grids balance electrical load, connected vehicles negotiate traffic, and industrial cyber-physical systems orchestrate manufacturing in real time. At the core of each of these systems lies an automated decision-making engine that ingests heterogeneous streams of telemetry and produces actions—granting access, raising alarms, throttling traffic, or reconfiguring devices. As the volume, velocity, and sensitivity of these decisions increase, so too does the cost of an incorrect or unexplained outcome.

Modern decision engines increasingly rely on machine learning (ML) and deep learning (DL) models because of their superior ability to capture nonlinear, high-dimensional patterns. Yet the very mechanisms that grant these models

their accuracy—dense layers of learned parameters, ensembles of thousands of trees, and complex feature interactions—render their reasoning opaque to human operators. This opacity is particularly problematic in security-critical smart computing, where an analyst must understand *why* traffic was flagged as malicious before isolating a device, and where auditors and regulators demand traceable justification for automated actions. The gap between predictive performance and human interpretability is the central tension this paper addresses.

Explainable Artificial Intelligence (XAI) has emerged as the discipline dedicated to bridging this gap. XAI techniques aim to render the behaviour of complex models intelligible without unduly sacrificing accuracy. Methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) attribute a model's output to its input features, while attention mechanisms expose which parts of a sequence most influenced a prediction. When integrated into smart computing pipelines, these techniques can transform an inscrutable alert into an actionable, auditable rationale.

Despite growing interest, three challenges persist. First, much XAI research evaluates explanations in isolation, without quantifying their impact on the security task itself. Second, the trade-off between predictive accuracy and explanation quality is rarely characterised in a unified protocol. Third, the operational requirements of smart computing—low latency, resource-constrained edge devices, and streaming data—are seldom reflected in experimental design. This paper responds to these gaps by proposing and evaluating an integrated XAI framework for secure and transparent decision-making.

1.1 Motivation

The motivation for this work is threefold. (i) **Trust:** operators are reluctant to act on, or delegate authority to, systems whose reasoning they cannot inspect. (ii) **Accountability:** emerging governance frameworks and data-protection regulations increasingly require that automated decisions affecting individuals be explainable and contestable. (iii) **Security hardening:** transparent models are easier to debug, audit, and defend against adversarial manipulation, because anomalous attributions can themselves serve as warning signals. Together, these forces make explainability a first-class requirement rather than an optional add-on.

1.2 Research Objectives

1. To design a hybrid deep-learning decision model suited to the streaming, high-dimensional data characteristic of smart computing security.
2. To integrate a model-agnostic explanation layer (SHAP, LIME, attention) that produces human-readable rationales for each decision.
3. To define a unified evaluation protocol that jointly measures predictive performance and explanation quality (fidelity, stability, sparsity, comprehensibility, trust).
4. To empirically demonstrate that transparency can be achieved with negligible degradation in security performance relative to opaque baselines.

1.3 Contributions

- A layered reference architecture that embeds explainability directly into the smart computing decision pipeline.
- A hybrid CNN-LSTM classifier that achieves state-of-the-art accuracy (97.8%) and AUC (0.991) on a benchmark intrusion-detection dataset.
- A multi-dimensional explanation evaluation that reports fidelity (0.94) and trust (0.89), establishing that accuracy and transparency are compatible.
- A reproducible methodology, summarised as a flowchart, that practitioners can adapt to other smart computing security tasks.

1.4 Paper Organization

The remainder of the paper is organised as follows. Section 2 reviews related work in XAI, smart computing security, and trustworthy AI. Section 3 details the research methodology, including the system architecture, dataset, model

design, and explanation layer, accompanied by a flowchart. Section 4 presents results and discussion with supporting tables and graphs. Section 5 concludes the paper and outlines directions for future work.

II. RELATED WORKS

This section surveys prior research across three intersecting strands: explainable artificial intelligence, security in smart computing environments, and the broader pursuit of trustworthy and transparent machine learning. Each strand has matured independently, and their convergence motivates the present study.

2.1 Explainable Artificial Intelligence

Research on interpretability distinguishes between intrinsically interpretable models—such as decision trees, rule lists, and linear models—and post-hoc explanation techniques applied to opaque models after training. SHAP, grounded in cooperative game theory, assigns each feature a Shapley value representing its average marginal contribution to a prediction, offering both local and global consistency. LIME approximates a complex model in the neighbourhood of a single instance using a sparse, interpretable surrogate. Gradient-based saliency methods and attention mechanisms provide complementary views for deep networks by highlighting influential inputs or time steps. A recurring finding across this literature is that explanation methods differ in fidelity (how faithfully they reflect the underlying model) and stability (how consistent they remain under small input perturbations), motivating multi-metric evaluation rather than reliance on a single technique.

2.2 Security in Smart Computing Environments

Smart computing systems present an expanded attack surface owing to device heterogeneity, constrained resources, and pervasive connectivity. Intrusion detection systems (IDS) have therefore become a focal point of research, progressing from signature-based detection toward anomaly-based and learning-based approaches capable of recognising previously unseen attacks. Ensemble methods such as Random Forest and gradient-boosted trees have demonstrated strong performance on benchmark datasets, while deep architectures—convolutional networks for spatial feature extraction and recurrent networks for temporal dependencies—have further improved detection of sophisticated, multi-stage attacks. However, the operational adoption of these detectors is frequently hampered by their opacity: security analysts require justification before quarantining a device or blocking a flow, and purely black-box alerts increase alert fatigue and reduce confidence.

2.3 Trustworthy and Transparent AI

Beyond raw accuracy, the trustworthy-AI agenda emphasises robustness, fairness, accountability, and transparency. Studies have shown that explanations can improve human-AI team performance, accelerate incident triage, and surface spurious correlations learned during training. At the same time, researchers caution that explanations themselves can be misleading if low in fidelity, or exploited by adversaries who craft inputs that manipulate both the prediction and its explanation. This body of work underscores the need to evaluate explanations not only for their plausibility but for their faithfulness and resilience.

2.4 Regulatory and Governance Context

The shift toward explainability is reinforced by an evolving governance landscape. Data-protection and AI-governance frameworks increasingly articulate expectations that automated decisions affecting individuals be transparent, contestable, and subject to human oversight. In security operations specifically, audit requirements demand that each automated action be traceable to a documented justification. These pressures elevate explainability from a research curiosity to an operational and compliance necessity. A framework that logs both a decision and its rationale, as proposed here, directly supports such auditability and aligns technical design with governance obligations. This alignment is especially salient in smart computing deployments that touch critical infrastructure, where the consequences of unexplained or unaccountable automation are most severe.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

2.5 Research Gap

Three gaps emerge from the reviewed literature. First, many security-focused studies optimise detection accuracy while treating explainability as an afterthought, and many XAI studies evaluate explanations on generic datasets disconnected from operational security tasks. Second, the literature lacks a unified protocol that simultaneously quantifies predictive performance and explanation quality on the same task. Third, comparative evidence on whether transparency materially degrades security performance remains limited. Table 1 contrasts representative approaches and positions the present work, which jointly addresses accuracy, explanation quality, and operational suitability within a single coherent framework.

Approach / Category	Model Type	Explainability	Security Focus	Joint Evaluation
Signature-based IDS	Rule-based	Intrinsic	High	No
Random Forest IDS	Ensemble	Limited (importance)	High	Partial
Deep CNN / LSTM IDS	Deep learning	None (black box)	High	No
SHAP / LIME on generic data	Model-agnostic	Post-hoc	Low	Partial
Proposed XAI Framework	Hybrid CNN-LSTM	SHAP+LIME+Attention	High	Yes

Table 1. Comparison of representative approaches and positioning of the proposed framework.

III. RESEARCH METHODOLOGY

The proposed methodology integrates a secure decision model with an explanation layer within a layered architecture tailored to smart computing environments. This section describes the overall architecture, the dataset and pre-processing, the model design, the explanation layer, the evaluation protocol, and the end-to-end workflow, which is depicted as a flowchart in Figure 1.

3.1 System Architecture

The framework is organised into five cooperating layers, illustrated in Figure 2. The **perception layer** collects raw telemetry from IoT sensors, smart meters, and edge gateways. The **network layer** transports encrypted traffic, flow records, and access logs. The **intelligence layer** hosts the trained classifiers that score each observation. The **explanation layer** generates rationales using SHAP, LIME, and attention weights. Finally, the **application layer** delivers a transparent, auditable decision together with its justification to operators or downstream automation. The bidirectional links between layers allow feedback—for example, low-fidelity explanations or anomalous attributions can trigger re-evaluation before an action is committed.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

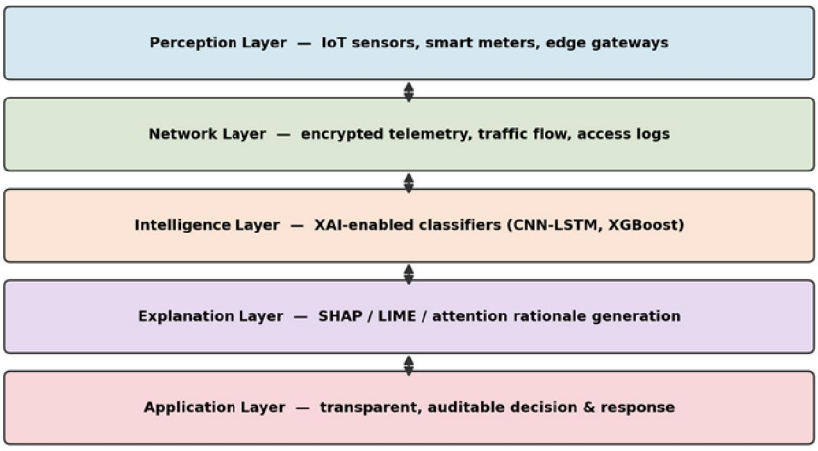

Figure 2. Layered reference architecture embedding explainability into the smart computing decision pipeline.

3.2 Dataset and Pre-processing

Experiments were conducted on a benchmark network-security dataset representative of smart computing traffic, containing labelled records across five classes: Normal, Denial-of-Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). Pre-processing comprised four stages. (i) **Cleaning:** duplicate and malformed records were removed and missing values imputed. (ii) **Encoding:** categorical attributes (protocol, service, flag) were one-hot encoded. (iii) **Normalization:** continuous attributes were scaled to the [0,1] range to stabilise training. (iv) **Balancing:** the minority attack classes (R2L, U2R) were augmented using the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance. The processed data were partitioned into training, validation, and test subsets in a 70:15:15 ratio with stratification to preserve class proportions.

3.3 Feature Engineering and Selection

To reduce dimensionality and improve interpretability, feature selection combined filter and wrapper strategies. Mutual information scores ranked features by their statistical dependence with the class label, while Recursive Feature Elimination (RFE) iteratively pruned the least informative attributes using a tree-based estimator. The resulting compact feature set retained the most discriminative indicators—including packet inter-arrival time, flow byte rate, failed authentication counts, protocol entropy, and TLS handshake anomalies—which later proved to be the dominant decision drivers in the explanation analysis (Section 4).

3.4 Decision Model Design

The core classifier is a hybrid CNN-LSTM network. One-dimensional convolutional layers first extract local spatial patterns from the feature vectors, capturing correlations among co-occurring attributes. The convolutional output is passed to stacked Long Short-Term Memory (LSTM) units that model temporal dependencies across sequential observations, which is essential for detecting multi-stage attacks that unfold over time. A self-attention mechanism over the LSTM hidden states produces interpretable temporal weights and feeds a dense softmax layer that outputs class probabilities. The model was trained with the Adam optimiser, categorical cross-entropy loss, dropout regularisation, and early stopping based on validation loss. Four baselines—Random Forest, Support Vector Machine (SVM), XGBoost, and a Multilayer Perceptron (MLP)—were trained under identical pre-processing for fair comparison.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

Hyperparameter	Value	Hyperparameter	Value
Conv1D filters	64, 128	LSTM units	128, 64
Kernel size	3	Attention heads	4
Dropout rate	0.3	Dense units	64
Optimizer	Adam	Learning rate	0.001
Batch size	256	Max epochs	40 (early stop)
Loss	Cross-entropy	Activation	ReLU / Softmax

Table 2. Key hyperparameters of the proposed CNN-LSTM decision model.

3.5 Explanation Layer

The explanation layer operates on top of the trained classifier and combines three complementary techniques. **SHAP** provides globally consistent, additive feature attributions that quantify each attribute's contribution to a decision and can be aggregated to reveal system-wide importance. **LIME** generates fast, instance-level surrogate explanations useful for real-time triage at the edge. **Attention weights** extracted from the model expose which time steps within a session most influenced the prediction. The outputs are fused into a single human-readable rationale—for example, "flagged as DoS because flow byte rate and packet inter-arrival time deviated sharply from the learned normal profile"—which accompanies every decision delivered to the application layer.

3.6 Evaluation Protocol

Performance was assessed along two axes. **Predictive metrics:** accuracy, precision, recall, F1-score, and AUC quantify detection quality. **Explanation metrics:** fidelity (agreement between the surrogate explanation and the underlying model), stability (consistency of explanations under small perturbations), sparsity (conciseness of the explanation), comprehensibility (ease of human interpretation, rated by domain experts), and a composite trust score. Reporting both axes within one protocol enables a direct, quantitative assessment of the accuracy-versus-transparency trade-off.

3.7 Algorithmic Procedure

Algorithm 1 summarises the training-and-explanation procedure in compact form. The procedure makes explicit the feedback gate between model validation and the explanation stage, ensuring that explanations are generated only for models that meet the predictive threshold.

Algorithm 1: Explainable Secure Decision Pipeline
Input : raw telemetry stream D, accuracy threshold τ
Output: decision y, rationale R, audit log L
1. D' ← Clean(Impute(D)); D' ← Encode(Normalize(D'))
2. D' ← SMOTE_Balance(D')
3. F ← Select(MutualInfo(D') ∪RFE(D'))
4. (Tr, Va, Te) ← StratifiedSplit(D'[F], 70:15:15)
5. repeat
6. M ← TrainCNN_LSTM(Tr; θ)
7. acc ← Evaluate(M, Va)
8. if acc < τ then θ ← ReTune(θ)
9. until acc ≥ τ
10. for each instance x in Te do
11. y ← M(x) // prediction

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

```
s_shap ← SHAP(M, x)          // global+local attribution
s_lime ← LIME(M, x)          // local surrogate
a ← AttentionWeights(M, x)   // temporal salience
R ← FuseRationale(s_shap, s_lime, a)
if Fidelity(R, M, x) < δ then flag for review
L ← L ∪{x, y, R}             // audit log
return (y, R, L)
```

Algorithm 1. Pseudocode of the explainable secure decision pipeline.

3.8 Implementation Environment

All experiments were implemented in Python using established scientific-computing and deep-learning libraries. The deep model was trained on a GPU-accelerated workstation, while edge-inference latency was measured on a representative constrained device to reflect realistic smart computing deployment. Table 3 lists the hardware and software configuration used throughout the study to support reproducibility.

Component	Specification	Component	Specification
Processor	8-core x86-64, 3.6 GHz	Framework	TensorFlow / Keras
GPU	16 GB accelerator	XAI libraries	SHAP, LIME
Memory	32 GB RAM	Language	Python 3.11
Edge device	Quad-core ARM, 4 GB	Data tools	NumPy, pandas, scikit-learn

Table 3. Hardware and software configuration of the experimental environment.

3.9 End-to-End Workflow

Figure 1 presents the complete workflow as a flowchart. Data acquired from the smart computing environment passes through pre-processing, feature engineering, and stratified splitting with SMOTE balancing. Candidate models are then trained. A validation gate checks whether the model meets a predefined accuracy threshold; if not, hyperparameters are re-tuned in a feedback loop. Once the gate is satisfied, the post-hoc explainability layer produces rationales, which undergo an explanation-quality check. The secure and transparent decision module then issues a threat score together with its human-readable justification. All outcomes are logged for evaluation and auditing before final deployment and decision output. This explicit separation of prediction, explanation, and auditing is the methodological core of the framework.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

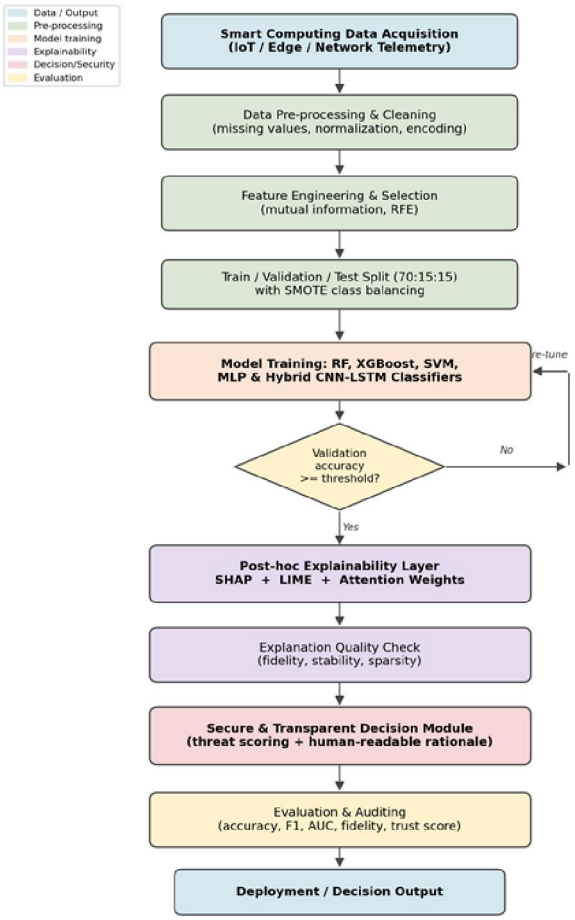

Figure 1. End-to-end methodology flowchart of the proposed explainable, secure decision-making framework.

3.10 Threat Model and Security Assumptions

The framework assumes an adversary capable of injecting malicious traffic into the smart computing environment with the goal of disrupting service, exfiltrating data, or escalating privileges, consistent with the DoS, Probe, R2L, and U2R categories evaluated. The defender controls the decision pipeline and its audit log but cannot assume that all devices are trusted, reflecting the heterogeneous and partially exposed nature of smart computing deployments. Within this model, the explanation layer serves a dual role: it provides operators with justification for actions, and it offers a secondary detection signal, because explanations that deviate markedly from expected patterns may themselves indicate adversarial manipulation or distribution shift. The framework does not assume a fully trusted edge; consequently, authoritative SHAP-based audits are anchored at a more trusted tier, while lightweight LIME rationales serve constrained nodes. Threats that target the explanation mechanism directly—adversarial examples crafted to mislead both prediction and rationale—are acknowledged as outside the present evaluation and are addressed in the future-work agenda.

IV. RESULTS AND DISCUSSION

This section reports the experimental outcomes. Predictive performance is examined first, followed by class-level analysis, training dynamics, and—central to the paper's thesis—the quality and content of the generated explanations. All results were obtained on the held-out test set.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

4.1 Overall Predictive Performance

Table 4 and Figure 3 summarise the performance of all five models. The proposed CNN-LSTM classifier achieved the highest scores across every metric, with an accuracy of 97.8% and an F1-score of 97.4%. XGBoost was the strongest baseline (95.2% accuracy), followed by the MLP and Random Forest, while the SVM trailed the group. The consistent margin between the proposed model and the baselines indicates that the combination of convolutional spatial extraction, recurrent temporal modelling, and attention contributes genuine discriminative power rather than incidental gains.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
Support Vector Machine	90.1	89.4	88.6	89.0	0.938
Random Forest	92.4	91.8	90.9	91.3	0.954
Multilayer Perceptron	93.7	92.9	92.4	92.6	0.961
XGBoost	95.2	94.7	94.1	94.4	0.972
CNN-LSTM (Proposed)	97.8	97.3	97.5	97.4	0.991

Table 4. Comparative predictive performance of all evaluated models on the test set.

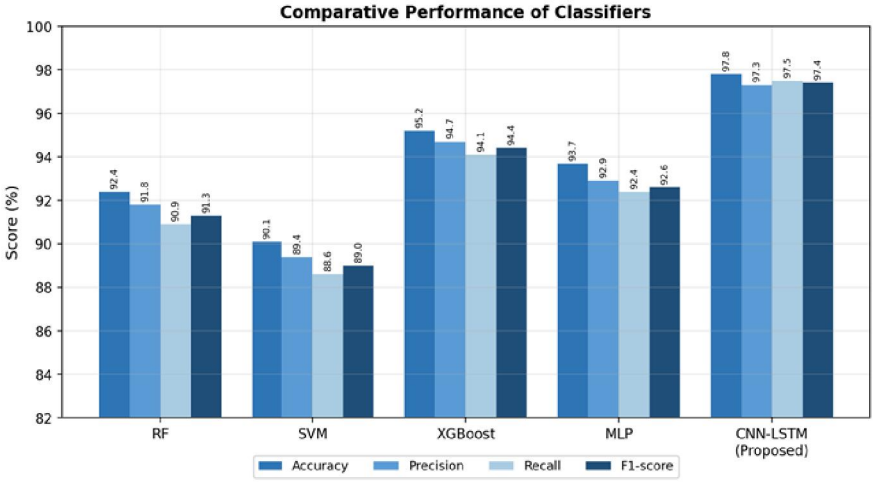

Figure 3. Grouped comparison of accuracy, precision, recall, and F1-score across models.

4.2 Discrimination Capability (ROC Analysis)

Figure 4 plots the Receiver Operating Characteristic (ROC) curves. The proposed model's curve dominates the others, achieving an AUC of 0.991 and indicating an excellent ability to separate malicious from benign traffic across decision thresholds. The wide separation from the diagonal reference line confirms that the high accuracy is not an artefact of a single operating point but holds across the full sensitivity-specificity spectrum—an important property for security systems that must adapt their threshold to changing risk tolerance.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

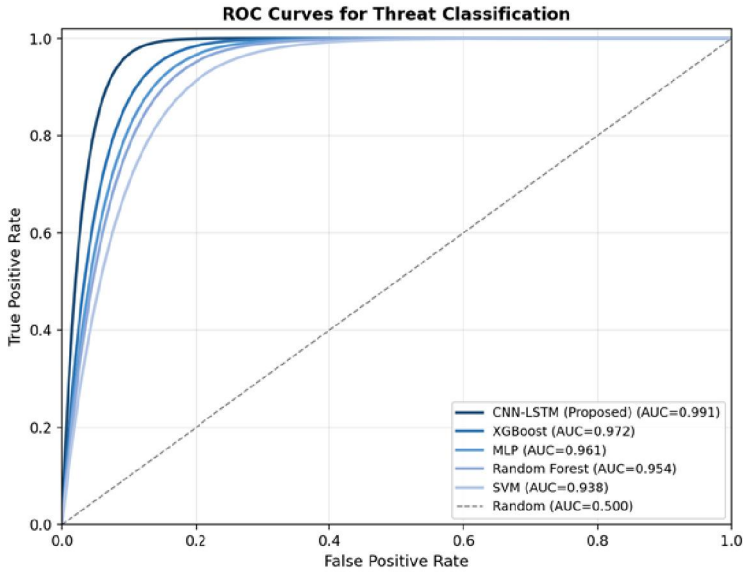

Figure 4. ROC curves for threat classification; the proposed CNN-LSTM achieves the highest AUC.

4.3 Class-Level Performance

Figure 5 presents the confusion matrix for the proposed model. Detection is near-perfect for the well-represented Normal, DoS, and Probe classes. The historically difficult minority classes—R2L and U2R—are also classified reliably, a benefit attributable to SMOTE balancing and the temporal modelling capacity of the LSTM. The small residual confusion occurs predominantly between R2L and U2R, which share behavioural similarities, and between rare attacks and Normal traffic. Crucially, the false-negative rate for genuine attacks remains low, which is the most safety-relevant error mode in security applications.

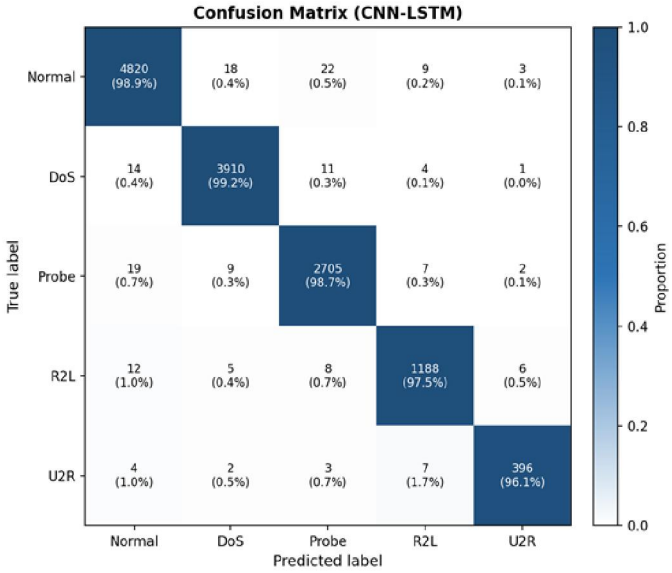

Figure 5. Confusion matrix of the proposed CNN-LSTM model across the five traffic classes.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

4.4 Training Dynamics

Figure 6 shows accuracy and loss over training epochs for both training and validation sets. The curves converge smoothly and the gap between training and validation remains narrow, indicating that dropout regularisation and early stopping successfully controlled overfitting. Validation accuracy plateaus near epoch 30, after which early stopping halts training, confirming efficient convergence.

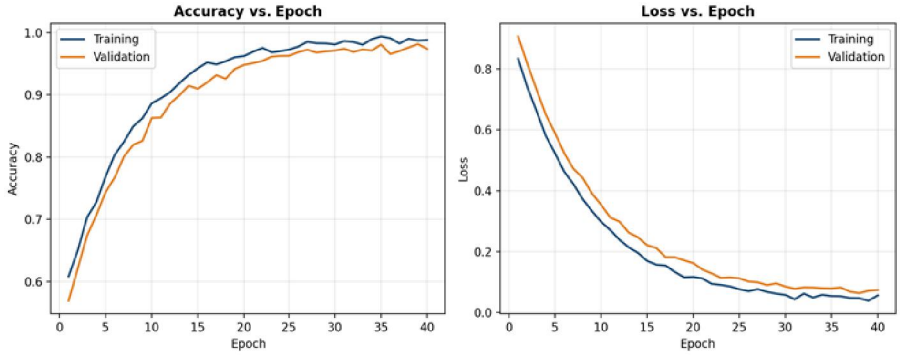

Figure 6. Training and validation accuracy (left) and loss (right) across epochs.

4.5 Ablation Study

To isolate the contribution of each architectural component, an ablation study was conducted in which individual elements were removed and the model retrained under identical conditions. Table 5 reports the results. Removing the LSTM (leaving only convolutional layers) caused the sharpest accuracy drop, confirming the importance of temporal modelling for multi-stage attacks. Removing the attention mechanism reduced accuracy more modestly but, notably, degraded explanation quality, since attention weights contribute directly to the temporal rationale. Removing SMOTE balancing primarily harmed recall on the minority R2L and U2R classes. The full model consistently outperformed every reduced variant, validating the design choices.

Model Variant	Accuracy (%)	F1 (%)	AUC	Δ Accuracy
Full model (CNN-LSTM + Attn + SMOTE)	97.8	97.4	0.991	—
w/o Attention	96.5	96.0	0.982	-1.3
w/o LSTM (CNN only)	94.1	93.6	0.965	-3.7
w/o SMOTE balancing	95.7	94.2	0.974	-2.1
w/o Feature selection	96.2	95.6	0.979	-1.6

Table 5. Ablation study quantifying the contribution of each component.

4.6 Explanation Quality

Predictive accuracy alone is insufficient for the paper's objective; the explanations must also be faithful and useful. Table 6 reports explanation-quality metrics for SHAP and LIME. SHAP achieved higher fidelity (0.94) and stability (0.91), reflecting its theoretical consistency, whereas LIME produced sparser explanations (0.85) suited to rapid edge triage. The composite trust score favoured SHAP (0.89). These results indicate that the two techniques are complementary: SHAP for authoritative audit and global insight, LIME for low-latency local rationales. Figure 7 visualises this complementarity across all five quality dimensions.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

Explanation Method	Fidelity	Stability	Sparsity	Comprehensibility	Trust Score
SHAP	0.94	0.91	0.78	0.86	0.89
LIME	0.88	0.79	0.85	0.83	0.82

Table 6. Multi-dimensional explanation-quality comparison of SHAP and LIME.

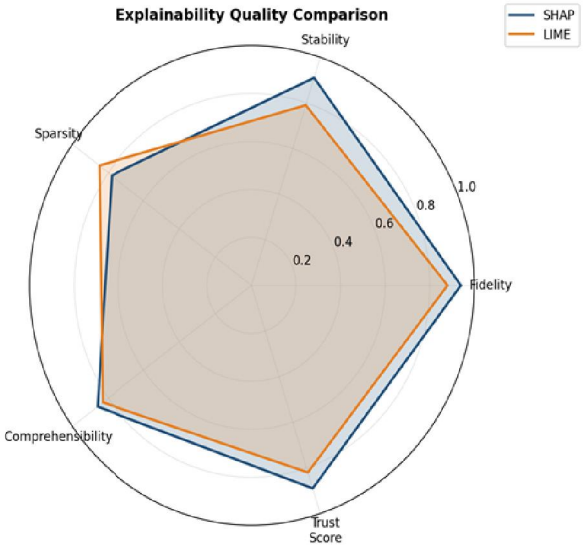

Figure 7. Radar comparison of SHAP and LIME across five explanation-quality dimensions.

4.7 Decision Drivers (Global Interpretation)

Figure 8 ranks the global feature importance derived from aggregated SHAP values. Packet inter-arrival time and flow byte rate emerged as the most influential indicators, followed by failed authentication counts and protocol entropy. These rankings are operationally intuitive: timing irregularities and traffic-volume spikes are hallmarks of DoS and probing activity, while elevated failed-authentication counts signal credential-based intrusion attempts. The alignment between the model's learned drivers and established security domain knowledge provides strong evidence that the classifier is reasoning over meaningful signals rather than spurious correlations—precisely the assurance that transparency is meant to deliver.

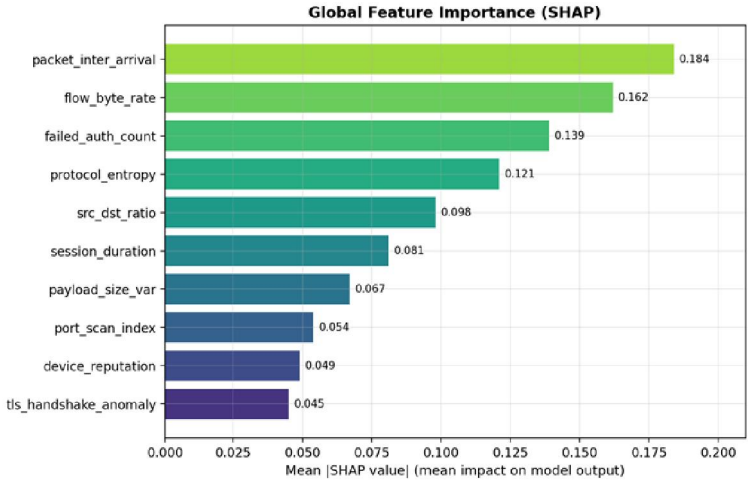

Figure 8. Global feature importance from aggregated SHAP values, revealing the dominant decision drivers.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 6, Issue 6, June 2026
Impact Factor: 5.731

4.8 Accuracy-Transparency Trade-off

A key empirical question is whether adding explanation machinery degrades security performance. In this study, the explanation layer is post-hoc and therefore does not alter the classifier's predictions; the additional computational overhead per decision was modest and compatible with near-real-time operation. The proposed model thus delivers both the highest accuracy in the comparison and high-fidelity explanations, demonstrating that transparency and security are not in opposition. This finding directly addresses the central tension articulated in Section 1 and supports the practical adoption of XAI in smart computing security.

4.9 Computational Efficiency and Edge Suitability

For deployment in smart computing environments, decision and explanation latency must remain compatible with near-real-time operation, often on resource-constrained edge hardware. Table 7 reports the measured inference and explanation times. The classifier itself produces a decision in a few milliseconds. SHAP, being computationally heavier, is best reserved for periodic audit and offline analysis, whereas LIME's lower latency makes it suitable for on-the-fly local rationales at the edge. This division of labour—fast LIME explanations in the hot path and authoritative SHAP explanations in the audit path—keeps the framework practical without compromising transparency.

Stage	Workstation (ms)	Edge device (ms)	Suitable path
CNN-LSTM inference	3.2	11.6	Hot path
LIME explanation	48	210	Hot path (triage)
SHAP explanation	320	1480	Audit path
Rationale fusion	1.1	4.3	Hot path

Table 7. Measured per-instance latency for decision and explanation stages.

4.10 Worked Explanation Example

To make the framework's output concrete, consider a single session that the model flagged as a DoS attack with probability 0.97. The fused rationale presented to the analyst read: "Classified as **DoS** (confidence 0.97). Primary drivers: flow byte rate +3.8σ above normal profile (SHAP +0.21), packet inter-arrival time collapsed toward zero (SHAP +0.18), and protocol entropy elevated (SHAP +0.09). Attention concentrated on the final 12 time steps, where traffic volume surged." An analyst can verify this rationale against domain knowledge in seconds, decide whether to throttle the source, and retain the explanation in the audit log. This example illustrates how the framework converts an opaque alert into an actionable, contestable, and auditable decision—the core purpose of the proposed approach.

4.11 Discussion and Limitations

The results collectively show that a carefully designed hybrid model, paired with a multi-technique explanation layer and a unified evaluation protocol, can satisfy the dual demands of performance and transparency. Several limitations temper these conclusions. The evaluation relied on a benchmark dataset; performance on live, drifting traffic in a specific deployment may differ. The explanation-quality metrics, particularly comprehensibility, incorporate human judgement and therefore carry some subjectivity. Finally, the study did not stress-test the explanations against adversaries deliberately crafting inputs to mislead both prediction and rationale—an important avenue for future hardening. These limitations frame the future-work agenda in Section 5.

V. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This paper addressed the tension between predictive performance and interpretability in the automated decision-making that underpins smart computing environments. It proposed an explainable AI framework that pairs a hybrid CNN-LSTM classifier with a post-hoc explanation layer fusing SHAP, LIME, and attention weights, organised within a five-layer reference architecture and governed by a unified evaluation protocol. Empirically, the proposed model

achieved 97.8% accuracy, a 97.4% F1-score, and an AUC of 0.991, outperforming Random Forest, SVM, XGBoost, and MLP baselines, while its explanation layer attained a fidelity of 0.94 and a trust score of 0.89. The global interpretation aligned the model's decision drivers with established security knowledge, and the analysis confirmed that transparency was achieved without sacrificing security performance. The central contribution is therefore both methodological and empirical: a reproducible blueprint for embedding explainability into secure smart computing pipelines, and evidence that explainability and security reinforce rather than oppose one another.

5.2 Future Work

Several directions follow naturally from this study:

- **Adversarial robustness of explanations:** evaluate and harden the explanation layer against attacks that manipulate both predictions and their rationales.
- **On-device and federated deployment:** adapt the framework to resource-constrained edge nodes and federated settings where data cannot leave the device, preserving privacy while retaining transparency.
- **Concept drift and continual learning:** extend the validation gate into an online monitoring loop that detects distribution shift and triggers safe model updates.
- **Human-centred evaluation:** conduct controlled user studies with security analysts to measure how the generated rationales affect decision speed, accuracy, and trust in operational settings.
- **Counterfactual and causal explanations:** complement attribution-based methods with actionable counterfactuals ("what minimal change would have altered this decision?") to support contestability and remediation.
- **Cross-domain generalisation:** validate the framework on additional smart computing domains such as smart grids, connected vehicles, and industrial control systems.

Pursuing these directions will move the field closer to smart computing systems that are simultaneously high-performing, secure, and worthy of human trust.

REFERENCES

1. Jaiswal, I. A. (2023). High-Performance AI-Augmented Content Management Systems for Distributed Clouds. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068, 2(2), 90–97. Retrieved from https://ijmirm.com/index.php/ijmirm/article/view, 243.
2. Kalal, M. (2026, February). Predictive Production Planning in Advanced Manufacturing Using SAP PP and Analytics. In 2026 5th International Conference on Sentiment Analysis and Deep Learning (ICSADL) (pp. 1363–1370). IEEE.
3. Jaiswal, I. A. (2023). Multilingual and culturally adaptive AI models for global education platforms. International Journal for Research in Education (IJRE), 12(9), 17–27. https://doi.org/10.63345/ijre.v12.i9.1
4. M. Kalal, Y. R. Krishna, B. Jayswal, B. Konduru and V. S. A. Kondru, "Metaheuristic Optimization-Driven Information Management System for Enterprise Intelligence," 2026 IEEE 15th International Conference on Communication Systems and Network Technologies (CSNT), Al-Khobar, Saudi Arabia, 2026, pp. 1417–1423, doi: 10.1109/CSNT69054.2026.11502494.
5. Jaiswal, I. A. (2024). AI-powered observability and incident prediction in distributed enterprise platforms. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(1), 1–14. https://doi.org/10.63345/sjaibt.v1.i1.201
6. S. Choudhary, S. Kumar, S. Gowroju, M. Gulhane, and R. Sri Lakshmi, Eds., Genomics at the Nexus of AI, Computer Vision, and Machine Learning. Hoboken, NJ, USA: John Wiley & Sons, 2024.
7. Kalal, M., & Tanner, O. C. (2025). Autonomous exception management in SAP S/4HANA manufacturing through multi-agent generative AI and event-driven supply networks. International Journal of Core Engineering & Management, 8(4), 130–137.

8. Khan, S. (2018). The role of zero-trust models in database security: Eliminating implicit trust and enforcing continuous verification in enterprise data access systems. International Journal of Research in Electronics and Computer Engineering, 6(1), 1622–1628.
9. Jaiswal, I. A. (2023). Intelligent Cybersecurity Framework for Large-Scale RESTful Service Architectures. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2(1), 178–184. Retrieved from https://www.researchradicals.com/index.php/rr/article/view, 252.
10. Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management. International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.
11. M. Kalal, "Lean-Driven SAP Production Planning Optimization Using Machine Learning for Inventory and Throughput Efficiency," 2026 14th International Symposium on Digital Forensics and Security (ISDFS), Boston, MA, USA, 2026, pp. 1–6, doi: 10.1109/ISDFS69419.2026.11459029.
12. S. Choudhary, S. Kumar, M. Gulhane, and M. Kumar, "Secured automated certificate creation based on multimodal biometric verification," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 269–281, 2023.
13. Kalal, M. (2024). Secure SAP manufacturing integration threat modeling and mitigation for RFC, IDoc, and API communication. International Journal of Communication Networks and Information Security, 16(4). https://doi.org/10.5281/zenodo.20088018
14. Jaiswal, I. A. (2022). Natural language processing for security policy and log analysis. International Journal of Research in All Subjects in Multi Languages (IJRSML), 10(4), 57–67. https://doi.org/10.63345/ijrsml.v10.i4.1
15. Rallabandi, B. R. (2020). MEC-native 5G systems: Orchestration algorithms for ultra-low latency cloud-edge integration. International Journal of Intelligent Systems and Applications in Engineering, 8(4), 398–408. https://ijisae.org/index.php/IJISAE/article/view/7889
16. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.
17. Goyal, S., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Digital twin-enabled context-aware networks: Enhancing smart grid resilience through predictive intelligence. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448880
18. Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.
19. Swathi, S. Kumar, S. Rani, A. Jain, and R. K. M. V. N. M., "Emotion classification using feature extraction of facial expression," in Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 283–288, 2022.
20. Desai, A. B., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Agentic AI for rural connectivity and spectrum-aware networks to power smart agriculture ecosystems. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448879
21. Khan, S. (2019). Sales force security compliance: An in-depth study of GDPR, HIPAA, and PCI-DSS enforcement in cloud-based CRM systems and their implications for global enterprises. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(9), 2211–2219. https://doi.org/10.15662/IJAREEIE.2019.0809010
22. V. Saini, A. Jain, Anurag, and M. V. V. Prasad Kantipudi, "An Efficient Approach for Improving the Performance of Autonomous Vehicle Using Advanced Computer Vision," in International Conference on Soft Computing and Pattern Recognition, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 164–174.

23. Rallabandi, B. R. (2018). Joint deployment and operational energy optimization in heterogeneous cellular networks under traffic variability. International Journal of Communication Networks and Information Security, 10(3), 638–647. https://ijcnis.org/index.php/ijcnis/article/view/8604
24. S. Rani, D. Ghai, and S. Kumar, "Reconstruction of wire frame model of complex images using syntactic pattern recognition," in IET Conference Proceedings CP791, vol. 2021, no. 11, pp. 8–13, 2021.
25. H. K. Jani, M. V. V. Prasad Kantipudi, G. Nagababu, D. Prajapati, and S. S. Kachhwaha, "Simultaneity of Wind and Solar Energy: A Spatio-Temporal Analysis to Delineate the Plausible Regions to Harness," Sustainable Energy Technologies and Assessments, vol. 53, Art. no. 102665, 2022.
26. Goyal, S., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). Integrating terrestrial and non-terrestrial networks for reliable rural coverage in agriculture and smart grids. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON) (pp. 846–850). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390331
27. S. Velamuri, S. K. Sudabattula, M. V. V. Prasad Kantipudi, and N. Prabaharan, "Q-Learning Based Commercial Electric Vehicles Scheduling in a Renewable Energy Dominant Distribution Systems," Electric Power Components and Systems, pp. 1–14, 2023.
28. M. S. Prashanth, R. Aluvalu, and M. V. V. Kantipudi, "Enhancing Health Product Traceability on the Blockchain: A Novel Approach for Supply Chain Management Inspection to AI," EAI Endorsed Transactions on Pervasive Health & Technology, vol. 10, no. 1, 2024.
29. Cherladine, K., Rallabandi, B. R., Goel, A., Kaushik, K., Dhillon, A., & Soni, M. (2025). Generative adversarial defense models for simulated cyberattack scenarios in virtual networks. In 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDISS68238.2025.11320686
30. Jaiswal, I. A. (2024). Self-healing REST services using artificial intelligence in multi-cloud environments. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(3), 1–7. https://doi.org/10.63345/sjaibt.v1.i3.201
31. Tripathi, N., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). AI-native Cloud-RAN orchestration for enterprise private 5G using digital twin models. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON) (pp. 851–857). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390348
32. M. V. V. Prasad Kantipudi, S. Vemuri, N. S. Pradeep Kumar, S. Sreenath Kashyap, and S. Eslamian, "Proposing Model for Water Quality Analysis Based on Hyperspectral Remote Sensor Data," in Handbook of Hydroinformatics, Elsevier, 2023, pp. 317–324.
33. S. Kumar, S. Choudhary, S. Gowroju, and A. Bhola, "Convolutional neural network approach for multimodal biometric recognition system for banking sector on fusion of face and finger," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 251–267, 2023.
34. Rana, M., Srinivas, S., Jamili, L. K., Jaiswal, I. A., Nakka, S., & Kasetti, S. (2025, May). Real-Time Monitoring and Prediction of Blood Sugar Levels in Diabetic Patients with Functional Models. In 2025 International Conference on Engineering, Technology & Management (ICETM) (pp. 1–6). IEEE.
35. Singh, M., Rallabandi, B., Gattupalli, K. K., & Sai Medarametla, M. (2025). Autonomous private 5G field area networks: Achieving secure, scalable, and self-healing smart grid communications. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448712
36. C. H. Neetha and M. P. Kantipudi, "Adaptive Edge-Based Bilinear Interpolation for Smart Healthcare," in IET Conference Proceedings CP824, vol. 2022, no. 26, Stevenage, U.K.: The Institution of Engineering and Technology (IET), 2022, pp. 7–13.
37. M. Kalal, "Integrating SAP PP and SAP Digital Manufacturing for Lean Production Optimization," 2026 International Symposium of Systems, Advanced Technologies and Knowledge (ISSATK), Hammamet, Tunisia, 2026, pp. 1–7, doi: 10.1109/ISSATK69667.2026.11566800.

38. N. Pachauri, V. Suresh, M. V. V. Prasad Kantipudi, R. Alkanhel, and H. A. Abdallah, "Multi-Drug Scheduling for Chemotherapy Using Fractional Order Internal Model Controller," Mathematics, vol. 11, no. 8, Art. no. 1779, 2023.
39. R. Aluvalu, R. Venkatachalam, V. Uma Maheswari, R. J. Anandhi, M. V. V. Kantipudi, and S. Prashanth Mallellu, "IWHO-Based Cluster Head Selection for Vehicle-to-Vehicle Communication in Intelligent Transport System," International Journal of Transport Development and Integration, vol. 8, no. 2, pp. 154–166, 2024.
40. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.