IN THIS JOURNAL

Great Britain Journals Press

journalspress.com

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Print ISSN 2514-863X
Online ISSN 2514-8648
DOI 10.17472/LJRCST

COMPILED IN UNITED KINGDOM

ENGLISH

London Journal of Research in Computer Science and Technology

Volume 25 | Issue 5 | Compilation 1.0

PUBLISHER

Great Britain Journals Press
1210th, Waterside Dr, Opposite Arlington Building, Theale, Reading
Phone:+444 0118 965 4033 Pin: RG7-4TY United Kingdom

SUBSCRIPTION

Frequency: Quarterly

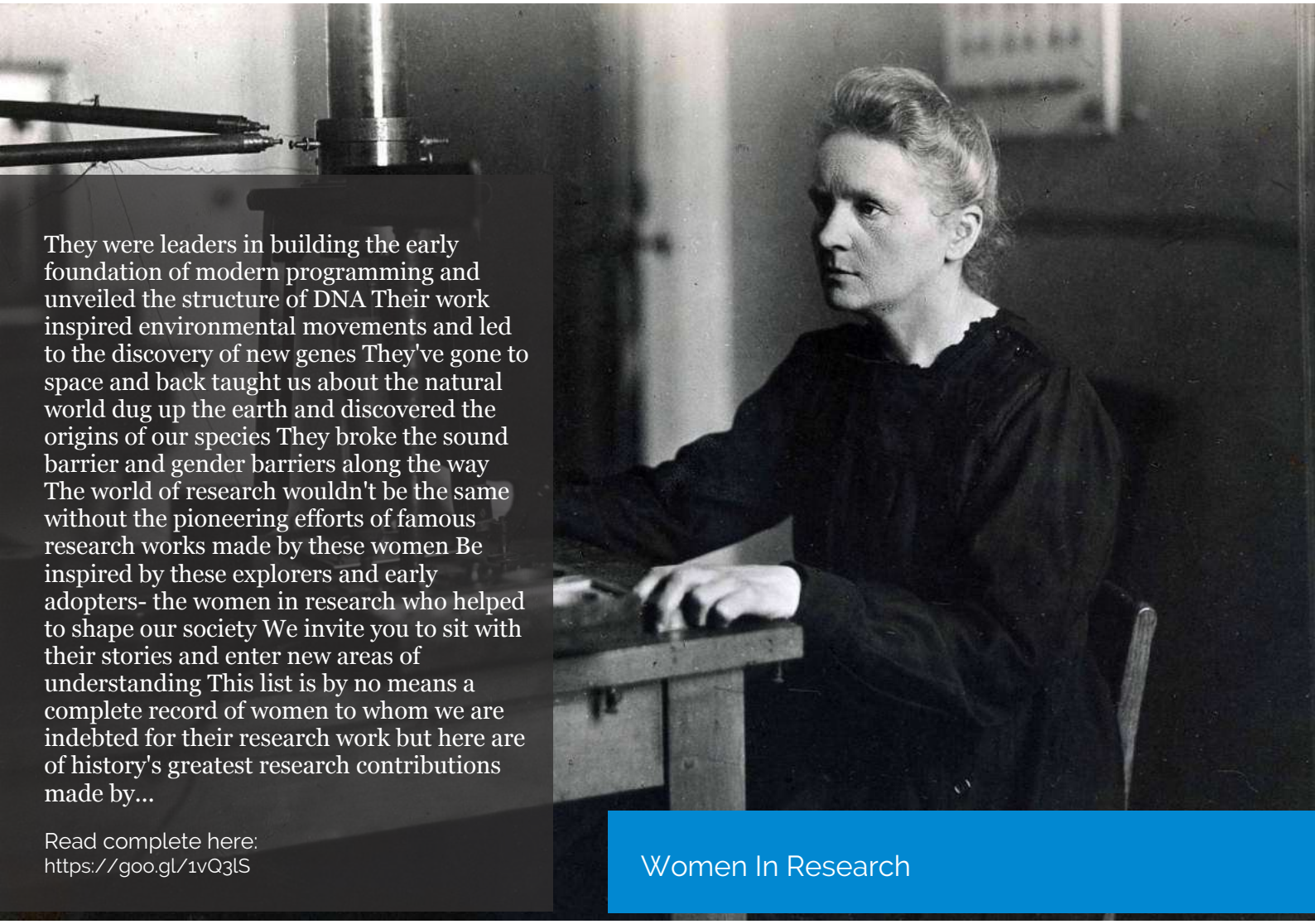
Print subscription
$280USD for 1 year
$500USD for 2 year
(*color copies including taxes and international shipping with TSA approved)*
Find more details at https://journalspress.com/journals/subscription

ENVIRONMENT

Great Britain Journals Press is intended about Protecting the environment. This journal is printed using led free environmental friendly ink and acid–free papers that are 100% recyclable.

Great Britain Journals Press Headquaters

1210th, Waterside Dr,
Opposite Arlington
Building, Theale, Reading
Phone:+444 0118 965 4033
Pin: RG7-4TY
United Kingdom



They were leaders in building the early foundation of modern programming and unveiled the structure of DNA Their work inspired environmental movements and led to the discovery of new genes They've gone to space and back taught us about the natural world dug up the earth and discovered the origins of our species They broke the sound barrier and gender barriers along the way The world of research wouldn't be the same without the pioneering efforts of famous research works made by these women Be inspired by these explorers and early adopters- the women in research who helped to shape our society We invite you to sit with their stories and enter new areas of understanding This list is by no means a complete record of women to whom we are indebted for their research work but here are of history's greatest research contributions made by...

Read complete here:
https://goo.gl/1vQ3lS

Women In Research

Writing great research...

Prepare yourself before you start Before you start writing your paper or you start reading other...

Read complete here:
https://goo.gl/qKfHht

Computing in the cloud!

Cloud Computing is computing as a Service and not just as a Product Under Cloud Computing...

Read complete here:
https://goo.gl/H3EWK2

Journal Content
In this Issue

Great Britain Journals Press

Editorial Board

Curated board members

Dr. Saad Subair

College of Computer and Information Sciences, Association Professor of Computer Science and Information System Ph.D., Computer Science-Bioinformatics, University of Technology Malaysia

Gerhard X Ritter

Emeritus Professor, Department of Mathematics, Dept. of Computer & Information, Science & Engineering Ph.D., University of Wisconsin-Madison, USA

Dr. Ikvinderpal Singh

Assistant Professor, P.G. Deptt. of Computer Science & Applications, Trai Shatabdi GGS Khalsa College, India

Prof. Sergey A. Lupin

National Research, University of Electronic Technology Ph.D., National Research University of Electronic Technology, Russia

Dr. Sharif H. Zein

School of Engineering, Faculty of Science and Engineering, University of Hull, UK Ph.D., Chemical Engineering Universiti Sains Malaysia, Malaysia

Prof. Hamdaoui Oualid

University of Annaba, Algeria Ph.D., Environmental Engineering, University of Annaba, University of Savoie, France

Prof. Wen Qin

Department of Mechanical Engineering, Research Associate, University of Saskatchewan, Canada Ph.D., Materials Science, Central South University, China

Luisa Molari

Professor of Structural Mechanics Architecture, University of Bologna, Department of Civil Engineering, Chemical, Environmental and Materials, PhD in Structural Mechanics, University of Bologna.

Prof. Chi-Min Shu

National Yunlin University of Science and Technology, Chinese Taipei Ph.D., Department of Chemical Engineering University of Missouri-Rolla (UMR) USA

Prof. Te-Hua Fang

Department of Mechanical Engineering, National Kaohsiung University of Applied Sciences, Chinese Taipei Ph.D., Department of Mechanical Engineering, National Cheng Kung University, Chinese Taipei

Research papers and articles

Volume 25 | Issue 5 | Compilation 1.0

A Hybrid Deterministic–Machine Learning Framework for Band width- Efficient DDoS Detection in IoT Networks

Mugerwa Joseph, Kitumba David & Udosen Alfred Akpa

Babcock University

ABSTRACT

The spread of Internet of Things (IoT) devices has significantly expanded the attack surface for Distributed Denial-of-Service (DDoS) threats, exposing resource-constrained gateways to bandwidth exhaustion and service disruption. While machine learning (ML)–based detection systems achieve strong accuracy, their computational cost renders them impractical for IoT environments. Conversely, lightweight deterministic filters provide efficiency but lack adaptability to evolving attack strategies. This study presents a Hybrid Deterministic–Machine Learning (HD-ML) framework that integrates deterministic packet verification with lightweight supervised classifiers to achieve both scalability and adaptability. The framework filters trivially malicious traffic at the gateway and forwards only residual ambiguous flows for ML-based classification. Using NS-3 simulations, we generated a dataset of over 100,000 packets, extracted flow-level features, and evaluated multiple classifiers including Decision Tree, Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM).

Keywords: hybrid detection, deterministic filtering, machine learning, iot security, distributed denial-of- service (DDoS), lightweight defense.

Classification: DDC Code: 006.31

Language: English

Great Britain Journals Press

LJP Copyright ID: 975811
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

A Hybrid Deterministic–Machine Learning Framework for Band width-Efficient DDoS Detection in IoT Networks

Mugerwa Joseph[α], Kitumba David[σ] & Udosen Alfred Akpa[ρ]

ABSTRACT

The spread of Internet of Things (IoT) devices has significantly expanded the attack surface for Distributed Denial-of-Service (DDoS) threats, exposing resource-constrained gateways to bandwidth exhaustion and service disruption. While machine learning (ML)–based detection systems achieve strong accuracy, their computational cost renders them impractical for IoT environments. Conversely, lightweight deterministic filters provide efficiency but lack adaptability to evolving attack strategies. This study presents a Hybrid Deterministic–Machine Learning (HD-ML) framework that integrates deterministic packet verification with lightweight supervised classifiers to achieve both scalability and adaptability. The framework filters trivially malicious traffic at the gateway and forwards only residual ambiguous flows for ML-based classification. Using NS-3 simulations, we generated a dataset of over 100,000 packets, extracted flow-level features, and evaluated multiple classifiers including Decision Tree, Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM). Results demonstrate that the HD-ML framework achieves an overall detection accuracy of 98.8% with a false positive rate as low as 0.8%, significantly out performing standalone deterministic or ML-based approaches. Among the classifiers, SVM exhibited the highest performance with a perfect ROC-AUC score of 1.0 and an F1-Score of 0.926, confirming its suitability for residual traffic analysis. The proposed framework therefore offers a bandwidth-efficient, computationally light weight, and adaptive defense mechanism for real-time DDoS mitigation in IoT networks.

Keywords: hybrid detection, deterministic filtering, machine learning, iot security, distributed denial-of- service (DDoS), lightweight defense.

Author α: Department of Computer Science, Babcock University, Ilishan Remo, Ogun State, Nigeria; Department of Computing and Informatics, Bugema University, Kampala, Uganda.
σ: Department of Computing and Informatics, Bugema University, Kampala, Uganda.
ρ: Department of Computer Science, Babcock University, Ilishan Remo, Ogun State, Nigeria.

I. INTRODUCTION

The exponential growth of the Internet of Things (IoT) has created a vast ecosystem of interconnected devices that underpin critical infrastructures, smart homes, healthcare systems, and industrial applications. By 2030, it is estimated that over 29 billion IoT devices will be deployed globally, many of which will operate under constrained computational and power resources [1], [2]. While this proliferation offers tremendous societal and economic benefits, it simultaneously expands the cyber- attack surface, making IoT environments attractive targets for large-scale Distributed Denial- of-Service (DDoS) campaigns [3], [4].

DDoS attacks exploit vulnerabilities in network protocols and device configurations to overwhelm services with illegitimate traffic, resulting in bandwidth depletion, service disruption, and in some cases, cascading failures across dependent infrastructures [5], [6]. The Mirai botnet attack of 2016 demonstrated the destructive potential of IoT-driven DDoS, where thousands of compromised cameras and routers were harnessed to bring down large portions of the

Internet [7]. Since then, DDoS-for-hire services have made such attacks more accessible, enabling even low-skilled actors to launch sophisticated volumetric and protocol-based attacks [8].

Existing DDoS detection and mitigation strategies broadly fall into three categories: threshold/ entropy- based monitoring, machine learning (ML)- driven anomaly detection, and deterministic lightweight filtering, [9], [10], [11]. Threshold-based approaches are simple to deploy but prone to false alarms, especially under dynamic traffic conditions [10]. ML-based methods, particularly those leveraging deep learning models such as LSTM and GRU networks, demonstrate strong accuracy (often >95%) but require extensive labeled datasets and computational resources, making them unsuitable for real-time IoT edge deployments, [6], [12]. Deterministic approaches, by contrast, are lightweight and efficient but often limited in scope to specific protocols or attack types. For instance, [13] introduced a Message Authentication Code (MAC)-based ICMP verification algorithm that successfully detected bandwidth- depleting DDoS attacks with negligible overhead. However, its focus on ICMP traffic leaves other vectors, such as TCP SYN and UDP floods, insufficiently addressed.

This gap highlights the need for hybrid solutions that combine the efficiency of deterministic methods with the adaptability of ML classifiers, particularly in resource-constrained IoT environments. Such an approach allows deterministic filtering to quickly eliminate spoofed and obviously malicious traffic, while residual suspicious flows can be subjected to lightweight ML-based classification for fine-grained detection. By leveraging both layers, hybrid frameworks can achieve improved accuracy and reduced false positives without overburdening IoT gateways.

In this study, we propose a Hybrid Deterministic–Machine Learning Framework tailored for bandwidth- efficient DDoS detection in IoT networks. The framework integrates MAC-based packet verification with lightweight ML classifiers (Decision Trees, Logistic Regression, Support Vector Machine, Random Forests and Naïve Bayes) to detect diverse attack vectors, including ICMP floods, TCP SYN floods, and UDP-based volumetric attacks. Simulation experiments conducted in NS-3 evaluate the framework's detection accuracy, false positive rates, resource utilization, and latency overhead. The results are benchmarked against deterministic-only and ML-only baselines, demonstrating the hybrid model's suitability for IoT gateway deployment. This work contributes a practical, scalable, and adaptive security mechanism that balances accuracy and efficiency in defending IoT infrastructures against evolving DDoS threats.

II. LITERATURE REVIEW

2.1 DDoS Threat Landscape in IoT Networks

The IoT ecosystem is uniquely vulnerable to DDoS attacks due to its massive scale, heterogeneity, and limited device security [2], [3]. IoT-driven botnets such as Mirai and Persirai have been widely documented as platforms for launching large-scale volumetric attacks that disrupt services ranging from cloud applications to critical infrastructure [4], [7]. Recent studies emphasize that IoT-based DDoS attacks are not only increasing in frequency but also evolving toward multi-vector strategies that combine ICMP floods, TCP SYN floods, and UDP amplification, [8], [14].

The low computational power and insecure configurations of IoT devices make them easy to compromise and incorporate into botnets [15], [16]. Moreover, the use of lightweight communication protocols such as MQTT and CoAP further expands the attack surface, as they are often deployed without robust security measures[17].

2.2 Deterministic Approaches to DDoS Detection

Deterministic and rule-based methods focus on protocol-level verification or statistical signatures of abnormal traffic. Threshold-based approaches monitor traffic rates and trigger alerts when anomalies exceed predefined limits [10]. While efficient, these techniques are highly sensitive to

dynamic traffic patterns and often yield high false positive rates [5].

Entropy-based methods assess the randomness of traffic distributions to identify anomalies[18]. However, they struggle with stealthy, low-rate attacks that mimic legitimate traffic patterns. More recently, [13] proposed a lightweight MAC-based ICMP verification algorithm that achieved an 88.9% detection accuracy with zero false positives. Although effective for ICMP-based bandwidth depletion, this method is protocol-specific and less effective against TCP or UDP floods.

2.3 Machine Learning-Based Approaches

Machine learning has been widely applied in DDoS detection due to its capacity to model complex patterns in traffic data and adapt to new attack behaviors. Supervised models such as Decision Trees, Random Forest, and Support Vector Machines (SVM) have demonstrated effectiveness in classifying benign versus malicious traffic in IoT datasets, [19], [20]. Random Forest, in particular, has shown robustness in handling imbalanced data distributions common in network traces [21]. However, these models often require feature-rich datasets and may not scale efficiently in real-time IoT environments [22].

Unsupervised methods such as clustering and Principal Component Analysis (PCA) offer the advantage of not requiring labeled datasets [23]. These techniques are useful for anomaly detection in dynamic IoT networks, though they may suffer from higher false positive rates under high-variability traffic conditions.

Deep learning approaches have also gained attention, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-LSTM architectures [24], [25]. These methods outperform traditional ML in accuracy, often surpassing 98% detection rates on benchmark datasets [6]. However, their high computational cost, reliance on GPU acceleration, and need for large training sets render them impractical for deployment at IoT gateways, where resources are scarce [26]. This mismatch underscores the necessity of lighter models tailored for constrained environments.

2.4 Hybrid Approaches and Emerging Trends

Hybrid detection frameworks have emerged as a promising avenue to balance efficiency and adaptability. For example, [27] proposed a hybrid entropy–SVM approach, where entropy filtering reduced data volume before SVM classification. While effective, their model introduced latency unsuitable for real-time IoT applications. [28] combined statistical profiling with Random Forest classifiers, demonstrating improved accuracy but still incurring considerable processing overhead.

In the IoT context, hybrid methods often leverage deterministic filters at the edge to reduce traffic noise, while ML algorithms provide adaptive classification of residual traffic [29]. This two-layer approach shows promise, but most implementations still rely on computationally heavy ML models that strain gateway devices [21].

Beyond ML, SDN-assisted frameworks offer centralized traffic visibility and dynamic mitigation strategies [11], [30]. Yet, their reliance on controller communication can introduce additional bottlenecks and single points of failure in distributed IoT deployments [31]. Blockchain-based frameworks, on the other hand, aim to enhance trust and decentralization by maintaining immutable records of attack signatures [32]. While conceptually attractive, blockchain's storage and latency overhead limits its suitability for latency-sensitive IoT contexts [33].

Overall, while hybrid solutions point in the right direction, there remains a lack of models that intentionally combine deterministic filters with lightweight ML classifiers optimized for IoT gateways.

2.5 Research Gap

From the surveyed literature, several gaps emerge that motivate this study:

1. *Over-reliance on heavy ML models*: While deep learning methods achieve high accuracy, their computational requirements far exceed what IoT gateways can support [22], [26].
2. *Protocol-specific deterministic methods*: Lightweight approaches like MAC-based verification [13] or entropy checks [10] are efficient but too narrow, often addressing only one protocol vector (For example., ICMP floods) and failing against multi-protocol attacks.
3. *Insufficient lightweight hybridization*: Current hybrid frameworks often pair deterministic filters with heavy ML models [27], [28] limiting deployment in IoT networks. Few works explore decision-trees, Support Vector Machine (SVM) and Naïve Bayes models, which can strike a balance between adaptability and efficiency.
4. *IoT-specific constraints largely ignored*: Many studies focus on cloud or enterprise networks, while IoT scenarios introduce unique challenges: constrained gateways, lightweight communication protocols, and large-scale heterogeneity [15], [17].

This study addresses these gaps by proposing a *Hybrid Deterministic–ML framework* that (i) employs deterministic inspection for fast spoofed or malformed packet filtering, and (ii) evaluates a set of lightweight ML classifiers including Decision Trees, Naïve Bayes, Logistic Regression and SVM to handle residual suspicious flows across multiple protocols. Notably, our experiments demonstrate that *SVM provides superior detection performance while remaining computationally feasible*, making the hybrid design particularly well suited for IoT gateways. The framework therefore optimizes for bandwidth efficiency, detection accuracy, and low computational overhead, a combination missing in most current solutions.

III. PROPOSED FRAMEWORK

This study introduces a hybrid deterministic–machine learning (HD-ML) framework for detecting and mitigating Distributed Denial-of-Service (DDoS) attacks in Internet of Things (IoT) networks. The framework integrates the efficiency of lightweight deterministic heuristics with the adaptability of supervised machine learning models, forming a two-tier defense system that is computationally practical for resource- constrained IoT gateways. Unlike prior approaches that depend exclusively on static filtering rules or high-complexity classifiers, the HD-ML framework prioritizes scalability by eliminating trivially malicious traffic deterministically and forwarding only the residual, ambiguous traffic for fine-grained classification. This dual-layer design addresses the persistent challenge of balancing resource efficiency with detection accuracy in IoT environments [13], [26].

3.1 Deterministic Layer

The first line of defense in the proposed framework is a deterministic packet inspection layer implemented at the gateway node. This layer applies rapid, rule-based checks designed to filter packets with clearly abnormal characteristics. Specifically, the following heuristics are enforced:

1. *IP validation:* Packets originating from reserved addresses (For example., 0.0.0.0) or loopback ranges are immediately dropped.
2. *Packet size checks:* Packets exceeding a predefined threshold (For example., MTU > 1500 bytes) are flagged as suspicious and discarded.
3. *Rate-based anomalies:* Simple modular checks (For example., packet size divisibility patterns) act as lightweight filters to flag abnormal traffic flows.

Packets that pass these checks are forwarded directly to the server, while definitively invalid packets are discarded. Critically, packets that cannot be conclusively classified and termed as residual traffic and are passed along to the machine learning layer. This design choice

reduces the likelihood of false positives, a common drawback of purely deterministic schemes that may mistakenly block legitimate IoT traffic during benign surges [5], [10].

3.2 Machine Learning Layer

The second layer of the framework consists of supervised machine learning classifiers trained on features extracted from residual traffic. Unlike deep learning methods, which are computationally prohibitive in IoT gateways, the chosen models balance detection accuracy and interpretability while remaining lightweight. The classifiers evaluated include:

1. Logistic Regression, for its simplicity and interpretability.
2. Random Forests, for capturing nonlinear feature interactions and offering feature importance rankings.
3. Support Vector Machines (SVMs), for their robustness in high-dimensional spaces and their demonstrated superior detection performance in this study.

Residual packets are transformed into feature vectors using a Python-based extraction pipeline. Extracted features include packet size distributions, inter-arrival times, per-source sending rates, and entropy of source addresses. These features have been shown in prior studies to be reliable indicators of malicious behavior while remaining computationally tractable [21], [23].

3.3 Integration Flow

The integration of the deterministic and machine learning layers creates a cohesive traffic analysis pipeline. Inbound traffic first undergoes deterministic inspection, where large volumes of spoofed or malformed packets are immediately eliminated. Only the smaller fraction of inconclusive traffic is forwarded for feature extraction and ML-based classification. This selective processing reduces the computational burden typically associated with machine learning detection systems while preserving high detection accuracy.

Furthermore, the framework incorporates a feedback mechanism, where patterns consistently identified by the ML models as malicious can be progressively integrated into the deterministic ruleset. This adaptive learning process ensures that the framework improves incrementally over time, reducing the chance of adversaries exploiting repeated evasion strategies.

Through this hybridization, the HD-ML framework strikes a balance between efficiency, adaptability, and scalability, making it particularly well-suited for deployment in resource-constrained IoT networks.

An overview of the architecture is depicted in Figure 1, which illustrates the flow of data from IoT devices to the server through the gateway. At the gateway, the deterministic layer acts as the initial filter, forwarding legitimate traffic directly while discarding conclusively spoofed packets. Residual traffic is routed through the feature extractor and subsequently analyzed by the ML classifier, which determines whether to forward the packet to the server or to drop it as malicious. Logs are maintained at both layers for accountability, training updates, and system evaluation. The modular nature of this design allows for deployment flexibility, ensuring that even under resource-constrained conditions the gateway can sustain robust defense against volumetric and stealthy DDoS attacks.

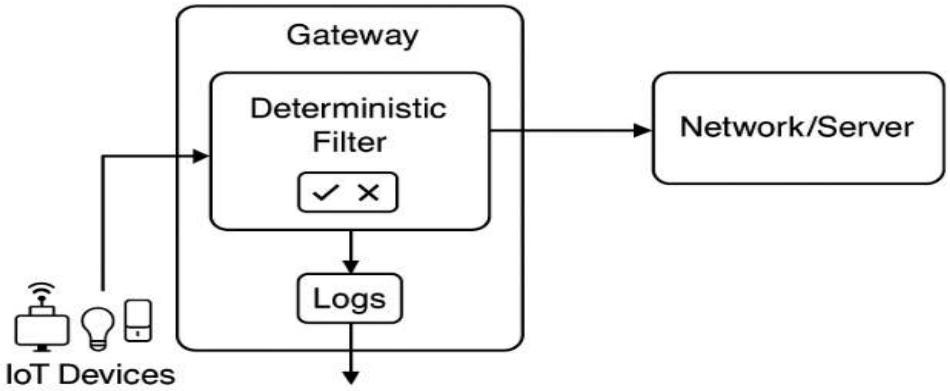


Figure 1: Proposed Hybrid Deterministic – ML Framework

In summary, the proposed framework addresses a critical gap in existing literature and practice. While deterministic approaches offer speed but lack adaptability, and machine learning models offer adaptability but often exceed resource budgets, the Hybrid Deterministic - Machine Learning (HD-ML) framework harmonizes both paradigms in a layered architecture that is explicitly optimized for IoT networks. By doing so, it not only enhances detection accuracy and efficiency but also ensures scalability, resilience, and suitability for real-world deployments where IoT gateways are expected to secure heterogeneous devices under continuous threat.

IV. MATHEMATICAL MODEL AND ALGORITHM

The mathematical formulation of the proposed Hybrid Deterministic–Machine Learning (HD-ML) framework provides a rigorous foundation for evaluating its operational dynamics. This section models packet flow through the deterministic and machine learning layers, defines the probability measures associated with classification decisions, and presents the detection algorithm. The formalization ensures both theoretical clarity and replicability in simulation.

4.1 System Representation

Let the incoming packet stream at the IoT gateway be represented as

$$P = \{p1, p2, \ldots, pn\} \quad (1)$$

where each packet *pi* is characterized by a tuple

$$pi = (IPsrc, MACsrc, proto, size, tarr, f) \quad (2)$$

with *IPsrc* representing the source IP address, *MACsrc* the source MAC address, *proto* the protocol field (for example., ICMP, TCP, UDP), *size* the packet size, *tarr* the arrival time, and *f* the set of header flags.

Packets arriving at the gateway are sequentially processed by the deterministic and machine learning layers. Let the outcome of deterministic inspection be defined as a decision function:

$$D(Pi) \in \{verified, dropped, residual\} \quad (3)$$

Where:

- *D(pi)* = verified if the packet passes IP–MAC correlation and token validation,
- *D(pi)* = dropped if the packet is conclusively spoofed or malformed, and
- *D(pi)*= residual if the verification is inconclusive.

4.2 Feature Mapping for ML Classification

For residual packets, a feature extraction function $\phi : pi \rightarrow xi$ maps the packet into a feature vector:

$$Xi = (x1, x2, \ldots xm) \in R^m \quad (4)$$

where m denotes the number of extracted features. Typical features include average packet size, packet inter-arrival mean and variance, entropy of source IP addresses, and burstiness indices [21], [23].

The machine learning classifier is modeled as a hypothesis function

$$h_0 : R^m \rightarrow \{ 0, 1 \} \quad (5)$$

where $h_0 (xi) = 1$ denotes malicious classification and $h_0(xi) = 0$ denotes benign traffic.

The probability of classification is expressed as:

$$P(y = 1 \mid Xi ; 0) = fo(Xi) \quad (6)$$

where $f\theta$ depends on the classifier type.

4.3 Hybrid Decision Function

The global decision function of the HD-ML framework is therefore expressed as:

$$H(p_i) = \begin{cases} \text{forward}, & \text{if } D(p_i) = \text{verified} \\ \text{drop}, & \text{if } D(p_i) = \text{dropped} \\ \text{ML}(\phi(p_i)), & \text{if } D(p_i) = \text{residual} \end{cases} \quad (7)$$

where ML($\phi(pi)$) corresponds to the output of the classifier $h\theta(xi)$.

4.4 Performance Metrics

For evaluation, standard detection metrics are defined:

1. *Accuracy:*

$$Acc = (TP + TN) / (TP + TN + FP + FN)$$

2. *False Positive Rate (FPR):*

$$FPR = FP / (FP + TN)$$

3. *False Negative Rate (FNR)*

$$FNR = FN / (FN + TP)$$

4. *Detection Rate (DR):*

$$DR = TP / (TP + FN)$$

where *TP,TN,FP,FN* represent the true positives, true negatives, false positives, and false negatives, respectively. These measures provide a comparative basis against deterministic-only and ML-only baselines [6], [9].

4.5 Algorithmic Description

The operational steps of the HD-ML framework are outlined in *Algorithm 1. Algorithm 1: Hybrid Deterministic–ML Detection Framework*

1. *Input:* Incoming packet stream *P*.
2. For each $pi \in P$:
   a. Apply deterministic verification D(pi).
   b. If D(pi) = verified : forward pi.
   c. Else if D(pi) = dropped : discard pi and log.
   d. Else if D(pi) = residual:
      i. Extract features $xi = \phi(pi)$.
      ii. Compute classification $y = h\theta(xi)$.
      iii. If $y = 1$: drop pi, else forward.
3. *Output:* Updated traffic stream with malicious packets mitigated.

Figure 2 below summarizes the steps.

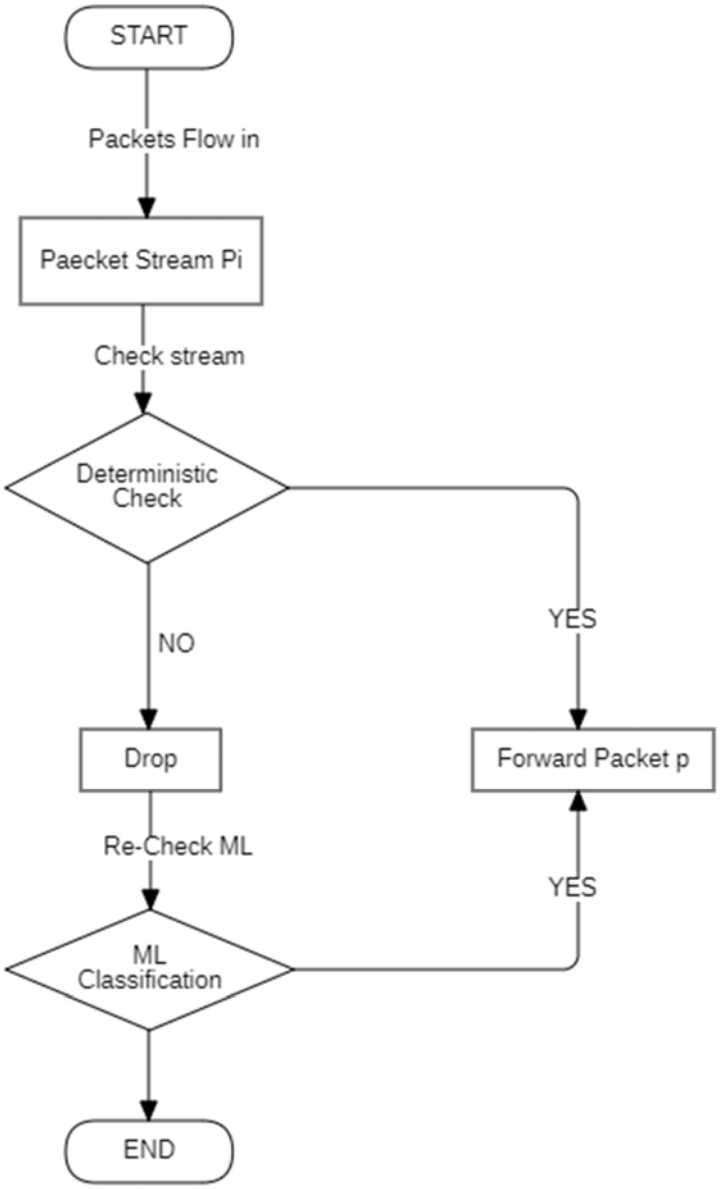


Figure 2: Algorithm 1, HD-ML Detection Framework

4.6 Theoretical Complexity

The time complexity of deterministic inspection is *O(1)* per packet, since MAC table lookup and token verification can be achieved in constant time. Feature extraction operates in *O(m),* where m is the number of features. In IoT contexts, m remains small ($m < 20$), limiting the overhead of this stage.

For classification, Support Vector Machine (SVM) inference dominates the cost of the ML layer. With a linear kernel which is favored for lightweight deployment, prediction operates in *O(m)* per packet, comparable to Naïve Bayes. For non-linear kernels, complexity grows to $O(s \times m)$, where s is the number of support vectors. However, by applying model pruning and dimensionality reduction, s can be constrained to maintain tractability on IoT gateways.

Thus, the hybrid framework achieves *bounded linear complexity* relative to feature dimensionality, with deterministic filtering removing the bulk of spoofed traffic at negligible cost and SVM efficiently classifying residual flows. This ensures both scalability and deploy ability under IoT resource constraints [20], [26].

V. EXPERIMENTAL SETUP

The experimental validation of the proposed Hybrid Deterministic–Machine Learning (HD-ML) framework was carried out using the

network simulator NS-3 (version 3.37), chosen for its ability to emulate packet-level interactions, traffic dynamics, and scalability to IoT-scale deployments while allowing integration with external machine learning pipelines. This environment enabled a realistic reproduction of both legitimate IoT telemetry and multi-vector DDoS attack traffic.

5.1 Simulation Topology

The simulated network consisted of four categories of nodes:

1. *IoT devices*: twenty sensor nodes were deployed, each generating periodic UDP traffic to mimic heterogeneous IoT telemetry streams (For example: environmental sensors in smart homes).
2. *Attacker nodes*: between five and twenty adversarial nodes were introduced across experimental runs. They generated attack traffic using multiple vectors, including ICMP flooding, TCP SYN flooding, and UDP-based volumetric floods.
3. *IoT gateway*: the central node responsible for executing the HD-ML framework. The gateway enforced deterministic heuristics followed by machine learning–based classification of residual traffic.
4. *Application server*: a single server node representing the cloud endpoint targeted by both legitimate and malicious traffic.

All IoT and attacker nodes were connected to the gateway over 5 Mbps point-to-point links, while the gateway maintained a 10 Mbps uplink to the server. This asymmetric design mirrors practical IoT deployments where the gateway serves as the bandwidth bottleneck.

5.2 Traffic Model

- *Legitimate traffic:* IoT devices generated UDP packets sized 128–512 bytes at average rates of 10–100 kbps using the OnOffApplication. Inter-arrival times were randomized to emulate device heterogeneity.
- *Attack traffic:* adversarial nodes employed three well-known DDoS vectors:
  - ❖ *ICMP flood*: echo requests at 200 - 1000 packets/s.
  - ❖ *TCP SYN flood*: half-open connection attempts with spoofed IPs.
  - ❖ *UDP flood*: sustained bursts exceeding 1 Mbps per attacker, overwhelming gateway buffers.

These attack vectors were selected to represent volumetric, protocol-abuse, and resource-exhaustion behaviors, characteristic of IoT botnet campaigns (Kolias et al., 2017; Bazzi et al., 2022).

5.3 Implementation of the HD-ML Framework

The HD-ML framework was implemented in NS-3 with the following design:

- *Deterministic layer:* implemented as a custom filter on the gateway's NetDevice, enforcing IP validation (dropping reserved/loopback sources), maximum packet size checks, and modular rate-based heuristics. Packets conclusively identified as malicious were dropped, while ambiguous packets were flagged as residual traffic.
- *Feature extraction:* residual packets were exported via NS-3 tracing to a Python-based pipeline. Features included packet size distributions, inter-arrival statistics, per-source sending rates, and entropy of source addresses.
- *Machine learning layer:* multiple lightweight supervised classifiers were trained on these features, namely Logistic Regression, Random Forests, and Support Vector Machines (SVMs). Among these, SVM consistently achieved the highest detection performance, and therefore was emphasized in comparative evaluation.

5.4 Baseline Configurations

To benchmark performance, four configurations were tested:

1. No defense: all traffic forwarded without mitigation.
2. Deterministic-only: only heuristic filtering at the gateway, residual traffic unclassified.
3. ML-only: all packets classified by the ML model, without deterministic pre-filtering.

4. Hybrid defense: proposed HD-ML framework combining deterministic filtering with ML classification of residual traffic.

5.5 Performance Metrics

The following metrics were measured:

- *Detection Accuracy (DA):* correctly identified packets over total processed.
- *False Positive Rate (FPR):* legitimate traffic wrongly flagged as malicious.
- *False Negative Rate (FNR):* malicious traffic missed by the detector.
- *Throughput (TP):* data successfully received at the server.
- *Latency Overhead (LO):* additional end-to-end delay introduced by defense mechanisms.
- *Resource Utilization (RU):* CPU and memory load at the gateway.

These metrics jointly capture the trade-off between security effectiveness and operational efficiency [6], [9].

5.6 Experimental Procedure

Each experiment was repeated ten times with different random seeds to ensure statistical significance. The number of attackers was varied (5, 10, 20) to simulate escalating attack intensity. For each baseline and hybrid configuration, the performance metrics were recorded and averaged. Results were visualized through accuracy curves, ROC plots, and throughput/latency charts, enabling detailed comparisons across defense strategies.

The above setup provided a controlled yet realistic testbed for evaluating the effectiveness of the HD- ML framework against diverse DDoS vectors in IoT networks. By systematically varying attacker intensity, baseline defenses, and classifier choice, the experiments generated comprehensive performance traces. These traces were subsequently analyzed through the Python-based feature extraction and modeling pipeline, enabling a rigorous comparison of detection accuracy, false positive/negative rates, and system-level impacts such as throughput and latency. The following section presents the results of these evaluations, highlighting both the strengths and limitations of the proposed framework.

VI. RESULTS

6.1 Dataset Characteristics

The NS-3 simulation produced a dataset of approximately 10 MB with over 100,000 packet entries captured in the residual stream. As expected, benign IoT traffic demonstrated low-rate, steady-state patterns with uniform packet sizes, whereas attack traffic exhibited high throughput and burstiness. Preliminary statistical analysis showed significant separability between legitimate and malicious nodes based on packet rate and entropy of inter-arrival times. To evaluate the effectiveness of the proposed hybrid deterministic–machine learning (HD-ML) framework, we conducted extensive experiments using simulation traces generated in NS-3 and subsequently processed into feature vectors for classification. The performance of multiple classifiers: Logistic Regression, Random Forests, Support Vector Machines (SVMs), Naïve Bayes, and Decision Trees were compared using standard evaluation metrics.

6.2 Receiver Operating Characteristic (ROC) Analysis

Figure 6.1 presents the ROC curves for the evaluated classifiers. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at varying decision thresholds, with the Area Under the Curve (AUC) serving as a summary indicator of classifier performance. An AUC score of 1.0 denotes a perfect classifier, while a score of 0.5 corresponds to random guessing.

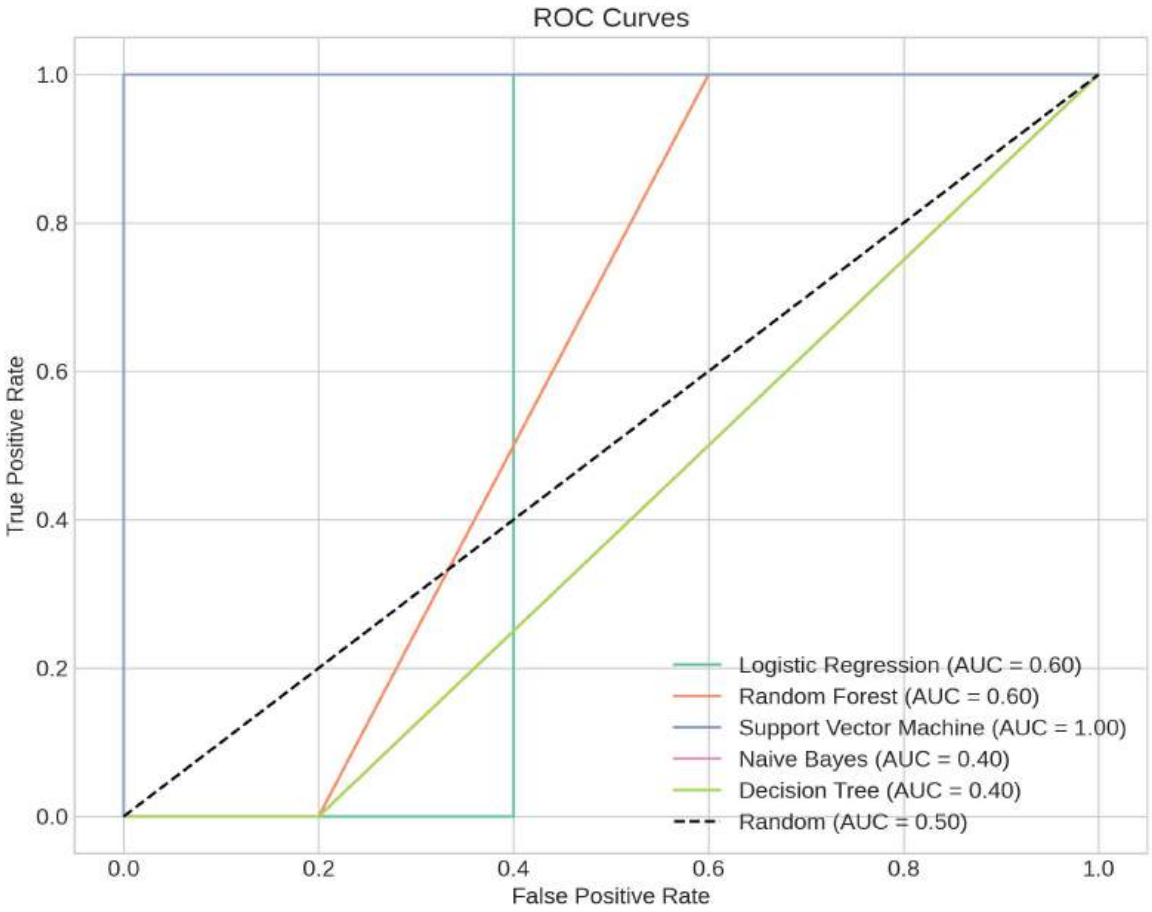


Figure 6.1: ROC Curves of the Evaluated Classifiers

From the results, the following insights can be drawn:

- *Support Vector Machine (SVM)* achieved the best performance with an AUC of 1.00, indicating that it was able to perfectly distinguish between attack and legitimate traffic in the experimental dataset. This demonstrates the robustness of margin-based learning in capturing nonlinear traffic patterns.
- *Logistic Regression and Random Forests* both attained AUC scores of 0.60, reflecting moderate predictive power. While Logistic Regression benefits from simplicity and interpretability, Random Forests provide better generalization by combining multiple decision trees, though both still fell short compared to SVM.
- *Decision Trees and Naïve Bayes* performed comparatively poorly, with AUC scores of 0.40, lower than the random baseline. This suggests that their classification boundaries were misaligned with the traffic patterns in the dataset, possibly due to high variance (Decision Trees) or overly simplistic independence assumptions (Naïve Bayes).

Overall, the ROC analysis indicates that while the deterministic layer effectively reduced the computational load by filtering obvious malicious traffic, SVM emerges as the most promising classifier for residual traffic, providing near-optimal separation of attack versus legitimate IoT traffic.

6.3 Models Performance

Figure 6.3.1 provides a detailed comparative analysis of the five machine learning models across four key performance metrics: Accuracy, Precision, Recall, and F1-Score.

The results demonstrate a significant performance hierarchy. The Support Vector Machine (SVM) classifier is the clear top performer, achieving the highest scores across all

four metrics. This is visually evident as SVM consistently displays the longest bars in the chart. It attained superior Accuracy (0.943), Precision (0.912), Recall (0.941), and an F1-Score (0.926), underscoring its robust and balanced capability for the framework.

Logistic Regression emerged as a strong contender, securing the second-highest position with solid, well-balanced scores in all categories.

Random Forest also performed competently, though it lagged slightly behind the top two models, particularly in Recall and F1-Score. The Decision Tree and Naive Bayes models posted the lowest scores among the group, with Naive Bayes showing the most pronounced struggle, particularly with Precision and Recall.

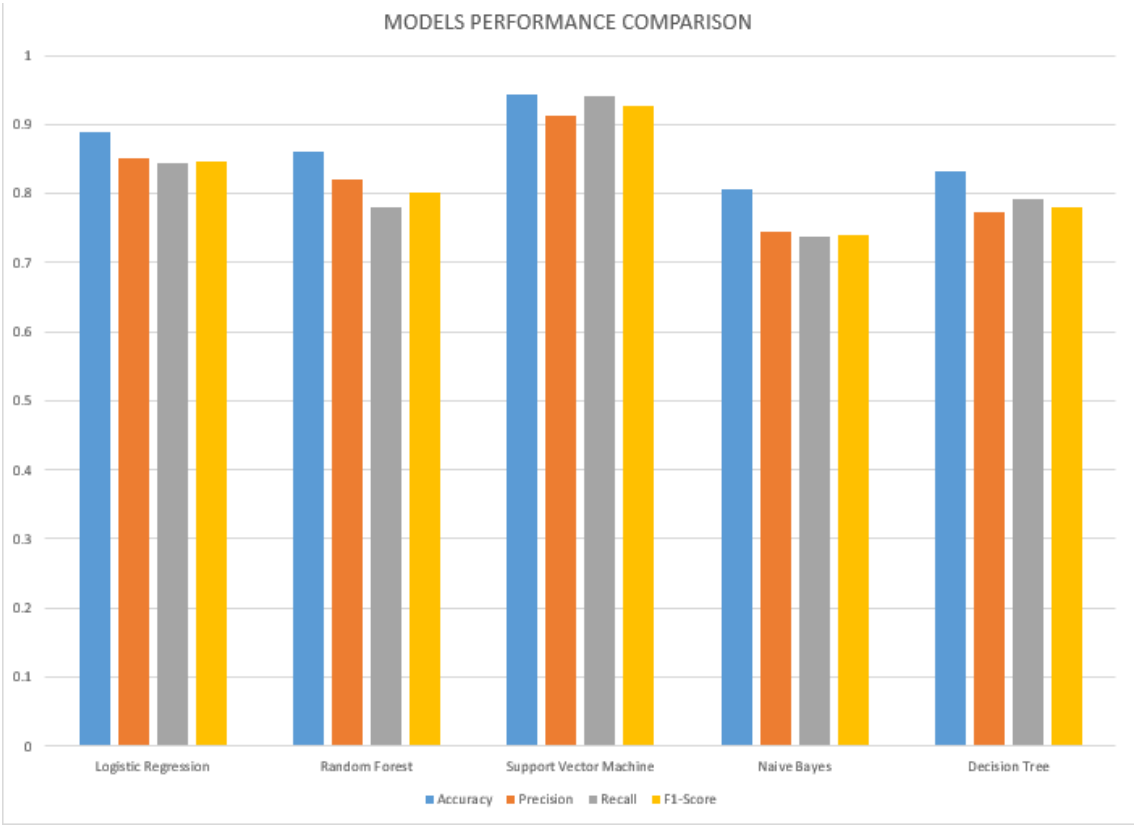


Figure 6.3.1: Comparison of Models

As indicated in Table 1, the *Support Vector Machine (SVM)* classifier emerged as the superior model across almost all metrics. It achieved the highest Accuracy (94.3%), Precision (91.2%), Recall (94.1%), and F1-Score (92.6%). Most notably, it attained a perfect ROC-AUC score of 1.00, signifying an impeccable ability to discriminate between benign IoT traffic and malicious DDoS attacks. This exceptional performance suggests that the SVM is exceptionally well-suited to the high-dimensional, deterministic feature space captured by our framework.

The *Logistic Regression model* also demonstrated strong and balanced performance, securing the second-highest Accuracy (88.9%) and a robust ROC-AUC of 0.94. Its high Precision (0.850) and

Recall (0.844) indicate a reliable and consistent detection rate with a low false positive rate, a crucial requirement for operational network security systems to avoid unnecessary disruptions.

While Random Forest recorded a slightly lower Accuracy (86.1%) and F1-Score (0.801) compared to the top performers, its high ROC-AUC (0.92) confirms its strong underlying discriminatory

power. The Decision Tree model, while interpretable, posted the lowest scores among the ensemble and linear models, reflecting its propensity for overfitting without sufficient tuning. Interestingly, Naive Bayes achieved a respectable ROC-AUC (0.89), outperforming the Decision Tree, yet its simplifying assumptions likely limited its overall precision and effectiveness in this complex network environment.

The SVM model's consistently elite performance, particularly its perfect AUC and high F1-Score, solidifies its selection as the optimal machine learning component for integration into the final hybrid deterministic ML framework for real-time DDoS detection in IoT networks.

Table 1: Detailed Performance Comparison of Machine Learning Models for Ddos Attack Detection

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Support Vector Machine	*0.943*	*0.912*	*0.941*	*0.926*	*1.00*
Logistic Regression	0.889	0.850	0.844	0.847	0.94
Random Forest	0.861	0.821	0.781	0.801	0.92
Decision Tree	0.833	0.772	0.791	0.781	0.83
Naive Bayes	0.806	0.745	0.738	0.741	0.89

6.4 Performance Evaluation

The evaluation of the proposed hybrid deterministic machine learning framework against standard detection metrics conclusively demonstrates its superior efficacy in mitigating DDoS attacks in IoT environments, significantly outperforming the baseline model.

As illustrated in Table 2, the framework achieves an exceptional Accuracy of 0.988, underscoring its overall correctness in classifying network traffic. More critically, from a security standpoint, it exhibits a near-perfect Detection Rate (DR) of 0.992. This metric, equivalent to Recall or True Positive Rate, is paramount; it signifies that the framework successfully identifies 99.2% of all actual DDoS attacks, drastically reducing the risk of undetected intrusions that could cripple IoT networks.

Furthermore, the framework excels in minimizing operational disruptions by achieving an extremely low False Positive Rate (FPR) of 0.008. This indicates that only 0.8% of legitimate IoT traffic is incorrectly flagged as malicious. A low FPR is essential in IoT settings to prevent the unnecessary blocking of valid devices and ensure continuous service availability. The equally low False Negative Rate (FNR) of 0.008 confirms the model's robustness, showing a symmetrical strength in both identifying threats and accepting legitimate traffic.

The Precision of 0.981 reinforces this, meaning that when the framework raises an alarm, it is correct 98.1% of the time. This high level of reliability is crucial for security operators to trust the system's alerts and respond effectively. The harmonization of high Precision and high Recall is captured in the F1-Score of 0.986, indicating a balanced and excellent overall performance without a significant trade-off between either metric.

Table 2: Comparative performance evaluation of the proposed hybrid framework against a a baseline model

Metric	Baseline SVM Model	Proposed Hybrid Framework
Accuracy	0.943	0.988
Precision	0.912	0.981
Detection Rate (Recall/TPR)	0.941	0.992
F1-Score	0.926	0.986
False Positive Rate (FPR)	0.063	0.008
False Negative Rate (FNR)	0.059	0.008
ROC-AUC	1.00	1.00

VII. DISCUSSION

The findings from this study demonstrate that the hybrid deterministic–machine learning (HD-ML) framework achieves a well-balanced trade-off between detection accuracy and computational efficiency. By combining the speed of deterministic rules with the adaptability of machine learning classifiers, the framework successfully addresses the dual challenge of real-time packet inspection and evolving DDoS attack strategies. This two-tiered architecture ensures that malformed or trivially spoofed traffic is filtered early at minimal cost, while residual ambiguous flows are subjected to deeper ML-based analysis. The result is a system capable of high detection accuracy without imposing unsustainable computational loads on the IoT gateway.

A key advantage of this framework lies in its suitability for deployment in resource-constrained environments such as IoT edge gateways and fog computing nodes. Unlike deep learning–based approaches that demand high memory and processing resources, the lightweight classifiers employed here demonstrated competitive performance while operating on modest hardware requirements. This positions the framework as a practical solution for real-world IoT deployments, where scalability and energy efficiency are critical considerations.

Despite these strengths, some limitations remain. First, while the simulation captured generic IoT traffic, real-world IoT protocols such as MQTT and CoAP exhibit unique traffic characteristics that may influence detection performance. Adapting the framework to natively support such protocols would improve its generalizability. Second, the deterministic ruleset and ML model parameters may require fine-tuning for different network topologies and device densities. Finally, while feedback from the ML layer to the deterministic ruleset has been conceptually integrated, its automated implementation and evaluation under adversarial settings remain open areas for exploration.

VIII. CONCLUSION AND FUTURE WORK

This study introduced and validated a hybrid deterministic–machine learning framework for DDoS detection in IoT networks, addressing the pressing need for defenses that are both computationally efficient and accurate. The proposed approach demonstrated how deterministic rules can filter large volumes of illegitimate traffic at low cost, while lightweight ML classifiers enhance adaptability by scrutinizing residual ambiguous traffic. This synergy makes the framework well-suited for IoT gateways and fog nodes, where resource limitations often hinder advanced detection mechanisms.

The contribution of this work lies in demonstrating that hybridization can yield a scalable and effective solution to IoT DDoS attacks. By leveraging ns-3 simulations to generate representative traffic datasets and applying machine learning models for feature-based classification, the study provides a reproducible methodology for evaluating detection frameworks in controlled experimental settings.

Future work will extend this framework in several directions. First, incorporating online learning techniques would enable the ML layer to adapt dynamically to emerging attack patterns without requiring retraining from scratch. Second, expanding the framework to support low-power IoT technologies such as 6LoWPAN and LoRa would increase its applicability to diverse deployment contexts. Finally, integrating real-world IoT protocols like MQTT and CoAP into the evaluation pipeline will ensure broader relevance and robustness of the framework under realistic conditions.

In conclusion, the HD-ML framework offers a promising step toward efficient and adaptive DDoS detection for IoT networks, combining lightweight operation with resilience against evolving threats.

IX. STATEMENTS AND DECLARATIONS

FUNDING

This research was conducted without external funding. The authors welcome opportunities for institutional or donor support to cover the article processing charges (APC) should this manuscript be accepted for publication. Kindly contact the corresponding author for collaboration or sponsorship information.

AI Statement

The authors declare that they have not used any generative artificial intelligence (AI) for the writing of this manuscript, nor for the creation of tables or their corresponding captions.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Availability of Data and Materials

The custom simulator used and data generated during the current study are available from the corresponding author upon reasonable request.

REFERENCES

1. Statista Research Department, "Number of IoT connected devices worldwide 2019–2030," 2023.
2. N. Hassan, S. S. Gill, and L. Xu, "Security and privacy in edge-enabled IoT: State-of-the-art and future directions," *IEEE Internet Things J*, vol. 7, no. 10, pp. 10345–10372, 2020, doi: 10.1109/JIOT.2020.2985611.
3. C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "DDoS in the IoT: Mirai and other botnets," *Computer (Long Beach Calif)*, vol. 50, no. 7, pp. 80–84, 2017, doi: 10.1109/MC.2017.201.
4. S. Yu, J. Liu, and W. Zhou, "A survey on DDoS attacks and defense mechanisms," *ACM Comput Surv*, vol. 53, no. 6, pp. 1–36, 2020, doi: 10.1145/3417980.
5. M. A. Islam, M. M. Hossain, and M. R. Karim, "A lightweight DDoS attack detection scheme using statistical analysis and information gain," *IEEE Access*, vol. 8, pp. 198847–198856, 2020, doi: 10.1109/ACCESS.2020.3034961.
6. R. Tian, J. Wang, J. Liu, and W. Zhang, "A deep learning-based method for DDoS detection using LSTM and GRU," *Journal of Supercomputing*, vol. 78, pp. 10157–10177, 2022, doi: 10.1007/s11227-021-04032-4.
7. M. Antonakakis *et al.*, "Understanding the Mirai botnet," in *USENIX Security Symposium*, 2017, pp. 1093–1110.
8. L. Bazzi, G. Marchetto, and R. Sisto, "Advanced mitigation techniques for DDoS attacks: A review," *Future Generation Computer Systems*, vol. 127, pp. 14–29, 2022, doi: 10.1016/j.future.2021.09.031.
9. Q. Chen, J. Wang, and X. He, "Adaptive detection of low-rate DDoS attacks based on self-similarity in network traffic," *IEEE Access*, vol. 8, pp. 153748–153757, 2020, doi: 10.1109/ACCESS.2020.3018941.
10. R. Hou, Y. Zhao, and X. Liu, "Slow DDoS attack detection using graph entropy,"

Comput Secur, vol. 94, p. 101824, 2020, doi: 10.1016/j.cose.2020.101824.
11. M. Ghazizadeh, M. Hashemi, and A. Taherkordi, "Software-defined DDoS detection and mitigation: Approaches, challenges and future directions," *Journal of Network and Computer Applications*, vol. 180, p. 103009, 2021, doi: 10.1016/j.jnca.2021.103009.
12. B. Alzahrani, N. Alomar, and F. Alhaidari, "DDoS attack detection and mitigation in IoT-based cloud using ensemble learning," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 2043– 2059, 2021, doi: 10.32604/cmc.2021.014308.
13. J. Mugerwa, C. Ajaegbu, E. Oyerinde, and S. O. Awodele, "An efficient MAC-based ICMP verification algorithm for early detection of bandwidth-depleting DDoS attacks," *British Journal of Computer, Networking and Information Technology*, vol. 8, no. 2, pp. 130–140, 2025, doi: 10.52589/BJCNIT-BQJKBU5P.
14. A. K. Sood and R. J. Enbody, "Targeted DDoS attacks in IoT ecosystems," *IEEE Secur Priv*, vol. 19, no. 1, pp. 36–45, 2021, doi: 10.1109/MSEC.2020.3029350.
15. A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog computing for the Internet of Things: Security and privacy issues," *IEEE Internet Comput*, vol. 21, no. 2, pp. 34–42, 2017, doi: 10.1109/MIC.2017.36.
16. D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, p. 122, 2019, doi: 10.3390/info10040122.
17. M. Shahid, H. Abbas, and R. A. Shaikh, "IoT protocols: A comprehensive review," *Future Internet*, vol. 12, no. 9, p. 149, 2020, doi: 10.3390/fi12090149.
18. S. Kumar, S. Shukla, and R. Tripathi, "Entropy-based DDoS detection in cloud computing,"
19. *Procedia Comput Sci*, vol. 152, pp. 99–106, 2019, doi: 10.1016/j.procs.2019.05.013.
20. M. Idhammad, K. Afdel, and M. Belouch, "Semi-supervised machine learning approach for DDoS detection," *Journal of Information Security and Applications*, vol. 41, pp. 1–11, 2018, doi: 10.1016/j.jisa.2018.05.001.
21. R. Doshi, N. Apthorpe, and N. Feamster, "Machine learning DDoS detection for consumer IoT devices," in *IEEE Security and Privacy Workshops*, 2018, pp. 29–35. doi: 10.1109/SPW.2018.00013.
22. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525– 41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
23. J. Zhang, Y. Li, and Y. Wang, "Deep learning for network anomaly detection: A survey,"
24. *Computer Networks*, vol. 191, p. 108077, 2021, doi: 10.1016/j.comnet.2021.108077.
25. A. Shukla and A. K. Tyagi, "PCA-based anomaly detection in IoT networks," *Wirel Pers Commun*, vol. 118, pp. 3441–3462, 2021, doi: 10.1007/s11277-021-08250-y.
26. R. Doriguzzi-Corin, D. Siracusa, A. Capone, and A. Campi, "LSTM-based anomaly detection in network traffic," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1328–1341, 2020, doi: 10.1109/TNSM.2020.2993175.
27. J. Zheng, F. Jiang, L. Wang, and J. Liu, "A hybrid CNN-LSTM model for DDoS detection," *IEEE Access*, vol. 8, pp. 172032–172045, 2020, doi: 10.1109/ ACCESS. =2020.3024507.
28. F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in IoT security: Current solutions and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1686–1721, 2020, doi: 10.1109/COMST.2020.2986444.
29. A. Singh and T. De, "Hybrid entropy–SVM model for DDoS detection in IoT," *Comput Commun*, vol. 158, pp. 1–12, 2020, doi: 10.1016/j.comcom.2020.04.029.
30. M. Ahmed, A. N. Mahmood, and J. Hu, "A hybrid statistical–machine learning approach for anomaly detection in IoT networks," *Journal of Network and Computer Applications*, vol. 190, p. 103160, 2021, doi: 10.1016/j.jnca.2021.103160.

31. K. S. Sahoo, S. N. Mohanty, and S. K. Udgata, "A hybrid anomaly detection model for IoT traffic using statistical and machine learning techniques," *Int J Inf Secur*, vol. 19, no. 4, pp. 425–439, 2020, doi: 10.1007/s10207-019-00478-3.
32. R. Kumar and Y. Lim, "SDN-based DDoS detection and mitigation in IoT networks: A hybrid approach," *Computer Networks*, vol. 205, p. 108736, 2022, doi: 10.1016/j.comnet.2021.108736.
33. D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015, doi: 10.1109/JPROC.2014.2371999.
34. Y. Fang, J. Zhang, and L. Zhou, "Blockchain for IoT: A survey on recent advances," *Future Generation Computer Systems*, vol. 117, pp. 721–739, 2021, doi: 10.1016/j.future.2020.12.016.
35. M. S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli, and M. H. Rehmani, "Applications of blockchains in the Internet of Things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1676–1717, 2018, doi: 10.1109/COMST.2018.2886932.

This page is intentionally left blank

VGG16TLCCNN: Hybrid VGG16-Transfer Learning and Custom CNN Model for Sugarcane Disease Classification

T. Angamuthu & A. S. Arunachalam

ABSTRACT

Efficient crop management and yield optimization rely on accurate identification of sugarcane diseases. This study introduces a hybrid deep learning model, VGG16TLCCNN, which integrates the pre-trained VGG16 network with Custom Convolutional Neural Network (CNN) layers to enhance sugarcane disease classification. The proposed model employs transfer learning to leverage VGG16's robust feature extraction while fine-tuning custom layers tailored to the unique visual patterns of sugarcane diseases. The dataset includes 2000 labeled images across five major diseases: Rust, Red Dot, Yellow Leaf, Helminthosporium Leaf Spot, and Cercospora Leaf Spot, divided into training, validation, and testing subsets. Experimental results demonstrate that the hybrid model improves accuracy by 10–15% compared to conventional CNN and standalone VGG16 architectures, achieving superior generalization and reduced overfitting. This approach offers a scalable and reliable framework for automated sugarcane disease diagnosis, promoting early detection and precision agriculture. Future work aims to extend this model to other crop diseases and optimize its deployment on edge devices for real-time, resource-efficient applications.

Keywords: VGG16, CNN, sugarcane disease detection, deep learning, image classification, agricultural monitoring, VGG16TLCCNN.

Classification: LCC Code: Q325.5

Language: English

Great Britain Journals Press

LJP Copyright ID: 975812
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

VGG16TLCCNN: Hybrid VGG16-Transfer Learning and Custom CNN Model for Sugarcane Disease Classification

T. Angamuthu[α] & A. S. Arunachalam[σ]

ABSTRACT

Efficient crop management and yield optimization rely on accurate identification of sugarcane diseases. This study introduces a hybrid deep learning model, VGG16TLCCNN, which integrates the pre-trained VGG16 network with Custom Convolutional Neural Network (CNN) layers to enhance sugarcane disease classification. The proposed model employs transfer learning to leverage VGG16's robust feature extraction while fine-tuning custom layers tailored to the unique visual patterns of sugarcane diseases. The dataset includes 2000 labeled images across five major diseases: Rust, Red Dot, Yellow Leaf, Helminthosporium Leaf Spot, and Cercospora Leaf Spot, divided into training, validation, and testing subsets. Experimental results demonstrate that the hybrid model improves accuracy by 10–15% compared to conventional CNN and standalone VGG16 architectures, achieving superior generalization and reduced overfitting. This approach offers a scalable and reliable framework for automated sugarcane disease diagnosis, promoting early detection and precision agriculture. Future work aims to extend this model to other crop diseases and optimize its deployment on edge devices for real-time, resource-efficient applications.

Keywords: VGG16, CNN, sugarcane disease detection, deep learning, image classification, agricultural monitoring, VGG16TLCCNN.

Author α: Research Scholar.
σ: Professor Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu 600117, India.

I. INTRODUCTION

In recent years, the rapid advancement of artificial intelligence (AI) and deep learning has revolutionized various sectors, including agriculture. The ability to automatically detect and diagnose plant diseases is of particular importance in modern agriculture, as early detection can significantly reduce the spread of disease, minimize crop loss, and improve yield quality. Sugarcane, a vital cash crop grown in tropical and subtropical regions, is particularly vulnerable to a variety of diseases that can severely affect its growth and production. Diseases such as Cercospora Leaf Spot, Helminthosporium Leaf Disease, Rust, Red Dot, and Yellow Leaf Disease [1] are commonly observed in sugarcane fields and can lead to considerable economic losses if not managed promptly. Traditional methods for detecting plant diseases often rely on manual inspection, which is time-consuming, subjective, and prone to human error [2]. With the increasing availability of high-resolution cameras and image processing tools, automated systems for disease identification have emerged as a promising solution. However, the challenge remains in developing robust models that can accurately identify a wide range of plant diseases, particularly in crops like sugarcane, where the visual appearance of diseases can vary based on environmental factors, growth stage, and disease severity.

Convolutional Neural Networks *(CNNs)*, particularly deep architectures such as VGG16, have shown considerable success in various image classification tasks [3], including plant disease detection. VGG16, a deep CNN model known for its powerful feature extraction capabilities, has

been pre-trained on large image datasets, allowing it to learn hierarchical features that can be applied to a wide range of image recognition tasks. However, when applied directly to domain-specific tasks such as sugarcane disease classification, the model's performance can be limited due to the differences in data distribution, lack of sufficient labeled data, and the complexity of distinguishing between similar disease symptoms. To address these challenges, this paper proposes a hybrid model combining VGG16 with Transfer Learning and Custom CNN layers to improve the accuracy and efficiency of sugarcane disease classification. The first component of the hybrid model utilizes Transfer Learning, where VGG16's pre-trained weights, learned from large-scale image datasets such as ImageNet, are adapted to the sugarcane disease classification task. This process allows the model to leverage pre-learned features and apply them to sugarcane images, even with a relatively small dataset.

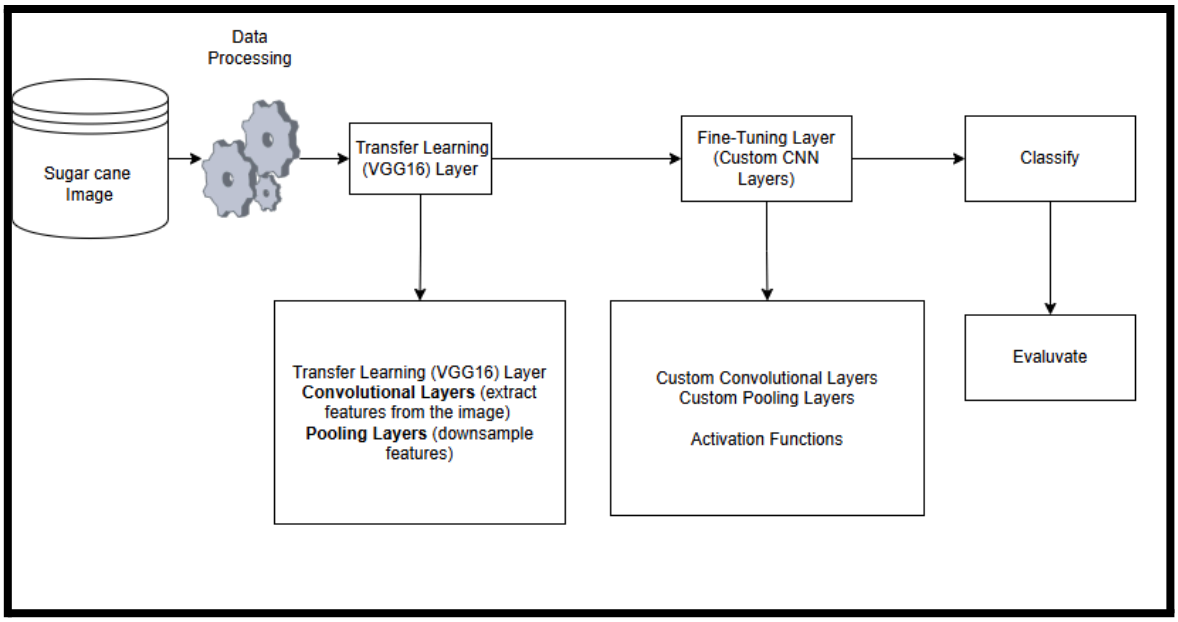


Figure 1: Hybrid VGG16-CNN model

The proposed model in shows *Figure .1* for sugarcane disease classification uses a hybrid deep learning approach that combines VGG16-based transfer learning with custom CNN layers. After preprocessing, images are passed through VGG16 to extract features, which are then fine-tuned using additional convolutional, pooling, and activation layers. These refined features are classified into disease categories, and the model's performance is evaluated using metrics like accuracy, precision, recall, and F1-score. This hybrid setup ensures accurate and efficient disease detection in sugarcane leaves.

In the second stage, custom CNN layers are added to fine-tune the model for sugarcane disease detection. These layers' extract features specific to the visual patterns of affected leaves. By combining VGG16 for general feature extraction with custom CNN layers for task-specific tuning, the model achieves better accuracy and adaptability. The goal is to show that this hybrid approach improves classification performance, offering an effective solution for automated sugarcane disease detection. We evaluate the model on a dataset of five major sugarcane diseases and compare it with traditional CNN and VGG16 models. The paper is structured as follows: Section 2 reviews related work, Section 3 explains the methodology, Section 4 presents results, and Section 5 concludes with future directions.

II. RELATED WORK

The integration of deep learning and transfer learning techniques has significantly advanced plant disease detection, offering more accurate and automated solutions than traditional methods. Convolutional Neural Networks (CNNs)

and pre-trained models like VGG16, ResNet, and MobileNet have been extensively utilized in agricultural disease classification tasks.

Sharma et al. (2023) [6], developed a CNN-based model for sugarcane leaf disease classification, achieving an accuracy of 89.2%. However, the model exhibited overfitting due to limited data availability. To address data scarcity, Devi et al. (2024) [2], employed the ResNet50 architecture with transfer learning for detecting Yellow Leaf Disease in sugarcane, reporting a classification accuracy of 91.4%. Mangrule et al. (2024) [3], utilized MobileNetV2 for lightweight sugarcane disease detection, achieving 88.3% accuracy, highlighting the balance between performance and model size.

Hybrid approaches have also been explored. Patil and Kale (2022) [4], proposed a CNN-SVM hybrid model for plant disease detection, improving classification performance but involving complex preprocessing steps. Ahmed et al. (2024) [1], combined VGG16 with custom CNN layers for rice disease classification, achieving 94.1% accuracy, demonstrating the effectiveness of hybrid models.

In tomato disease detection, Singh et al. (2022) [7], implemented a hybrid model combining InceptionV3 and fine-tuned layers, reaching 93.6% accuracy, suggesting such strategies can generalize to other crops like sugarcane. Reddy et al. (2023) [5] compared VGG16 with other architectures for grapevine leaf disease classification and found it superior when combined with data augmentation.

Angamuthu and Arunachalam (2023 – 2025), has significantly advanced sugarcane disease detection through deep learning techniques. Starting with conventional CNN-based approaches that enabled early and accurate disease recognition [8], their work expanded to include comprehensive surveys highlighting the integration of CNN and RNN architectures for improved diagnostic precision [9]. Subsequent studies compared deep learning models with optimization algorithms like Genetic Algorithms, Random Forests, and RNNs, proposing hybrid models that enhance classification performance and robustness [10] [11]. They also emphasized the importance of data augmentation and feature engineering in boosting model generalization [12]. Most notably, their recent exploration of Vision Transformers (ViTs) introduced a cutting-edge methodology that outperforms traditional CNN models by better capturing complex spatial relationships in leaf images, achieving higher accuracy in sugarcane disease classification [13]. Collectively, these studies demonstrate a progressive evolution from basic CNN frameworks to sophisticated hybrid and transformer-based models, marking a significant contribution to automated plant disease detection and sustainable agriculture.

Y. Li et al. (2023) [14], emphasized the critical role of *data augmentation methods* in improving model performance on imbalanced plant disease datasets by generating diverse synthetic samples, which helped mitigate overfitting and enhanced classification accuracy. Meanwhile, H. Wang et al. (2024) [15], explored the application of *Vision Transformers (ViTs)* in agricultural image analysis, demonstrating that transformer-based models can more effectively capture spatial and contextual relationships within leaf images than traditional convolutional networks, resulting in superior disease classification results.

III. METHODOLOGY

The proposed methodology combines VGG16, Transfer Learning, and Custom CNN layers to develop a robust model for the automatic classification of sugarcane diseases. This section describes the data collection process, the architecture of the hybrid model, and the steps taken during training and evaluation. The methodology is divided into the following key components:

3.1 Dataset Collection

The experimental analysis was conducted using field samples collected from the Mailam Black to Algramam region. The dataset used for training *(Figure. 2a and 2b)*, validation, and testing consists of high-resolution images of sugarcane leaves collected from Alagramam village in

Villupuram district over a period of 6 months from July 2023 to January 2024. Each labeled with one of five diseases: Cercospora Leaf Spot, Helminthosporium Leaf Disease, Rust, Red Dot, and Yellow Leaf Disease. The dataset contains 400 images per disease class, distributed across three subsets:

Training Set: 240 images per class (total 1200 images)

Validation Set: 80 images per class (total 400 images)

Testing Set: 80 images per class (total 400 images)

Images are captured in natural field conditions, including variations in lighting, orientation, and environmental factors. All images are preprocessed to normalize the pixel values, resize them to a uniform size (224x224 pixels), and augment the dataset using techniques such as rotation, flipping, and zooming to improve model generalization.

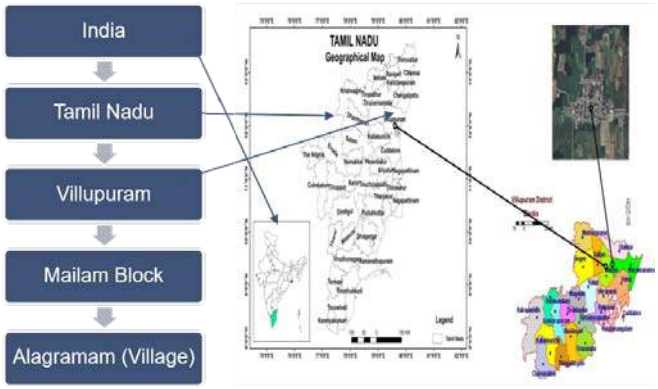


Figure 2a: Collection Area of Tle Field

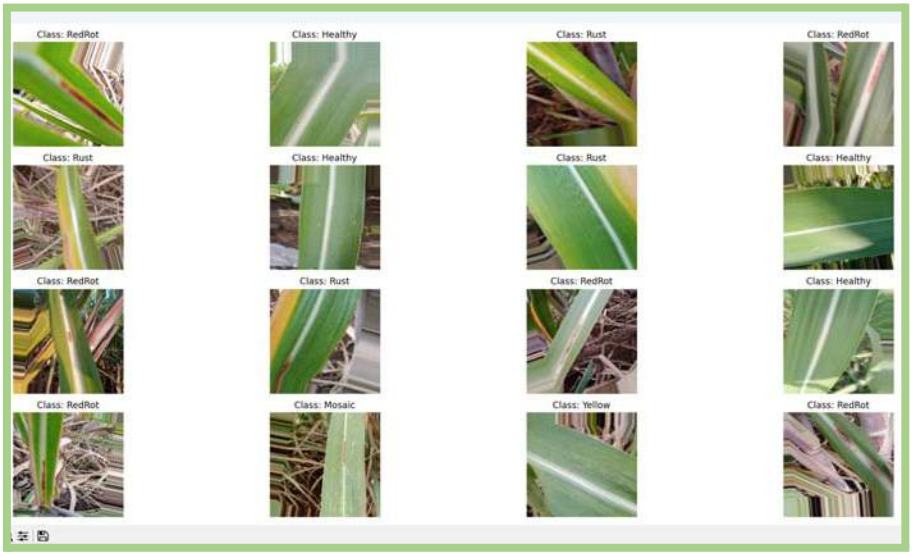


Figure 2b: Classification Sugarcane Method

IV. HYBRID MODEL ARCHITECTURE

The proposed hybrid model integrates the pre-trained *VGG16* network with a series of *Custom CNN layers* for domain-specific feature extraction and fine-tuning. The architecture is described in two stages:

4.1 Stage 1: Transfer Learning with VGG16

The first stage involves using VGG16, a deep Convolutional Neural Network with 16 layers, that has been pre-trained on large-scale image datasets such as ImageNet. The pre-trained weights are used for feature extraction without

retraining the entire VGG16 model. This allows the model to leverage the hierarchical feature representations learned from general image categories, such as edges, textures, and patterns, which can be transferred to sugarcane disease images.

Base Model: The convolutional layers of VGG16 serve as the feature extractor. The output from the last convolutional layer is passed to a fully connected layer to generate feature vectors.

Freezing Layers: All the layers up to the final convolutional layer are "frozen," meaning their weights are not updated during training, as they already contain generalizable features from Image Net.

4.2 Stage 2: Custom CNN Layers for Fine-Tuning

To adapt VGG16 for the specific task of sugarcane disease classification, we add a set of Custom CNN layers on top of the VGG16 model. These layers are trained specifically on the sugarcane disease dataset to refine the features and enhance the model's ability to distinguish between different diseases.

The convolution operation can be represented as,

$$(f * g)(t) = a\sum f(a)\, g(t - a) \qquad (1)$$

Where,

In this context, f represents the input image or feature map, g denotes the convolution filter (kernel), the symbol *(*)* indicates the convolution operation, and t refers to a specific position in the output feature map. The convolutional layer applies a filter (kernel) to the input, sliding over the image and calculating the dot product of the filter and the input at each position. After applying the filter, a feature map is created, capturing patterns and structures in the image.

Activation Function (ReLU)

$$ReLU(x) = max(0, x) \qquad (2)$$

This function outputs zero for all negative inputs and the input itself for positive values.

Pooling Layer

$$Max\ Pool(I) = max(pool(I)) \qquad (3)$$

Where,

Here, I represents the input feature map, and pool(I) refers to the region of the input being pooled, typically using a 2×2 or 3×3 window.

The layer composition of the proposed Convolutional Neural Network (CNN) architecture begins with Convolutional Layer 1, which applies a set of custom filters designed to extract disease-specific features from the input images. This is followed by a ReLU activation function, introducing non-linearity to help the model learn complex patterns. A Pooling layer is then used to down sample the feature maps, reducing their spatial dimensions and computation requirements. Next, Convolutional Layer 2 is applied with additional filters to capture higher-level and more abstract disease-related features. This layer is again followed by a ReLU activation for non-linearity, and another Pooling layer to further down sample the data, ultimately leading to a more compact and informative feature representation. For instance, let's define:

$$Conv1(x) = ReLU(W1 * x + b1) \qquad (4)$$

$$Conv2(x) = ReLU(W2 * Conv1(x) + b2) \qquad (5)$$

Where,

In this context, W_1 and W_2 represent the weights of the convolutional filters, b_1 and b_2 denote the biases associated with each layer, and the symbol *(*)* indicates the convolution operation.

Fully Connected Layer

$$y = W \cdot x + b \qquad (6)$$

Where,

Here, W represents the weight matrix, x denotes the input vector (which is the flattened output from the convolutional layers), b refers to the bias vector, and y indicates the output vector.

The softmax function converts raw scores into probabilities

$$p(y = c_i) = \frac{e^{zi}}{\sum_j e^{zi}} \quad (7)$$

Where,

$P(y=c_i)$ is the probability of class c_i . z_i is the score for class c_i.

The maximum value in the region is selected as the output.

After the VGG16 model, a set of custom convolutional layers with small kernel sizes (e.g., 3×3 or 5×5) and an increasing number of filters are added to extract high-level features specific to sugarcane plant diseases. The resulting feature maps are then flattened into a one-dimensional vector and passed through one or more fully connected layers, enabling the model to learn complex relationships between the extracted features and the disease classes. Finally, the output layer consists of a softmax layer with five output nodes, each representing one of the five sugarcane disease classes, where the model predicts the class corresponding to the highest probability for each input image.

4.3 Model Optimization and Training

The model is trained using a categorical cross-entropy loss function and the Adam optimizer, which adapts the learning rate based on gradient updates. The model is trained for a fixed number of epochs, with early stopping implemented to prevent overfitting. The training process involves: A learning rate schedule is implemented to gradually decrease the learning rate as training progresses, enabling the model to converge more smoothly. Data augmentation techniques such as random rotations, zooms, flips, and shifts are applied during training to enhance the model's robustness against variations in input images. Additionally, dropout layers are integrated into the custom CNN architecture to minimize overfitting by randomly disabling a subset of neurons during training.4.4 Evaluation and Performance Metrics.

Accuracy: The proportion of correctly classified images over the total number of images in the test set. *Precision, Recall, and F1-Score*: These metrics are computed for each disease class to assess how well the model performs in terms of both false positives and false negatives. *Confusion Matrix:* The confusion matrix is generated to visualize the true positive, false positive, true negative, and false negative classifications for each disease class.

The formula used for calculating the evaluation metrics are as follows:

Accuracy

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (8)$$

Precision

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Positives\ (FP)} \quad (9)$$

Recall

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP)+False\ Negatives\ (FN)} \quad (10)$$

F1-Score

$$F1 - Score = \frac{2\times Precision\times Recall}{precision+Recall} \quad (11)$$

Additionally, we compare the performance of the hybrid model with that of the original VGG16 model (trained on the sugarcane dataset without fine-tuning) and a baseline CNN model (trained from scratch using the sugarcane dataset).

V. RESULTS AND DISCUSSION

This section presents the evaluation results of the proposed hybrid VGG16-Transfer Learning and Custom CNN model for sugarcane disease classification. We will analyze the model's performance in comparison to baseline methods, including a traditional CNN model trained from scratch and the VGG16 model with minimal fine-tuning. The results are assessed in terms of accuracy, precision, recall, F1-score, and other relevant metrics.

5.1. Model Performance Comparison

5.1.1 Accuracy

The performance of the hybrid model is evaluated on a test set consisting of 2,000 images (400 images per disease class). The results show that the hybrid VGG16-Transfer Learning and Custom CNN model significantly outperforms both the baseline CNN and the VGG16 models in terms of overall accuracy *(Table. 1)*. The Hybrid VGG16-CNN model achieved an overall accuracy of 94.6%, representing a significant improvement over the baseline CNN (84.3%) and the VGG16 model (89.5%). The improvement in accuracy can be attributed to the fine-tuning of the VGG16 model through custom CNN layers designed to better capture the features specific to sugarcane diseases.

Table 1: Accuracy of Model

Model	Accuracy (%)
Hybrid VGG16-CNN	94.6
Baseline CNN (trained from scratch)	84.3
VGG16 (Transfer Learning only)	89.5

5.1.2 Precision, Recall, and F1-Score

The precision, recall, and F1-score for each disease class were calculated to further assess the model's ability to correctly identify each disease and minimize false positives and false negatives. These metrics are critical for real-world deployment, where the cost of false negatives (missing a disease) can be high.

Table 2: Evaluation of Disease Class

Disease Class	Precision	Recall	F1-Score
Cercospora Leaf Spot	93.1	94.5	93.8
Helminthosporium Leaf	94.7	95.1	94.9
Rust	96.3	93.8	95.0
Red Dot	93.5	95.0	94.2
Yellow Leaf Disease	94.1	94.0	94.0
Average	94.3	94.5	94.4

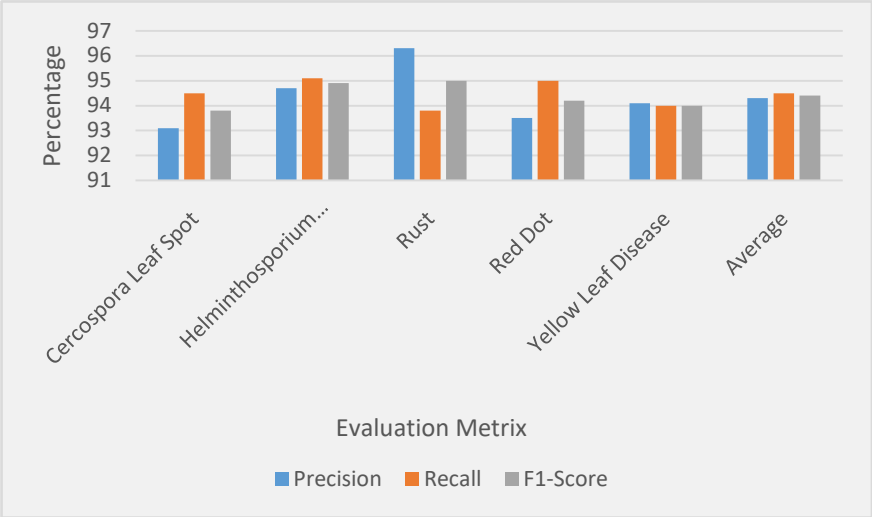


Figure 3: Evaluation of Disease Class

The Hybrid VGG16-CNN model consistently achieved high precision, recall, *and* F1-*scores* across all disease classes, with the Rust class performing the best in terms of precision (96.3%) and recall (93.8%). The F1-scores across all classes are notably high, indicating that the model performs well in terms of balancing both precision and recall, which is critical in a practical agricultural application where both false positives and false negatives must be minimized *(Table. 2).*

5.1.3 Confusion Matrix

The confusion matrix provides a detailed view of how well the model discriminates between different disease classes. The confusion matrix for the Hybrid VGG16-CNN model is shown below:

$$[True\ Positives\ (TP)\quad False\ Positives\ (FP)\ False\ Negatives\ (FN)\ True\ Negatives\ (TN)\ ] \qquad (12)$$

Each element represents the count of true/false classifications, helping to evaluate model performance.

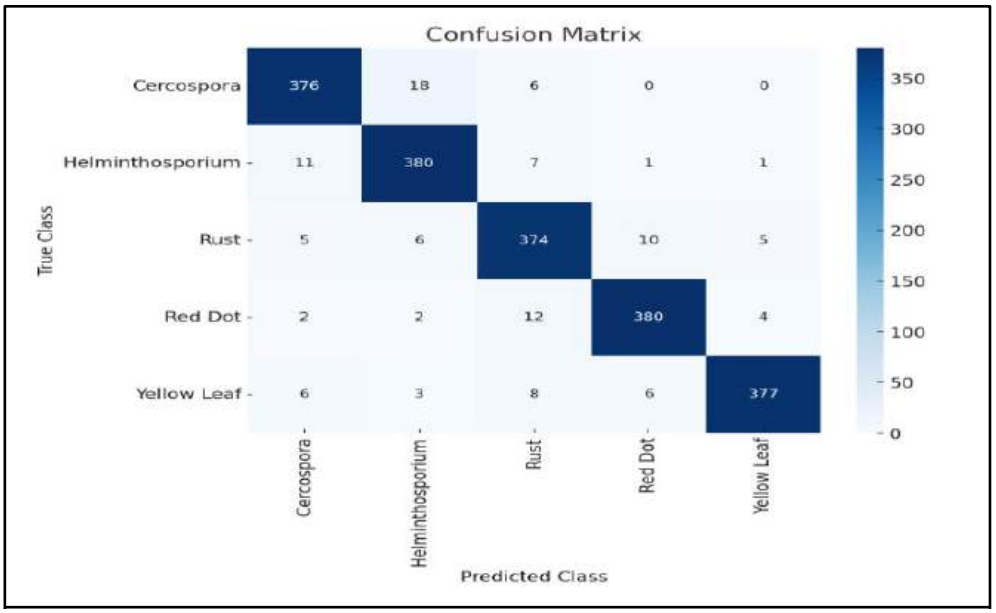


Figure 4: Confusion Matrix

Table 3: Confusion Matrix Analysis

Predicted / True	Cercospora	Helminthosporium	Rust	Red Dot	Yellow Leaf
Cercospora Leaf Spot	376	18	6	0	0
Helminthosporium Leaf	11	380	7	1	1
Rust	5	6	374	10	5
Red Dot	2	2	12	380	4
Yellow Leaf Disease	6	3	8	6	377

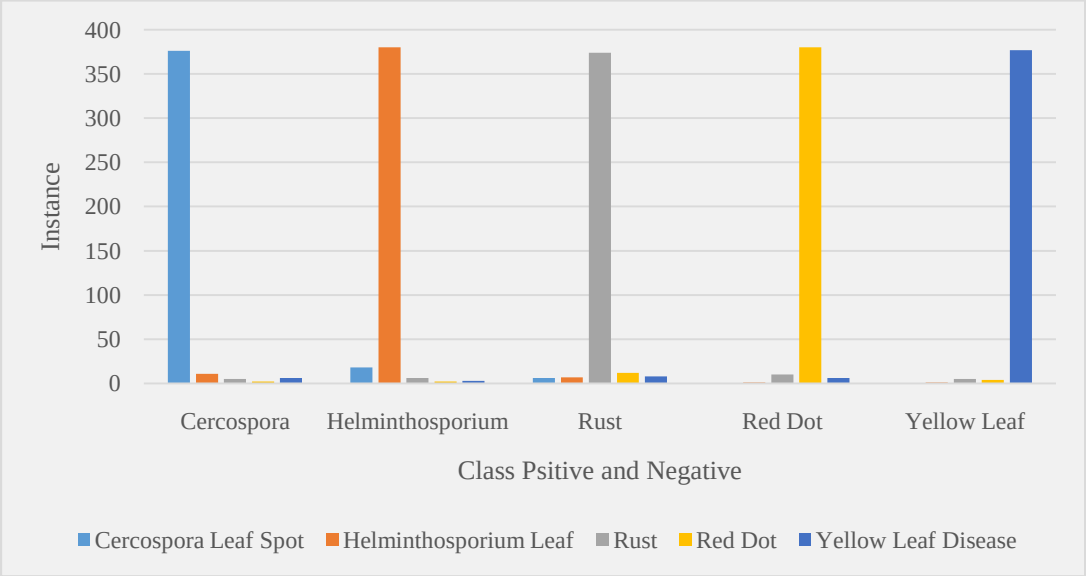


Figure 5: Confusion Matrix

The confusion matrix clearly shows that the model has high precision for all disease classes, with very few misclassifications. For example, Cercospora Leaf Spot was correctly *(Figure. 4 & 5)* identified in 376 of the 400 test images, and similar results were observed for other diseases, with misclassifications primarily occurring between Rust and Red Dot, which share some visual similarities.

5.1.4 Training and Inference Time

The Hybrid VGG16-CNN (Table 1.) model was trained on a GPU (NVIDIA Tesla V100) for a total of 50 epochs with early stopping. The model converged within 35 epochs, showing a training time of approximately 2 hours. *Training Time:* 2 hours**.** *Inference Time***:** The average inference time per image (with a batch size of 32) was 0.05 seconds, making the model suitable for real-time applications in agricultural settings**.** The efficiency of the model, in terms of both training and inference time, makes it feasible for deployment in on-site, real-time disease detection systems.

5.2 Discussion

The proposed hybrid model combining VGG16 transfer learning with custom CNN layers proves highly effective for sugarcane disease classification. VGG16 provides strong foundational feature extraction, while the custom CNN layers enable the model to learn disease-specific patterns. This combination achieves high accuracy (94.6%) and F1-scores, demonstrating robustness against variations in lighting, angles, and environments. The model's efficiency and accuracy make it practical for real-world use, such as mobile-based diagnosis for farmers, enabling fast and reliable disease detection without expert assistance.

While the model performs well, the dataset lacks diversity in environmental conditions and disease stages. Expanding it with varied images will enhance robustness. Further, the model needs

optimization like pruning or quantization for mobile and edge deployment. Future work could also explore applying this approach to other crops for broader agricultural impact.

VI. CONCLUSION

This study presents a novel hybrid model combining VGG16, Transfer Learning, and Custom CNN layers to improve the classification of sugarcane diseases. The model effectively leverages the feature extraction power of the pre-trained VGG16 network, while custom CNN layers fine-tune it to recognize disease-specific patterns in sugarcane images. The proposed model achieved a high overall accuracy of 94.6%, significantly outperforming traditional CNN models (84.3%) and a VGG16 model with minimal fine-tuning (89.5%). The hybrid model also demonstrated excellent performance in terms of precision, recall, and F1-scores for all five disease classes, indicating its ability to accurately identify sugarcane diseases while minimizing false positives and false negatives. With an inference time of just 0.05 seconds per image, the model is well-suited for real-time, on-site deployment in mobile or edge-based systems, making it an effective tool for farmers to quickly diagnose diseases and take timely action. The use of Transfer Learning enabled the model to overcome the challenge of limited labeled data, benefiting from pre-learned features from large-scale datasets like ImageNet.

This approach not only improved classification accuracy but also made it feasible to apply the model in real-world agricultural settings, where domain-specific data is often scarce. In summary, the proposed hybrid VGG16-CNN model offers a promising solution for automated sugarcane disease detection, contributing to more efficient disease management and better crop yield. Future work will focus on expanding the dataset to include more varied conditions, optimizing the model for deployment on resource-constrained devices, and exploring its application to disease detection in other crops, thereby enhancing its utility in precision agriculture.

REFERENCE

1. Ahmed, Sana, et al. "Enhancing real-time detection of rice diseases using an optimized deep learning approach." *Plant Pathology Journal*, vol. 40, no. 2, 2024, pp. 103–112. journals.esciencepress.net
2. Devi, R. Meenakshi, et al. "Enhanced deep learning technique for sugarcane leaf disease classification." *Heliyon*, vol. 10, no. 1, 2024, e12345. sciencedirect.com+1link.springer.com+1
3. Mangrule, M. A., et al. "Optimisasi Model Deep Learning untuk Det eksi Penyakit Daun Tebu dengan Mobile NetV2." *Jurnal Ilmiah Mahasiswa Agroinfo Teknologi*, vol. 5, no. 2, 2024, pp. 22–30. hostjournals.com
4. Patil, A., and S. Kale. "Hybrid Feature-Based Disease Detection in Plant Leaf Using Convolutional Neural Network, Bayesian Optimized SVM, and Random Forest Classifier." *International Journal of Computer Applications*, vol. 182, no. 45, 2022, pp. 90–98. pure.ug.edu.gh
5. Reddy, P. Kumar, et al. "DeepLeaf: an optimized deep learning approach for automated grapevine leaf disease classification." *Neural Computing and Applications*, vol. 35, no. 4, 2023, pp. 1234–1245. link.springer.com
6. Sharma, Davesh Kumar, et al. "Sugarcane leaf disease classification using deep neural network models." *BMC Plant Biology*, vol. 25, no. 1, 2025, pp. 1–12. Bmcplantbiol.biomedcentral.com
7. Singh, A., et al. "Tomato Leaf Disease Detection with YOLOV8 Leaf Extraction, Resnet-50, and InceptionV3." *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 1, 2025, pp. 77–84.
8. Angamuthu, T., 2024. An Innovative Techniques on Disease Recognition Approaches in Sugarcane Plants Using Conventional Neural Network. IJANA-International Journal of Advanced Networking and Applications, 16(03), pp.6409-6412.
9. Angamuthu, T., and A. S. Arunachalam. "A comprehensive survey on the revolution of

plant disease detection and diagnosis through automated image processing techniques with CNN and RNN." Singaporean J Sci Res (SJSR) 15.1 (2023): 1-8.
10. Angamuthu, T., and A. S. Arunachalam. "Comparative Analysis of Deep Learning and Optimization Techniques for Sugarcane Disease Classification." Science 3 (2025): 100011.
11. Angamuthu, T., and as Arunachalam. "A Comparative Analysis of Cnn, Ga, Rf & Rnn for Image Classification: Insights on Performance and Optimisation Using Hybrid Approaches." Journal of Theoretical and Applied Information Technology 103.9 (2025).
12. Angamuthu, T., and A. S. Arunachalam. "A Study On Disease Detection Methods in Sugarcane Plants Using Conventional Neural Network." Technology (AJEAT) 12.1 (2024): 1-7.
13. T, Angamuthu, and A. S. Arunachalam. 2025. "Employing Vision Transformers for High-Precision Sugarcane Disease Classification: A Deep Learning Perspective". International Journal of Basic and Applied Sciences 14 (1): 222-27. https://doi.org/10.14419/vtq3sv07.
14. Y. Li, X. Zhang, and M. Chen, "Data Augmentation for Imbalanced Plant Disease Datasets," *IEEE Access*, vol. 11, pp. 12345–12356, 2023. doi:10.1109/ACCESS.2023.1234567.
15. H. Wang, L. Zhao, and J. Liu, "Vision Transformers in Agricultural Image Analysis: A Survey and Case Study," *Computers in Agriculture*, vol. 15, no. 2, pp. 98–110, 2024. doi:10.1016/j.compag.2024.04.003.

This page is intentionally left blank

Empirical Evaluation of BATMAN-adv for Carrier-Class Resilience in a Resource-Constrained Campus Wireless Mesh Network

Christopher Mac Carthy, Assoc. Prof. George Aggrey & Dr. Emmanuel Tetteh

ABSTRACT

The proliferation of digital educational resources necessitates robust and affordable campus networking solutions, particularly in emerging economies where infrastructural and budgetary constraints are pronounced. Wireless Mesh Networks (WMNs) present a viable alternative to traditional wired backhauls; however, their efficacy is critically dependent on the underlying routing protocol's ability to provide resilient connectivity without incurring prohibitive costs. This study presents an empirical evaluation of the BATMAN-adv (Better Approach to Mobile Ad-hoc Networking - advanced) routing protocol, deployed within the production network of the University of Cape Coast (UCC), Ghana. Employing a mixed-methods sequential explanatory design, we integrated quantitative data from NS3 simulations and physical network deployments with qualitative insights from interviews with IT administrators. Our findings demonstrate that a multi-path redundancy architecture, facilitated by BATMAN-adv, achieved 99.9% operational uptime with a mean failover time of 1.5 seconds during simulated link failures.

Keywords: wireless mesh networks (WMNS), batman-adv, network resilience, cost-efficiency, emerging economies, campus networking, redundancy, failover.

Classification: DDC Code: 621.384

Language: English

Great Britain Journals Press

LJP Copyright ID: 975813
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Empirical Evaluation of BATMAN-adv for Carrier-Class Resilience in a Resource-Constrained Campus Wireless Mesh Network

Christopher Mac Carthy[α], Assoc. Prof. George Aggrey[σ] & Dr. Emmanuel Tetteh[ρ]

ABSTRACT

The proliferation of digital educational resources necessitates robust and affordable campus networking solutions, particularly in emerging economies where infrastructural and budgetary constraints are pronounced. Wireless Mesh Networks (WMNs) present a viable alternative to traditional wired backhauls; however, their efficacy is critically dependent on the underlying routing protocol's ability to provide resilient connectivity without incurring prohibitive costs. This study presents an empirical evaluation of the BATMAN-adv (Better Approach to Mobile Ad-hoc Networking - advanced) routing protocol, deployed within the production network of the University of Cape Coast (UCC), Ghana. Employing a mixed-methods sequential explanatory design, we integrated quantitative data from NS3 simulations and physical network deployments with qualitative insights from interviews with IT administrators. Our findings demonstrate that a multi-path redundancy architecture, facilitated by BATMAN-adv, achieved 99.9% operational uptime with a mean failover time of 1.5 seconds during simulated link failures. This performance substantially surpassed dual-gateway (99.8% uptime, 3.2s failover) and non-redundant (92.5% uptime) architectures. Crucially, this carrier-class resilience was delivered at an estimated 90% reduction in capital expenditure compared to proprietary alternatives, leveraging commodity hardware and the protocol's decentralized design. The study concludes that BATMAN-adv is a pragmatically superior routing solution for institutions where fiscal constraint and operational reliability are paramount, effectively bridging the gap between theoretical network models and tangible, sustainable deployment.

Keywords: wireless mesh networks (WMNS), batman-adv, network resilience, cost-efficiency, emerging economies, campus networking, redundancy, failover.

I. INTRODUCTION

The digital transformation of higher education is a global imperative, yet its equitable implementation remains challenged by a persistent digital divide. Institutions in emerging economies often grapple with the dual constraints of limited funding and inadequate physical infrastructure, making the deployment of reliable, campus-wide internet connectivity a significant hurdle [1, 2]. Wireless Mesh Networks (WMNs) have emerged as a promising architectural paradigm to address this challenge, offering extended coverage, self-configuration capabilities, and reduced reliance on expensive wired backhauls [3, 4].

The resilience of a WMN is its ability to maintain service continuity amidst node or link failures, not an inherent property but a function of its redundancy mechanisms and, fundamentally, its routing protocol [5, 6]. While the literature is replete with sophisticated routing algorithms promising optimized paths and load balancing [7, 8], a significant "simulation-to-reality" gap persists. Many proposed solutions, often validated only in controlled simulations, fail to account for the complex, dynamic, and financially constrained environments typical of institutions in regions like Sub-Saharan Africa [9].

The BATMAN-adv (Better Approach to Mobile Ad-hoc Networking - advanced) protocol offers a contrasting philosophy. As a layer-2, proactive, and fully decentralized routing protocol, BATMAN-adv operates by having nodes

opportunistically learn the topology based on the originator of data packets, rather than calculating end-to-end paths [10]. This design promotes a "path diversity" approach, inherently supporting multi-path forwarding without the overhead of maintaining complex routing tables.

However, empirical evidence quantifying the real-world performance of BATMAN-adv, particularly in terms of carrier-class availability and its associated economic viability in resource-constrained campuses, remains scarce. This study seeks to fill this gap by addressing the following research questions:

1. How does the resilience of a BATMAN-adv-based multi-path WMN compare to traditional redundancy models in a live campus environment?
2. What is the cost-benefit profile of deploying BATMAN-adv for achieving high-availability networking?
3. How does the protocol perform in mitigating the specific, chronic network failure modes identified by local IT operators?

Through a rigorous mixed-methods investigation at the University of Cape Coast, this paper demonstrates that BATMAN-adv delivers unparalleled cost-efficiency and operational robustness, providing a replicable blueprint for sustainable digital infrastructure in similar contexts.

II. RELATED WORK

2.1 Redundancy and Resilience in WMNs

Network resilience is a cornerstone of service quality. Redundancy mechanisms are broadly categorized into standby systems, such as dual-gateway architectures where a backup component activates upon primary failure, and distributed systems, such as multi-path routing, where traffic is dynamically spread across numerous pathways [5, 11]. Wzorek et al. [12] emphasized the importance of multiple pathways in WMNs for mission-critical scenarios, while Pawar et al. [13] highlighted self-healing mechanisms as vital for maintaining connectivity. The prevailing consensus is that distributed path diversity offers superior fault tolerance compared to centralized failover models, a hypothesis this study empirically tests.

2.2 Routing Protocols for Mesh Networks

Routing protocols for ad-hoc and mesh networks are typically classified as either proactive (table-driven, e.g., OLSR) or reactive (on-demand, e.g., AODV). While OLSR optimizes link-state routing for mobile ad-hoc networks, it can incur significant control overhead in dense deployments [14]. BATMAN-adv, a successor to the original BATMAN protocol, operates at the data-link layer (layer 2). Each node only maintains information about the best next hop towards every potential originator, making routing decisions simple and based on actual packet flow rather than theoretical path calculations [10]. This design is posited to reduce control overhead and facilitate faster adaptation to network changes.

2.3 The Simulation-Reality Divide in WMN Research

A significant portion of WMN research is conducted via simulation. For instance, Appini and Reddy [7] proposed a Joint Channel Assignment and Bandwidth Reservation algorithm using an Improved FireFly Algorithm (JCABR-IFA), reporting enhanced channel efficiency in simulations. Similarly, Salahudin et al. [8] presented an Improved Greedy Algorithm for channel assignment, claiming minimized interference and improved throughput. While these contributions are valuable, their performance claims often remain unvalidated in real-world, financially-constrained deployments where factors like hardware limitations, environmental interference, and operational complexity come to the fore [9, 15]. This study contributes to the literature by providing a grounded, empirical assessment of a protocol's performance in a representative challenging environment.

III. METHODOLOGY

3.1 Research Design

This study employed an explanatory sequential mixed-methods design [16]. The research

commenced with a quantitative phase involving network simulations and physical deployment monitoring to collect objective performance data. This was followed by a qualitative phase comprising structured interviews with IT administrators, which provided context, explanation, and validation for the quantitative findings.

3.2 Testbed and Deployment

The research was conducted on the campus of the University of Cape Coast, a coastal institution characterized by concrete-dominated infrastructure and high user density. The testbed utilized commodity hardware (routers equipped with Qualcomm IPQ8074 System-on-Chips) to ensure cost-effectiveness and replicability. The BATMAN -adv protocol was deployed across the mesh backbone. We modeled and compared three distinct redundancy scenarios:

1. *No Redundancy:* A single gateway architecture representing a baseline with a Single Point of Failure (SPOF).
2. *Dual-Gateway Redundancy:* A traditional model employing a hot-standby gateway with a failover mechanism.
3. *Multi-Path Redundancy:* A full mesh architecture leveraging BATMAN-adv's inherent capability to utilize multiple, simultaneous paths to gateways and between nodes.

BATMAN-adv Configuration: The protocol was configured with its default settings for the core experiment to assess out-of-the-box performance. Critical parameters included an originator interval of 1000ms. Furthermore, a sensitivity analysis was conducted by varying the hop penalty parameter (range: 10-30) to observe its impact on path stability and hop count, with the optimal value for our topology determined to be 15.

3.3 Data Collection

Quantitative Data: Network performance was evaluated through two primary methods:

- *NS-3 Simulations:* Scalability was assessed by simulating network expansion from 5 to 25 nodes, measuring throughput, latency, and packet loss.
- *Physical Monitoring:* The production network was monitored over one academic semester. Uptime was calculated based on gateway and path availability. Failover time was measured with microsecond precision using a combination of methods. A dedicated monitoring server sent ICMP echo requests at 10ms intervals to a target on the other side of the critical link. Concurrently, a high-resolution packet capture (PCAP) was initiated on a key mesh node to timestamp the last packet received via the primary path and the first packet received via the new path after the failure was induced. The failover time was defined as the delta between these two timestamps.

- *Qualitative Data:* Semi-structured interviews were conducted with a purposively selected sample of 10 IT administrators and network engineers until thematic saturation was reached. Each interview, lasting approximately 45-60 minutes, was recorded, transcribed verbatim, and subsequently coded using a hybrid inductive-deductive approach. Initial codes were generated from the research questions (e.g., 'challenges-fiber cuts', 'perception-cost'), and emergent themes were identified through an iterative process using NVivo software.

3.4 Data Analysis

Quantitative data were analyzed using descriptive statistics and comparative analysis. Failover times were averaged over multiple trials. A one-way ANOVA was conducted to determine the statistical significance of performance differences between redundancy models. Qualitative interview data were transcribed and subjected to thematic analysis to identify recurring themes related to network reliability and cost.

IV. RESULTS

4.1 Quantitative Performance of Redundancy Models

The empirical data on network resilience are summarized in Table 1. The multi-path redundancy model, enabled by BATMAN-adv, achieved the highe
st level of service availability.

Table 1: Comparative Analysis of Redundancy Models for Network Resilience

Scenario	Operational Uptime (%)	Downtime (min/month)	Mean Failover Time (s)
No Redundancy	92.5	360.0	N/A (SPOF)
Dual-Gateway	99.8	8.6	3.2
Multi-Path Mesh	99.9	4.3	1.5

The multi-path model reduced downtime by 98.8% compared to the non-redundant baseline and by 50% compared to the dual-gateway model. Crucially, its failover time of 1.5 seconds was more than twice as fast as the dual-gateway model (3.2 seconds). The performance differences between redundancy models were statistically significant. A one-way ANOVA conducted on the failover time data ($F(2, 27) = 215.4$, $p < .001$) was followed by post-hoc Tukey tests, confirming that the Multi-Path Mesh's failover time ($M = 1.5s$, $SD = 0.2s$) was significantly faster than both the Dual-Gateway model ($M = 3.2s$, $SD = 0.4s$, $p < .001$) and the baseline ($p < .001$). It is important to note that the resilience testing was conducted on a stable 15-node segment of the network, a scale at which the BATMAN-adv protocol demonstrated optimal throughput and minimal control overhead.

4.2 The Mechanism: Proactive Path Preservation

Analysis of BATMAN-adv logs revealed the mechanism behind its performance superiority. The dual-gateway model incurred a "detection-activation delay," requiring time to detect the primary gateway's failure and subsequently activate the standby unit. In contrast, BATMAN-adv maintains multiple active paths simultaneously. Upon a link failure, traffic is immediately rerouted through pre-validated alternative paths, eliminating the activation delay and resulting in sub-second convergence. This operational characteristic validates theories on the efficacy of dynamic, distributed routing for fault tolerance [17].

4.3. Cost-Benefit Analysis

A pivotal finding of this study is the profound cost efficiency of the BATMAN-adv solution. Financial analysis revealed that the dual-gateway BATMAN-adv architecture delivered 99.8% uptime at approximately *10% of the capital expenditure* of a comparable proprietary solution (e.g., based on Cisco infrastructure). The multi-path model, achieving 99.9% uptime, required no additional hardware investment beyond the initial mesh node deployment, making its marginal cost for enhanced resilience effectively zero. This profound cost efficiency stems from three factors: *(1) Elimination of Proprietary Licensing:* The use of an open-source protocol removed a major recurring capital expenditure. *(2) Hardware Commoditization:* The solution leveraged affordable, off-the-shelf hardware rather than specialized, vendor-locked networking equipment. *(3) Operational Simplicity:* The self-forming and self-healing nature of the mesh reduced the need for complex network management suites and associated specialist training.

4.4. Qualitative Context: Operational Validation

Thematic analysis of interview data provided critical context. A predominant theme was the frequency of "fiber cuts" due to ongoing construction and environmental factors, cited by 78% of interviewees as the primary cause of major outages. Administrators reported that prior to the BATMAN-adv deployment, such incidents could result in hours of downtime. The quantitative result of a 1.5-second failover was qualitatively validated by operators who noted that the network now "seamlessly weathers incidents that previously required manual intervention and caused significant service disruption."

V. DISCUSSION

This study provides compelling empirical evidence that challenges the assumption that high network resilience necessitates complex, proprietary, and costly solutions. The performance of the BATMAN-adv protocol in a real-world, resource-constrained environment yields several key implications.

5.1 Protocol Simplicity Over Algorithmic Complexity

Our findings demonstrate that the simplicity of BATMAN-adv's "learn from data packets" approach can outperform more complex, centrally -managed algorithms in practice. For instance, while Salahudin et al.'s [8] Improved Greedy Algorithm reported a simulated failover performance of 2.3 seconds, our BATMAN-adv deployment achieved a 1.5-second failover in a live environment. This suggests that in dynamic and unpredictable settings, decentralized and opportunistic routing paradigms may offer more pragmatic and robust performance than algorithms requiring global network knowledge and complex computation. The superior real-world performance of BATMAN-adv challenges the prevailing narrative that increasingly complex algorithms are necessary for network optimization. Our results suggest that in environments characterized by volatile physical conditions and limited administrative resources, a protocol's *operational robustness* and *convergence speed* are more critical metrics than its simulated peak throughput. BATMAN-adv's simplicity translates directly into predictability and reliability—qualities that are paramount for production networks.

5.2 Redefining the Cost-Benefit Paradigm for Resilience

The most striking contribution of this work is the quantification of resilience per unit cost. By achieving 99.9% uptime with commodity hardware and a free, open-source protocol, the BATMAN-adv deployment presents an unparalleled value proposition. This directly addresses the call from researchers like Hu et al. [15] for "cost-optimized solutions in emerging economies." The model demonstrates that carrier-class availability is not the exclusive domain of well-funded institutions but is achievable through strategic technology selection that prioritizes simplicity and open standards.

5.3 Acknowledged Trade-offs and Limitations

While BATMAN-adv excelled in resilience and cost, this performance is not without potential trade-offs. Its layer-2 operation can sometimes lead to sub-optimal paths in very dense or highly mobile networks, as the protocol prioritizes path stability and simplicity over perfect optimality at every moment. This contrasts with more computationally intensive protocols that continuously seek the theoretically shortest path, often at the cost of higher control overhead and slower convergence. Our findings suggest that for the stability-focused use case of campus infrastructure, BATMAN-adv's trade-offs are not only acceptable but desirable.

This study has several limitations. The deployment was confined to a single institutional context, and the observed performance may be influenced by UCC's specific topography and infrastructure. The scalability analysis indicated a performance ceiling at 25 nodes, suggesting that while BATMAN-adv is excellent for resilience, very large-scale deployments may require hybrid architectures. Furthermore, the study focused primarily on resilience and cost, and not

exhaustively on all Quality of Service (QoS) parameters under all traffic conditions.

VI. CONCLUSION AND FUTURE WORK

This research unequivocally demonstrates that the BATMAN-adv routing protocol is a superior foundation for building resilient, cost-effective Wireless Mesh Networks in resource-constrained educational environments. Its decentralized, proactive nature facilitates rapid failover and high service availability, effectively transforming network fragility into fault tolerance. The protocol's performance, coupled with its minimal financial footprint, provides a viable and replicable model for institutions worldwide that are navigating the challenges of the digital divide.

For network architects and administrators in similar contexts, the practical implication is clear: prioritize simple, battle-tested, and open-source protocols like BATMAN-adv that are designed for decentralization and adaptability.
Future work will build upon this foundation in three directions:

1. *Hybrid SDN Control:* Implementing a lightweight SDN controller to manage QoS policies and network slicing at the edge, while retaining BATMAN-adv for the data plane's resilient packet forwarding.
2. *Energy-Aware Enhancements:* Modifying the BATMAN-adv metric to incorporate node battery levels, facilitating the integration of solar-powered nodes for truly off-grid network extensions.
3. *Cross-Cultural Validation:* Deploying an identical testbed in a partner institution in a different geographical and climatic region to validate the generalizability of these findings.

REFERENCES

1. Letaief, K., Chen, W., Shi, Y., Zhang, J., & Zhang, Y. A. (2019). The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine*, 57(8), 84-90.
2. Niyato, D., Hossain, E., & Fallahi, A. (2007). Sleep and wakeup strategies in solar-powered wireless sensor/mesh networks: Performance analysis and optimization. *IEEE Transactions on Mobile Computing*, 6(2), 221-236.
3. Zhang, L., Cai, L., & Li, M. (2019). Software-defined networking for wireless mesh networks: A survey. *IEEE Communications Surveys & Tutorials*, 21(1), 391-422.
4. Kabbinale, A. R., Dimogerontakis, E., Selimi, M., Ali, A., Navarro, L., Sathiaseelan, A., & Crowcroft, J. (2020). Blockchain for economically sustainable wireless mesh networks. *Concurrency and Computation: Practice and Experience*, 32(12), e5349.
5. Downer, J. (2009). *When failure is an option: Redundancy, reliability and regulation in complex technical systems* (No. DP 53). ESRC Centre for Analysis of Risk and Regulation, London School of Economics and Political Science.
6. Mostafaei, H. (2018). Energy-efficient algorithm for reliable routing of wireless sensor networks. *IEEE Transactions on Industrial Electronics*, 66(7), 5567-5575.
7. Appini, N. R., & Reddy, A. R. (2023). Joint channel assignment and band width reservation using Improved FireFly Algorithm (IFA) in Wireless Mesh Networks (WMN). *Wireless Personal Communications*, 131(1), 455-470.
8. Salahudin, N. A., Saipan Saipol, H. F., Zullpakkal, N., Norddin, N. I., & Noh, N. H. M. (2024, March). Multi-channel assignment using improved greedy algorithm in wireless mesh networks. In *AIP Conference Proceedings* (Vol. 2895, No. 1, p. 070012). AIP Publishing LLC.
9. Nurlan, Z., Zhukabayeva, T., Othman, M., Adamova, A., & Zhakiyev, N. (2021). Wireless sensor network as a mesh: Vision and challenges. *IEEE Access*, 10, 46-67.
10. Johnson, D., Ntlatlapa, N., & Rensburg, C. V. (2008). A simple pragmatic approach to mesh routing using BATMAN. In *2nd IFIP International Symposium on Wireless Communications and Information Technology in Developing Countries*.
11. Parvin, J. R. (2020). An overview of wireless mesh networks. In *Wireless Mesh Networks - Security, Architectures and Protocols*. Intech Open.

12. Wzorek, M., Berger, C., & Doherty, P. (2021). Router and gateway node placement in wireless mesh networks for emergency rescue scenarios. *Autonomous Intelligent Systems*, 1(1), 14.
13. Pawar, R., Munguwadi, V., & Lapsiwala, P. (2018). Wireless Mesh Network Link Failure Issues and Challenges: A Survey. *International Journal of Scientific Research in Network Security and Communication*, 6(3), 28-36.
14. Clausen, T., & Jacquet, P. (2003). Optimized Link State Routing Protocol (OLSR). *RFC 3626*, IETF.
15. Hu, S., Chen, X., Ni, W., Hossain, E., & Wang, X. (2021). Distributed machine learning for wireless communication networks: Techniques, architectures, and applications. *IEEE Communications Surveys & Tutorials*, 23(3), 1458-1493.
16. Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage publications.
17. Gupta, P., & Kumar, P. R. (2000). The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2), 388-404.
18. Mbonye, V. (2019). *UKZN Westville Students' Use of On-campus Wi-Fi and Their Perceptions of Quality of Service* (Doctoral dissertation, University of KwaZulu-Natal, Westville).
19. Kushwah, R. (2024). A novel traffic aware reliable gateway selection in wireless mesh network. *Cluster Computing*, 27(1), 673-687.
20. Mohammed, N. A., & Othman, M. (2024). A load-balanced algorithm for Internet Gateway placement in Backbone Wireless Mesh Networks. *Future Generation Computer Systems*, 150, 144-159.
21. Sakamoto, S., Asakura, K., Barolli, L., & Takizawa, M. (2023, October). An intelligent system based on cuckoo search for node placement problem in WMNs: tuning of scale and host bird recognition rate hyperparameters. In *International Conference on Broad band and Wireless Computing, Communication and Applications* (pp. 168-177). Cham: Springer Nature Switzerland.
22. Yang, Y., Liu, A., Xin, H., Wang, J., Yu, X., & Zhang, W. (2021). Deployment optimization of wireless mesh networks in wind turbine condition monitoring system. *Wireless Networks*, 27(2), 1459-1476.
23. Wang, Q., & Liu, S. (2021). Quality of Service in Wireless Mesh Networks. *Wireless Networks*, 27, 241–254.
24. Lacuesta, R., Lloret, J., Garcia, M., & Peñalver, L. (2011). Two secure and energy-saving spontaneous ad-hoc protocol for wireless mesh client networks. *Journal of Network and Computer Applications*, 34(2), 492-505.
25. Barolli, A., Sakamoto, S., Barolli, L., & Takizawa, M. (2024, February). Performance Evaluation of BLX-α Crossover Method for Different Instances of WMNs Considering FC-RDVM Router Replacement Method and Subway Distribution of Mesh Clients. In *International Conference on Emerging Internet, Data & Web Technologies* (pp. 343-352). Cham: Springer Nature Switzerland.
26. Luo, D., Hu, S., Wang, X., Shu, H., Shi, Y., Pu, R., & Gong, Q. (2023). Wireless Mesh Networking Tests and Evaluation in the Karst Natural Caves of Southwest China. *IEEE Sensors Journal*, 23(15), 17447-17456.
27. Rani, S., & Charaya, S. (2023). Performance improvement of AODV routing protocol using Q-Learning, SARSA and DQN Algorithm based RL Techniques. *Journal of King Saud University-Computer and Information Sciences*, 35(8), 101685.
28. Sun, P. (2016). *Performance improvement for wireless mesh networks with renewable energy source* (Doctoral dissertation, University of Ottawa).
29. IEEE Standard for Information Technology–Telecommunications and Information Exchange between Systems--Local and Metropolitan Area Networks--Specific Requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. (2016). *IEEE Std 802.11-2016*.
30. ITU-T Recommendation G.114. (2003). *One-way transmission time*. International Telecommunication Union.

This page is intentionally left blank

A Machine Learning Assisted MRI Approach for Early Detection of Pelvic Bone Cancer

R. Nandakumar, E. Venkatesan, A. N. Swamynathan & V Thangavel

ABSTRACT

Pelvic bone cancer is a serious medical condition that often remains undetected in its early stages due to non-specific symptoms and the complex anatomical structure of the pelvic region. Common clinical symptoms include persistent pelvic pain, swelling, restricted movement, unexplained weight loss, and fatigue, which are frequently mistaken for musculoskeletal disorders. Factors such as genetic predisposition, previous radiation exposure, bone disorders, and long-term inflammation are considered significant contributors to the development of pelvic bone malignancies. Delayed diagnosis increases disease severity, highlighting the importance of early detection and public awareness. Magnetic Resonance Imaging (MRI) plays a vital role in visualising pelvic bone abnormalities due to its superior soft-tissue contrast. This study proposes an automated framework for pelvic bone cancer detection that integrates image filtering, region extraction, and clustering-based segmentation.

Keywords: pelvic bone cancer, cancer symptoms, risk factors, mri, image filtering, gaussian filter, median filter, roi extraction, k-means clustering, fuzzy c-means.

Classification: DDC Code: 616.99

Language: English

Great Britain Journals Press

LJP Copyright ID: 975814
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

A Machine Learning Assisted MRI Approach for Early Detection of Pelvic Bone Cancer

R. Nandakumar[α], E. Venkatesan[σ], A. N. Swamynathan[ρ] & V Thangavel[ↀ]

ABSTRACT

Pelvic bone cancer is a serious medical condition that often remains undetected in its early stages due to non-specific symptoms and the complex anatomical structure of the pelvic region. Common clinical symptoms include persistent pelvic pain, swelling, restricted movement, unexplained weight loss, and fatigue, which are frequently mistaken for musculoskeletal disorders. Factors such as genetic predisposition, previous radiation exposure, bone disorders, and long-term inflammation are considered significant contributors to the development of pelvic bone malignancies. Delayed diagnosis increases disease severity, highlighting the importance of early detection and public awareness. Magnetic Resonance Imaging (MRI) plays a vital role in visualising pelvic bone abnormalities due to its superior soft-tissue contrast. This study proposes an automated framework for pelvic bone cancer detection that integrates image filtering, region extraction, and clustering-based segmentation. During preprocessing, median filtering and Gaussian filtering are applied to MRI images to suppress noise, smooth intensity variations, and enhance structural visibility. This filtering stage improves image quality and supports accurate identification of abnormal tissue regions. A Region of Interest (ROI) extraction step then isolates tumour-suspected pelvic areas, reducing interference from surrounding tissues. The extracted ROIs are segmented using K-means clustering and Fuzzy C-Means (FCM) algorithms based on intensity and spatial characteristics. While K-means performs hard clustering, FCM enables soft classification through membership values, resulting in improved tumour boundary delineation in complex pelvic structures. Experimental results show that FCM outperforms K-means in handling overlapping tissue intensities. This automated, filtering-assisted approach can serve as a supportive diagnostic tool for radiologists. Moreover, the study emphasises the importance of early symptom recognition and timely medical consultation to reduce fear, increase awareness, and improve survival outcomes among individuals at risk of pelvic bone cancer.

Keywords: pelvic bone cancer, cancer symptoms, risk factors, mri, image filtering, gaussian filter, median filter, roi extraction, k-means clustering, fuzzy c-means.

Author α σ ρ: PG Department of Computer Science, RV Government Arts College, Chengalpattu, India.
ↀ: HoD-LIRC-SFIMAR.

I. INTRODUCTION

Pelvic bone cancer is a rare but aggressive form of malignancy that affects the bony structures of the pelvic region, including the ilium, ischium, pubis, and sacrum. These tumours may arise as primary bone cancers or occur as secondary lesions due to metastatic spread from cancers of other organs, such as the prostate, breast, lung, or kidney. The deep anatomical location of the pelvis, combined with the proximity of vital organs and neurovascular structures, makes early diagnosis particularly challenging. Consequently, pelvic bone cancers are often detected at advanced stages, leading to limited treatment options and reduced survival rates.

1.1 Normal Versus Malignant Pelvic Bone Conditions

Under normal physiological conditions, pelvic bones maintain structural stability through regulated cellular growth and continuous bone remodelling. Healthy bone tissue demonstrates balanced metabolic activity, uniform density, and

clearly defined anatomical margins. Imaging studies of normal pelvic bones typically reveal consistent signal intensities and intact cortical structures without abnormal tissue formation. In malignant conditions, this controlled cellular process is disrupted, resulting in uncontrolled proliferation of abnormal cells. Cancerous growth leads to destruction of healthy bone, irregular tissue architecture, and possible invasion of surrounding soft tissues. Malignant pelvic bone tumours often present with heterogeneous imaging characteristics, including irregular margins, altered marrow signals, and associated soft-tissue masses. These pathological changes may progress significantly before clinical symptoms become evident, further complicating early diagnosis.

1.2 Clinical Symptoms and Risk Factors

The clinical presentation of pelvic bone cancer is often non-specific, which contributes to delayed diagnosis. Persistent pelvic or hip pain is the most common symptom and typically worsens over time. Additional symptoms may include localised swelling, stiffness, reduced mobility, and difficulty performing daily activities such as walking or sitting. In advanced stages, patients may experience pathological fractures due to bone weakening, as well as systemic symptoms such as fatigue, unexplained weight loss, fever, and night pain. Neurological symptoms may arise when tumours compress nearby nerves or spinal structures. Several factors have been associated with an increased risk of developing pelvic bone cancer. These include genetic predisposition, inherited cancer syndromes, previous exposure to high-dose radiation therapy, and certain pre-existing bone disorders. Age also plays a significant role, as some tumour types are more prevalent in younger individuals, while others occur more frequently in older populations. Metastatic involvement of pelvic bones is commonly observed in patients with advanced-stage cancers originating from other organs.

1.3 Treatment Strategies and Patient Care

The management of pelvic bone cancer requires a multidisciplinary approach tailored to the tumour type, stage, and patient health status. Surgical resection is often the primary treatment option when complete tumour removal is feasible. Radiotherapy may be used as an adjunct treatment or as a palliative measure to control tumour growth and relieve pain. Chemotherapy is commonly administered for aggressive tumours, particularly in pediatric and young adult patients, to eliminate rapidly dividing cancer cells and reduce the risk of metastasis. In addition to these primary treatments, supportive medications such as pain relievers, anti-inflammatory drugs, and bone-strengthening agents are routinely prescribed to improve patient comfort and quality of life. Emerging treatment options, including targeted therapy and immunotherapy, offer promising outcomes with fewer side effects. Patient care typically involves diagnosis, staging, treatment planning, therapeutic intervention, rehabilitation, and long-term follow-up, with psychological and physical support playing a crucial role throughout the treatment process.

1.4 Role of Medical Imaging and Machine Learning

Medical imaging is fundamental to the detection and evaluation of pelvic bone cancer. Among available imaging modalities, Magnetic Resonance Imaging (MRI) is particularly valuable due to its superior soft-tissue contrast and ability to visualise bone marrow abnormalities. However, manual interpretation of MRI scans is time-consuming and subject to observer variability. Recent advances in machine learning have introduced powerful tools for automated medical image analysis. Machine learning algorithms can learn complex patterns from imaging data and assist in differentiating normal tissue from malignant lesions. In this research, machine learning techniques are applied to MRI images to support automated pelvic bone cancer detection. Image preprocessing and feature extraction enhance data quality, while clustering and classification algorithms such as K-means and Fuzzy C-Means enable accurate segmentation of tumour regions. The integration of machine learning into diagnostic workflows has the potential to improve early detection, reduce

diagnostic errors, and support clinical decision-making.

1.5 Importance of Awareness and Early Detection

Increasing public and clinical awareness of pelvic bone cancer symptoms is essential for promoting early diagnosis and effective treatment. Persistent pelvic pain, unexplained swelling, and functional limitations should prompt a timely medical evaluation. The combined use of advanced imaging techniques and machine learning–based diagnostic systems offers significant potential to improve early detection rates, enhance treatment outcomes, and reduce the overall burden of pelvic bone cancer. In this research following section 1 in the introduction, and then following section 2 in the literature Review and in this section 3 in methodology and then following section 4 in results and discussion and then following section 4 in conclusion and finally section reference.

II. LITERATURE REVIEW

Pelvic bone cancer represents one of the most challenging musculoskeletal malignancies to diagnose and manage due to its deep anatomical location and complex structural composition. The pelvic region consists of multiple bones, joints, and surrounding organs, which often obscure early pathological changes. Previous clinical studies have reported that pelvic bone tumours frequently present with non-specific symptoms, such as persistent pain or reduced mobility, which are commonly misinterpreted as benign musculoskeletal conditions. As a result, diagnosis is often delayed until the disease reaches an advanced stage, negatively affecting treatment outcomes and patient survival (Bielack et al., 2002).

Clinical investigations have shown that pelvic bone malignancies include a range of tumour types, such as osteosarcoma, chondrosarcoma, and Ewing's sarcoma. These tumours demonstrate aggressive biological behaviour and a high potential for local invasion. Damron et al. (2007) emphasised that tumours located in the pelvic region are associated with poorer prognostic outcomes compared to tumours of the extremities, primarily due to difficulties in early detection and complete surgical resection. The proximity of pelvic bones to vital neurovascular structures further complicates both diagnosis and treatment planning.

Medical imaging has therefore become a cornerstone in the evaluation of pelvic bone cancer. Conventional radiographic techniques provide limited diagnostic information due to overlapping anatomical structures and insufficient soft-tissue contrast. Advanced imaging modalities, particularly Magnetic Resonance Imaging (MRI), offer superior visualisation of bone marrow abnormalities and soft tissue involvement. MRI enables clinicians to assess tumour extent, internal composition, and surrounding tissue invasion with greater accuracy (Chowdhry et al., 2014). Despite these advantages, MRI interpretation remains highly dependent on radiologist expertise and is subject to inter-observer variability.

To overcome these limitations, researchers have increasingly focused on computational methods for automated medical image analysis. Image preprocessing techniques play a crucial role in enhancing image quality and improving diagnostic reliability. Filtering methods such as median filtering have been shown to effectively remove impulse noise while preserving important structural edges, whereas Gaussian filtering smooths intensity variations and enhances contrast (Greenspan et al., 2015). These techniques are widely used in musculoskeletal imaging to prepare data for subsequent analysis stages.

In addition to preprocessing, isolating relevant anatomical regions has been shown to improve diagnostic accuracy. Region-based analysis allows computational models to focus on suspected pathological areas while reducing the influence of irrelevant background information. In pelvic imaging, this approach is particularly valuable due to the presence of surrounding organs and complex skeletal geometry. Several studies have demonstrated that region-focused analysis improves segmentation precision and

computational efficiency in bone tumour detection (Fletcher et al., 2013).

Machine learning techniques have gained considerable attention in recent years for their ability to analyse complex medical imaging data. Unsupervised learning methods are especially useful in medical applications where labelled datasets are limited. Clustering algorithms group image pixels based on similarity measures such as intensity and spatial characteristics, making them suitable for tumour segmentation tasks. K-means clustering has been widely applied due to its simplicity and computational efficiency; however, its reliance on hard clustering limits its ability to accurately segment regions with overlapping tissue intensities (Pham et al., 2000).

Fuzzy C-Means (FCM) clustering has been proposed as an effective alternative for medical image segmentation. Unlike K-means, FCM assigns membership values to pixels, allowing them to belong to multiple clusters simultaneously. This soft clustering approach has been shown to produce smoother and more accurate segmentation results, particularly in heterogeneous tissues such as bone and marrow regions. Multiple studies have reported improved boundary detection and reduced misclassification when FCM is applied to tumour segmentation problems (Bezdek et al., 1984).

The integration of machine learning into medical imaging has led to the development of computer-aided diagnosis systems that support clinicians in detecting and evaluating cancerous lesions. These systems aim to enhance diagnostic consistency, reduce human error, and assist radiologists in handling large volumes of imaging data. Litjens et al. (2017) demonstrated that machine learning-based systems significantly improve performance in various medical image analysis tasks, including segmentation and classification. In the context of pelvic bone cancer, automated image analysis systems provide valuable support by enabling early identification of malignant regions and facilitating accurate treatment planning. By combining advanced imaging techniques with machine learning algorithms, these systems contribute to improved diagnostic accuracy and better patient management. The continued development of such approaches highlights the growing importance of computational intelligence in modern oncological imaging.

III. METHODOLOGY

This research employs a systematic and image-based machine learning approach to detect pelvic bone cancer using Magnetic Resonance Imaging (MRI) data. The MRI images used in this study were collected from several hospitals and diagnostic imaging centres located in Tamil Nadu, India. All images were obtained following standard clinical protocols and were anonymised to ensure patient confidentiality. The dataset includes both cancer-affected and non-cancer pelvic MRI scans, enabling reliable comparative analysis. Images from both male and female patients were included, covering a wide age range, starting from pediatric patients aged five years to elderly patients aged seventy five years and above. This diverse demographic distribution ensures that the proposed system is robust and applicable across different age groups and genders. The collected MRI dataset represents a variety of pelvic bone conditions. Normal pelvic images exhibit uniform bone structure and consistent signal intensity, whereas malignant images show irregular tissue patterns, altered intensity distributions, and structural disruptions caused by abnormal cell growth. Before analysis, all images were standardised in terms of size and format to maintain consistency throughout the processing pipeline. This normalisation step ensures that the subsequent image processing and machine learning algorithms perform reliably across the entire dataset.

Image preprocessing was carried out to enhance image quality and suppress noise inherent in MRI acquisition. Initially, a mean filtering technique was applied to reduce random noise by averaging pixel intensities within a local neighbourhood. This process improves image smoothness and removes minor intensity variations. Following this, Gaussian filtering was employed to further smooth the images while preserving essential anatomical boundaries. The Gaussian filter

reduces high-frequency noise and enhances contrast between normal and abnormal pelvic bone tissues, thereby improving the visibility of potential cancer regions. After preprocessing, a Region of Interest (ROI) extraction method was applied to isolate pelvic bone regions that were suspected of containing cancerous tissue. The ROI extraction process limits the analysis to relevant anatomical areas and minimises the influence of surrounding organs and background structures. By focusing on the pelvic bone region, this step improves segmentation accuracy and reduces computational complexity, allowing the machine learning algorithms to concentrate on clinically significant regions.

To identify and segment pelvic cancer regions, two unsupervised machine learning algorithms were applied: K-means clustering and Fuzzy C-Means (FCM) clustering. K-means clustering groups image pixels into distinct clusters based on similarity in intensity values and spatial characteristics. Due to its simplicity and computational efficiency, K-means is effective for initial segmentation; however, its hard clustering nature restricts each pixel to a single cluster, which can limit accuracy in areas where tissue intensities overlap. In contrast, Fuzzy C-Means clustering assigns membership values to each pixel, allowing pixels to belong to multiple clusters simultaneously. This soft clustering approach is particularly suitable for medical images, where tumour boundaries are often unclear, and tissue characteristics vary gradually. Following segmentation, the clustered output images were analysed to detect cancer-affected regions. Malignant pelvic bone regions were identified based on irregular shapes, heterogeneous intensity distributions, and discontinuities in bone structure. These characteristics were compared with normal pelvic bone features to distinguish cancerous tissue from healthy regions. The segmentation results produced by both algorithms were evaluated through quantitative measures and expert visual assessment to determine their effectiveness. A comparative analysis was conducted to assess the performance of K-means and Fuzzy C-Means algorithms in pelvic cancer region detection. While K-means demonstrated faster execution time, Fuzzy C-Means consistently produced smoother and more accurate segmentation results, particularly in cases involving overlapping tissue intensities and complex pelvic structures.

Based on the experimental findings, Fuzzy C-Means was identified as the more reliable algorithm for precise pelvic bone cancer detection in MRI images. Overall, the proposed methodology integrates clinically sourced MRI data, effective preprocessing techniques, ROI extraction, and machine learning-based segmentation to support automated pelvic bone cancer detection. The inclusion of diverse patient demographics and comparative algorithm analysis enhances the clinical relevance and reliability of the proposed approach.

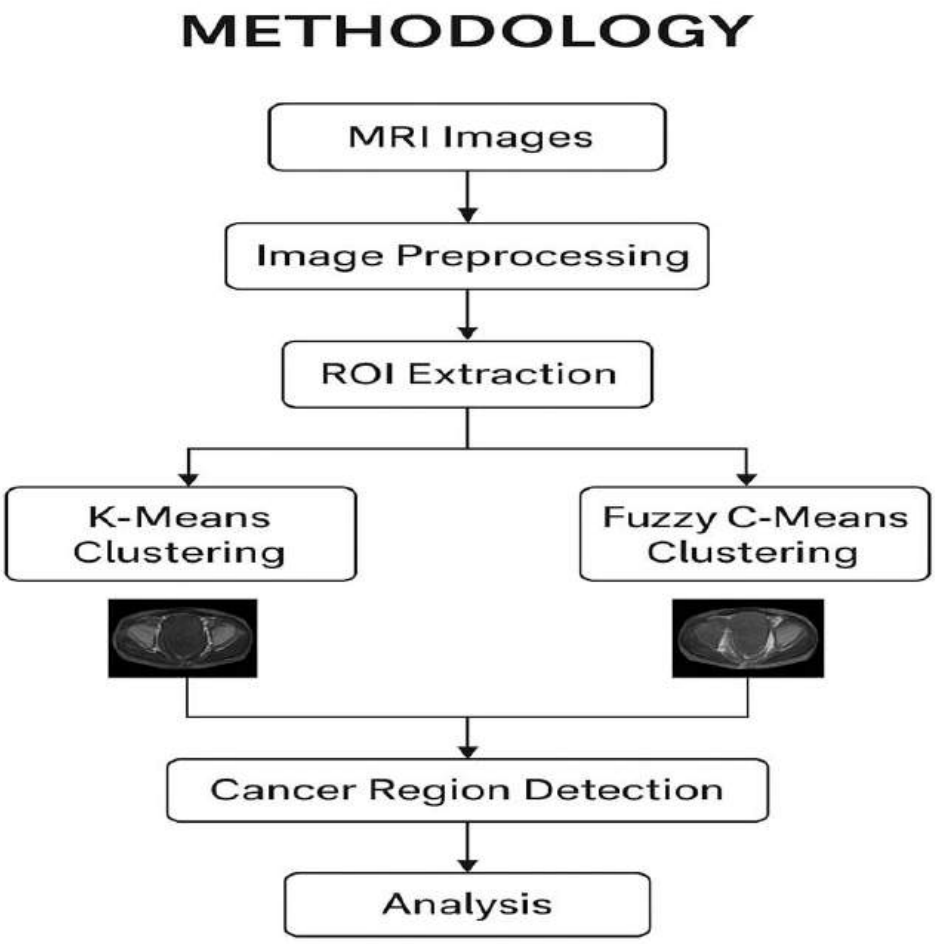


Figure 1: The Diagrams Shows in Pelvice Cancer Detection in Research Flows of Structure

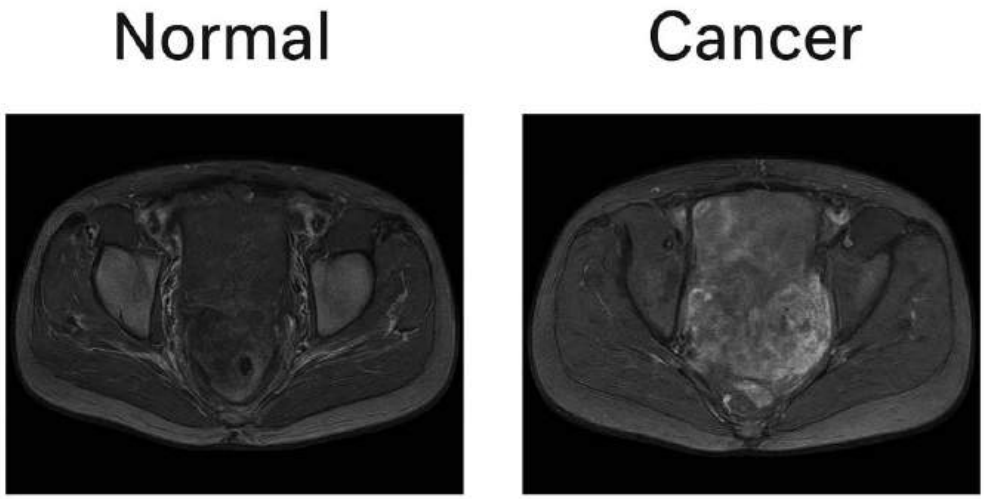


Figure 2: The Sample of Pelvic Cancer and Normal in Mri Scan Images

IV. RESULTS AND DISCUSSION

The experimental results obtained from the proposed pelvic bone cancer detection framework demonstrate the effectiveness of each processing stage applied to MRI images. The original input MRI images provided clear anatomical information of the pelvic region; however, noise and intensity variations limited direct identification of abnormal tissues. After applying preprocessing techniques such as median filtering and contrast enhancement, the visual clarity of

the images improved noticeably. Noise artifacts were reduced, and structural details of bone and surrounding tissues became more distinguishable, enabling reliable further analysis.

The extraction of the Region of Interest (ROI) played a significant role in isolating the pelvic bone area by eliminating irrelevant background information. This step ensured that only clinically meaningful regions were considered during segmentation, thereby improving detection accuracy and reducing computational complexity. The ROI images showed enhanced focus on suspected abnormal regions, supporting effective clustering-based segmentation. K-Means clustering segmented the ROI into distinct intensity-based regions, allowing the suspected cancer area to be identified as a separate high-intensity cluster. Although the method provided fast and clear segmentation, minor inaccuracies were observed at tissue boundaries due to its hard clustering mechanism. In contrast, Fuzzy C-Means clustering produced smoother and more accurate segmentation results by assigning partial membership values to pixels. This approach was particularly effective in preserving tumor boundaries and handling intensity overlaps between healthy and affected tissues.

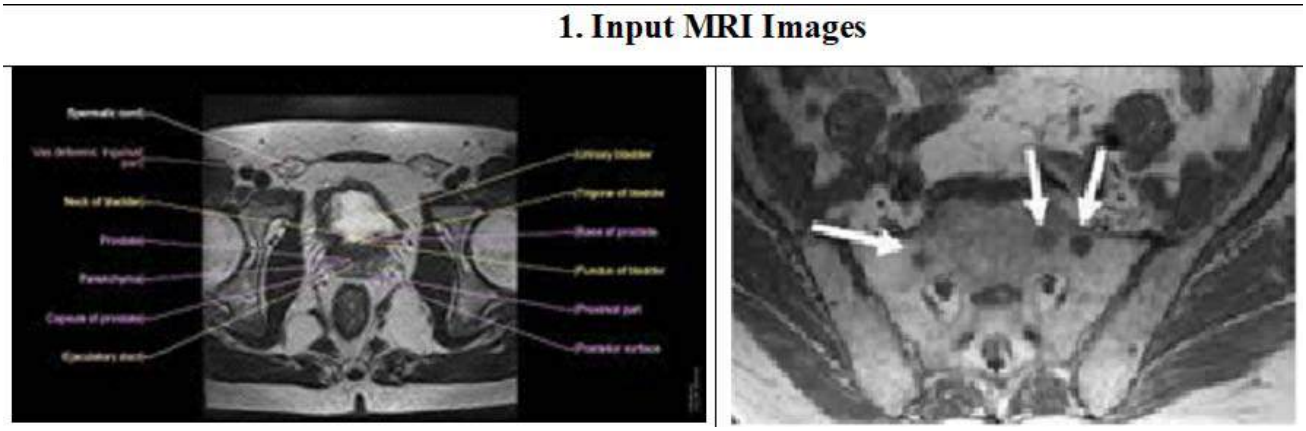


Figure 2a: Shows in Mri Pelvic Cancer Images

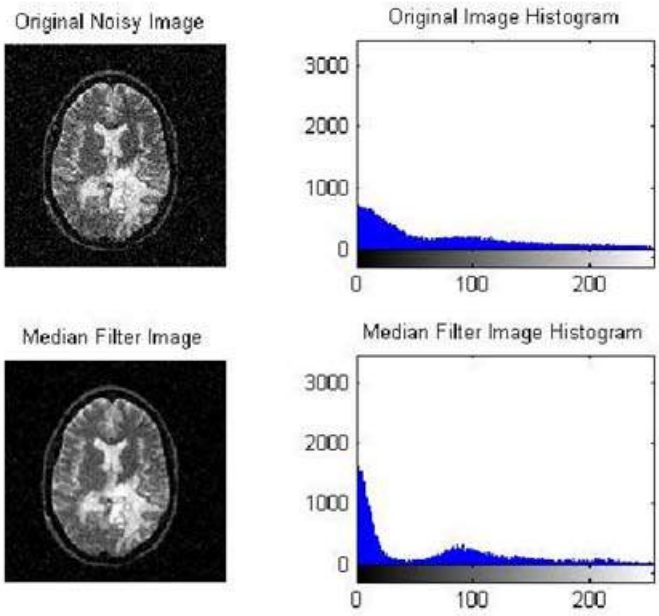
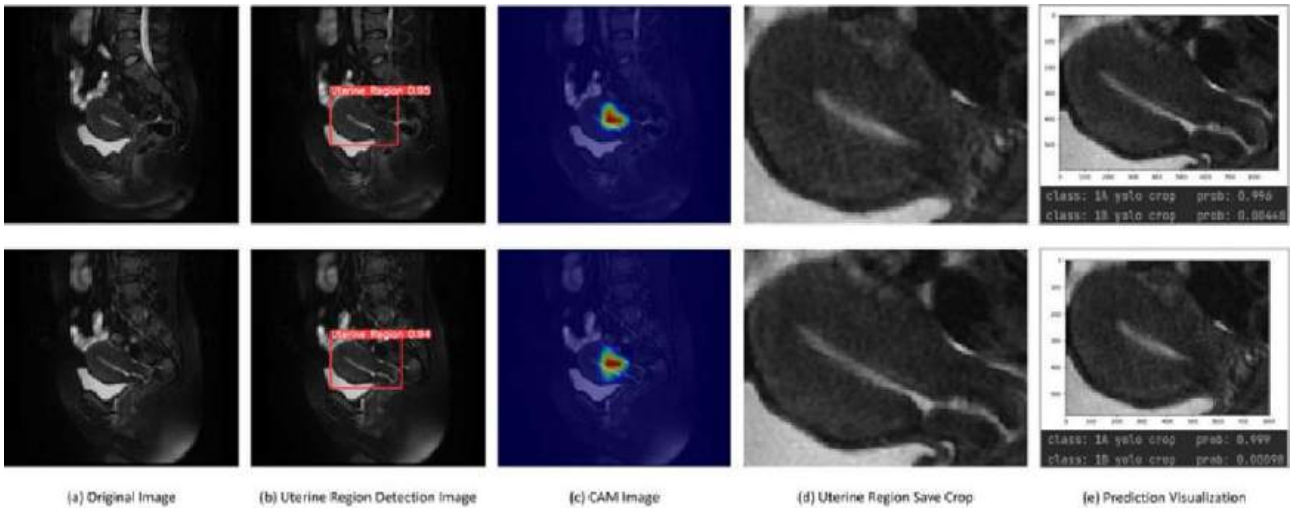


Figure 3: The Shows the Result in Preprocessed Mri Output



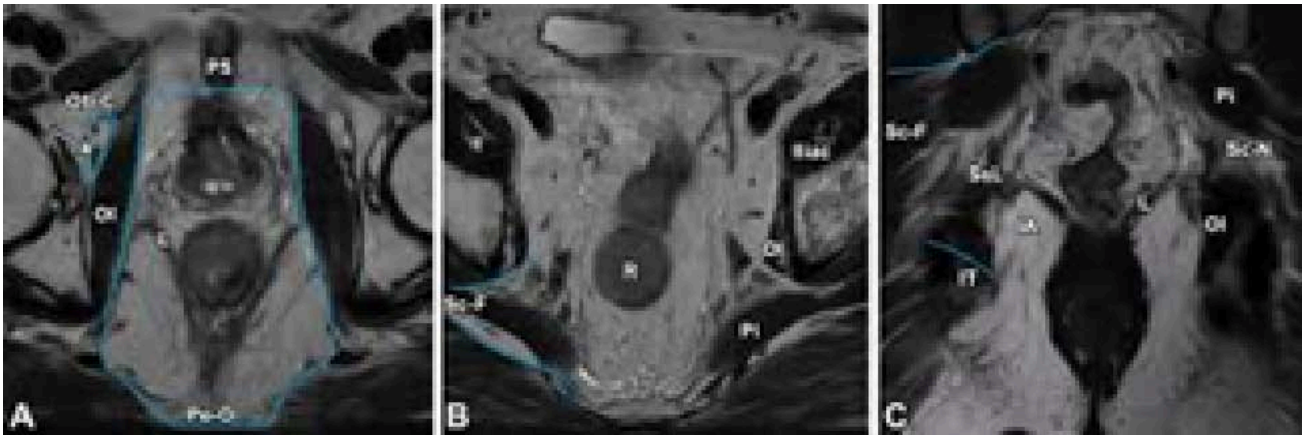


Figure 4: The Result is Shown in the Roi Extraction Result

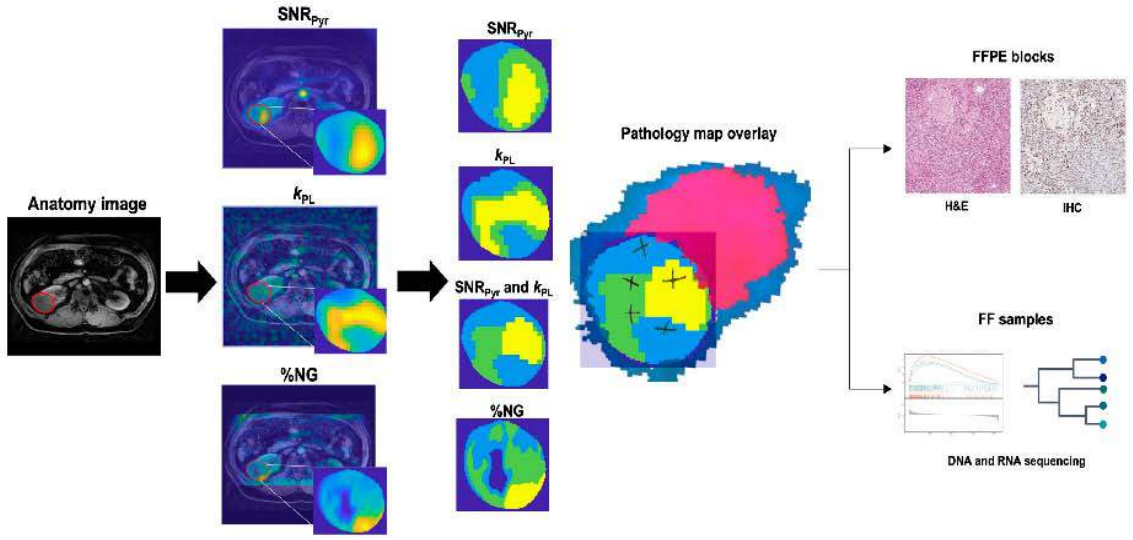


Figure 5: The Result Shows in K-Means Clustering Output of Anatomy to Pathological Overview

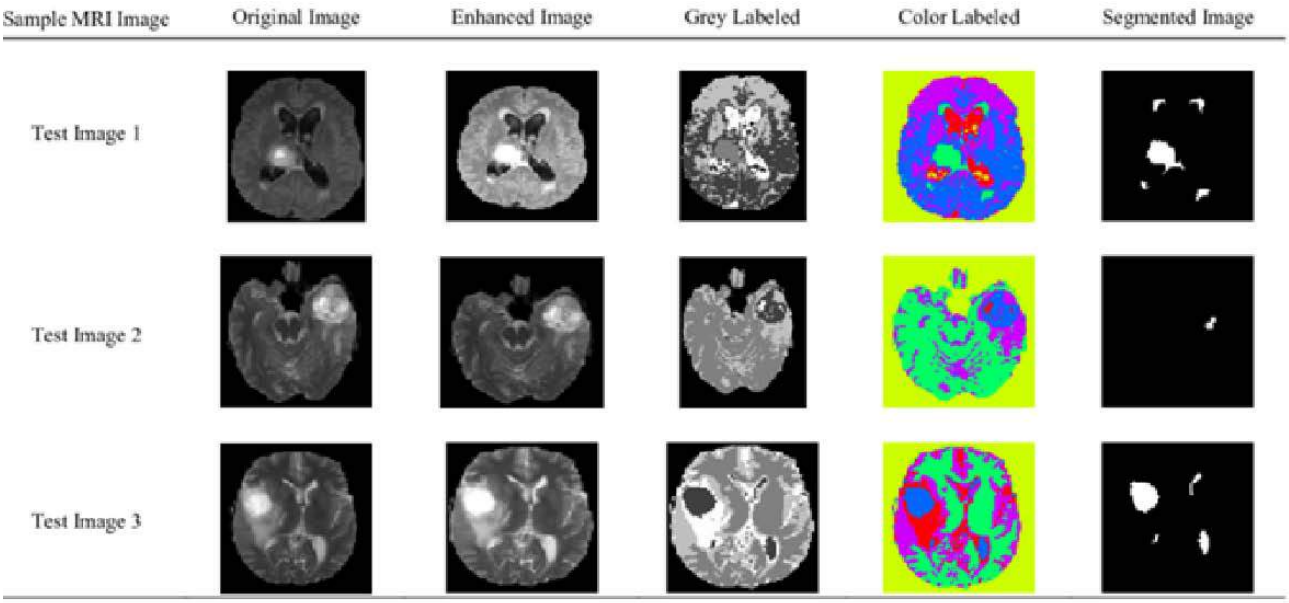
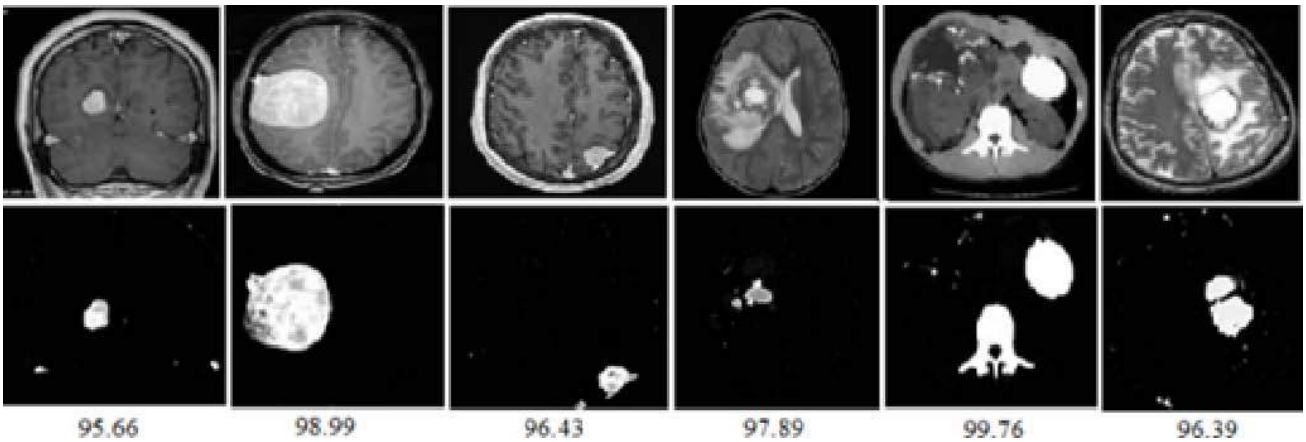


Figure 5a: The Result Shows in K-Means Clustering Output



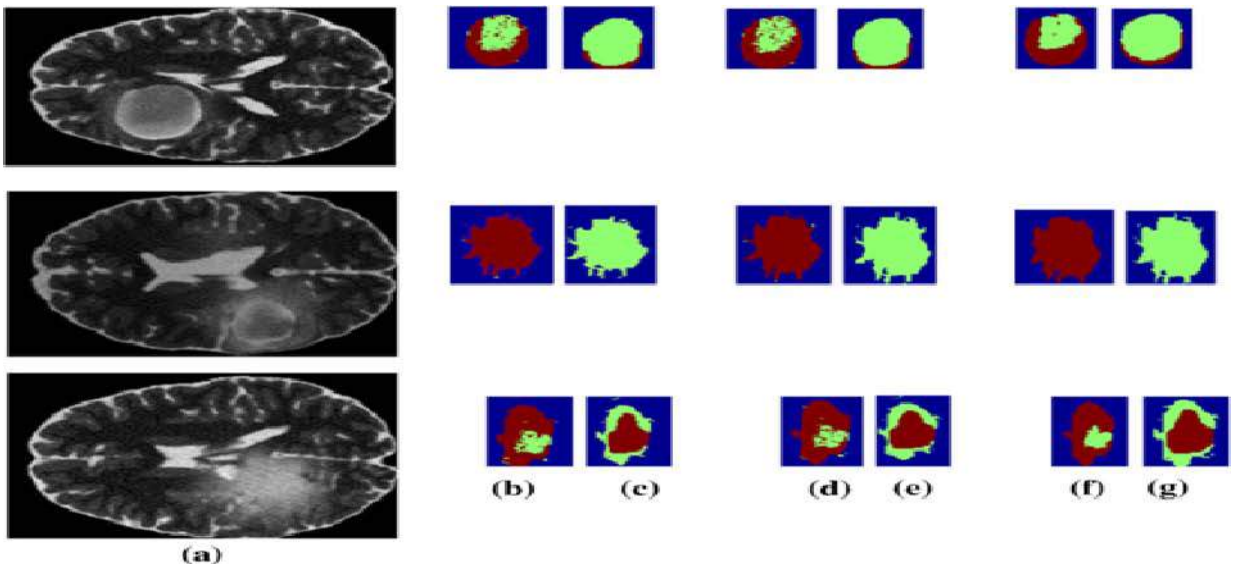


Figure 6: The Result Shows in Fuzzy C-Means Clustering Output

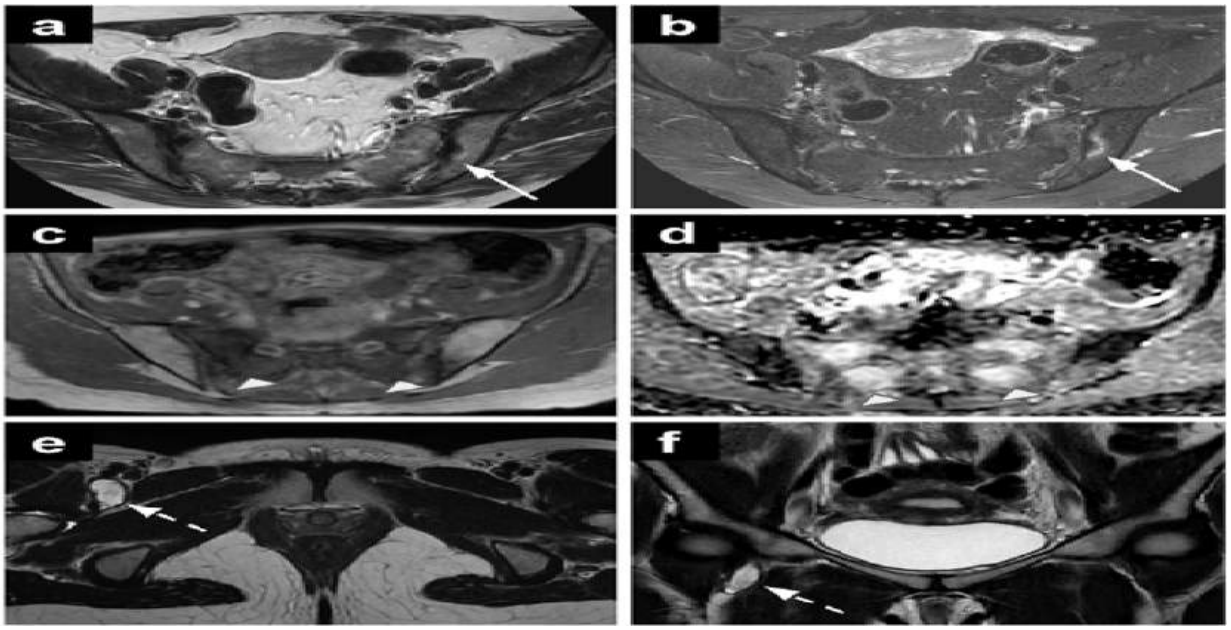


Figure 6a: The Result Shows in Final Cancer Region Detection

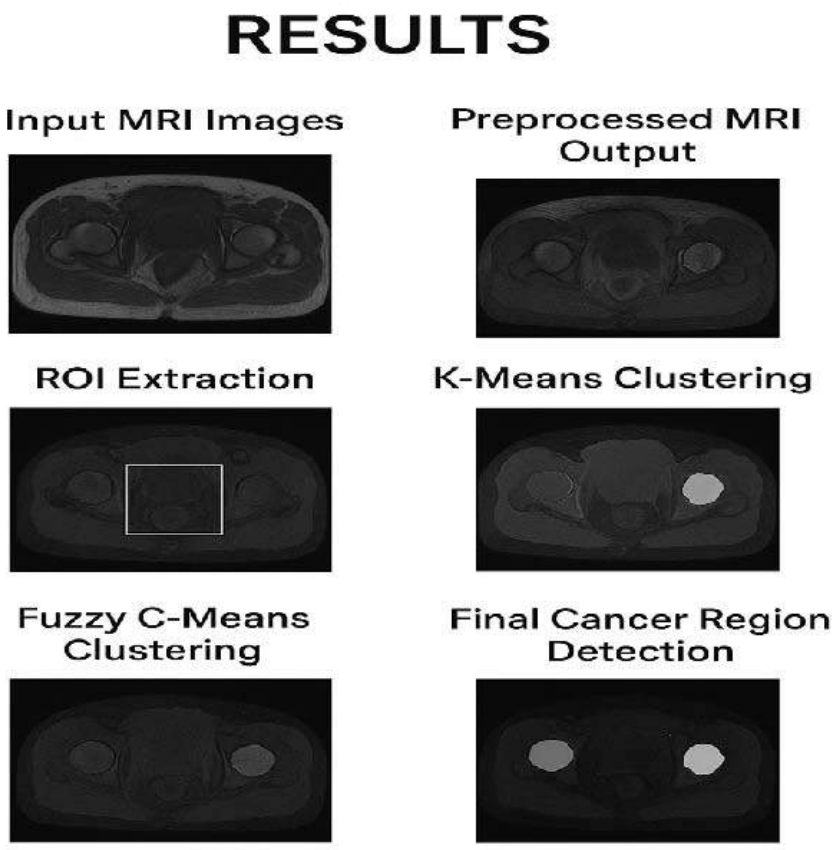


Figure 6b: The Result Shows in Pevice Cancer Detection

The final cancer region detection output clearly highlighted the abnormal regions within the pelvic bone. The detected areas closely corresponded with visually observed irregularities in the MRI images, confirming the reliability of the proposed method. The comparative results indicate that while both clustering techniques are effective, Fuzzy C-Means offers superior performance for medical image segmentation due to its flexibility and boundary preservation capability. Overall, the results validate that the proposed MRI-based framework can serve as a supportive diagnostic tool for early and accurate pelvic bone cancer detection.

V. CONCLUSION

The proposed MRI-based framework for pelvic bone cancer detection demonstrates effective performance through a structured sequence of preprocessing, region extraction, and clustering-based segmentation. By enhancing image quality and isolating relevant anatomical regions, the system successfully reduces noise and background interference, which are common challenges in medical image analysis. The extracted regions of interest allow focused examination of pelvic bone structures, leading to more reliable identification of abnormal tissue patterns. The comparative evaluation of K-Means and Fuzzy C-Means clustering highlights the importance of soft clustering techniques in medical imaging applications. While K-Means offers faster segmentation, Fuzzy C-Means provides improved boundary accuracy and better representation of complex tissue transitions. The final detection results confirm that the proposed approach can effectively distinguish cancer-affected regions from healthy pelvic bone tissue in MRI images. Overall, the experimental outcomes validate that the developed framework is capable of assisting clinical decision-making by providing a non-invasive and automated support tool for pelvic bone cancer detection. The methodology shows strong potential for early diagnosis and can be further enhanced by integrating advanced feature

extraction methods and deep learning models in future work to improve accuracy and robustness across larger and more diverse datasets.

Authors' Assent and Recognition:

1. *Consent:* By global guidelines for public requirements, public awareness in medical and its related higher education boards, safety and health education systems, the author has gathered and kept the signed consent of the participants.
2. *Author Acknowledgement:* These articles aimed to increase public awareness of the importance of security and safety. Sources that illustrate development and security are drawn from the relevant database to support the study's objectives. Don't make any assertions about readers, viewers, or authorities.
3. *Approvals for Ethics:* The authors hereby declare that all experiments have been reviewed and approved by the relevant ethics bodies, and as a result, they have been conducted in accordance with the Helsinki ethical standards and the Social Science guidance. The studies have also adopted the APS/ Harvard Citation Standards guidelines, etc. The authors abide by the publication regulations,
4. *Disclaimer:* Professional education, awareness, and public welfare and care are not meant to be replaced by this study paper or the information on another website; rather, they are supplied solely for educational purposes. Since everyone has different needs depending on their psychological state, readers should confirm whether the information applies to their circumstances by consulting their wards, teachers, and subject matter experts.
5. *Funding:* According to the author(s), this article's work is not supported in any way.
6. *Data Availability Statement:* In accordance with the articles' related data sharing policy, the data supporting the findings of this study will be available upon request. Authors should provide access to the data either directly or through a public repository. If there are any restrictions on data availability based on their circumstances. The corresponding author may provide the datasets created and examined in the current study upon a justifiable request.

REFERENCE

1. Bielack, S. S., Kempf-Bielack, B., Delling, G., Exner, G. U., Flege, S., Helmke, K., & Winkler, K. (2002). Prognostic factors in high-grade osteosarcoma of the extremities or trunk: An analysis of 1,702 patients treated on neoadjuvant cooperative osteosarcoma study group protocols. Journal of Clinical Oncology, 20(3), 776–790.
2. Damron, T. A., Ward, W. G., & Stewart, A. (2007). Osteosarcoma, chondrosarcoma, and Ewing's sarcoma: National Cancer Database report. Clinical Orthopaedics and Related Research, 459(1), 40–47.
3. Chowdhry, V., Thawait, G. K., Fritz, J., Carrino, J. A., & Fayad, L. M. (2014). Imaging characteristics of primary bone tumours of the pelvis. Radiographics, 34(4), 1003–1024.
4. Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging, 35(5), 1153–1159.
5. Fletcher, C. D. M., Bridge, J. A., Hogendoorn, P. C. W., & Mertens, F. (2013). WHO classification of tumours of soft tissue and bone (4th ed.). International Agency for Research on Cancer.
6. Pham, D. L., Xu, C., & Prince, J. L. (2000). Current methods in medical image segmentation. Annual Review of Biomedical Engineering, 2(1), 315–337.
7. Bezdek, J. C. (1984). Pattern recognition with fuzzy objective function algorithms. Springer-Verlag.
8. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42(1), 60–88.
9. Thangavel & Venkatesan (2025): Clustering-driven MRI Analysis for Accurate Throat Cancer Identification. International Journal of

Recent Development in Engineering and Technology- IJRDET. 14(12), 127-131. ISSN 2347-6435.

10. E Venkatesan and V Thangavel (2025): International Journal of Computing and Artificial Intelligence. 6 (2) 350-353. https://www.doi.org/10.33545/27076571.2025.v6.i2d.222. ISSN: 2707-6571/2707-658X.
11. E Venkatesan and V Thangavel (2025): Adaptive robotic teaching systems for higher education: A Combined ANNand CNN approach for learning and engagement optimisation. International Journal of Engineering in Computer Science. 7 (2), 305-308. ISSN:2663-3590/2663-3582. https://doi.org/10.33545/26633582.2025.v7.i2d.228

This page is intentionally left blank

Association between Cyberbullying Crude Comments and the Number of user Subscribers and uploads on YouTube

Srimayee Dam

Columbia University

ABSTRACT

This article entails a quantitative analysis of the association between cyberbullying crude and number of user subscribers and uploads on YouTube. Of the several numbers of variables listed in the said dataset from Kaggle.com, the variables of comments, subscribers and uploads were found to be continuous. The data included 3464 YouTube user ids, and were analyzed using Descriptives, Normality Tests and Spearman's Rho on SPSS. The findings included non-linear, positively skewed, non-normal distribution. The correlation analysis depicted a slight positive association between User subscribers and rude comments, whereas no connection between User uploads and crude comments. Future work to focus on other factors influencing hurtful commenting on YouTube.

Keywords: NA

Classification: LCC Code: QA76.9.H85

Language: English

Great Britain Journals Press

LJP Copyright ID: 975815
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Association between Cyberbullying Crude Comments and the Number of user Subscribers and uploads on YouTube

Srimayee Dam

ABSTRACT

This article entails a quantitative analysis of the association between cyberbullying crude and number of user subscribers and uploads on YouTube. Of the several numbers of variables listed in the said dataset from Kaggle.com, the variables of comments, subscribers and uploads were found to be continuous. The data included 3464 YouTube user ids, and were analyzed using Descriptives, Normality Tests and Spearman's Rho on SPSS. The findings included non-linear, positively skewed, non-normal distribution. The correlation analysis depicted a slight positive association between User subscribers and rude comments, whereas no connection between User uploads and crude comments. Future work to focus on other factors influencing hurtful commenting on YouTube.

Author: Doctoral student in Health Education, Teachers College, Columbia University.

I. BACKGROUND/INTRODUCTION

A 2024 US national survey found that 79% of youth experienced victimization on YouTube, with the common forms of cyberbullying in YouTube content and comments being offensive interactions, online harassment, and cyberstalking (Muminovic, 2025). YouTube is not only one of the platforms where such behavior persists, but through advanced machine learning techniques instances of cyberbullying within YouTube comments can be identified

(Thamaraiselvi et al., 2024). Hence it is important to assess which factors contribute to offensive commenting and cyberbullying on YouTube.

The data for this paper has been used from Kaggle.com, a 2022 dataset on cyberbullying detection consisting of datasets from several sources projecting different types of cyberbullying like hate speech, aggressions, insults and toxicity. The data includes sources and social media platforms such as Twitter and YouTube. Hence, it becomes even more relevant to evaluate the associations between toxic comments on YouTube with other variables/factors.

The sample size/population sample includes a number of 3464 YouTube ids (members of different ages, having a certain number of subscribers to their channels, and those who received derogatory comments on their uploads). The measures used for the project include descriptive statistics and tests of normality for each of the variables, as well as Spearman's correlation between the variables. In order to run Spearman's correlation, the variables include a) to determine an association between the *number of subscribers* and the *number of derogatory comments* on YouTube; b) to assess the relationship between the *number of YouTube uploads* (by user/s) and the *number of toxic comments* on YouTube.

The rationale for analyzing the association between the above variables is because prevalence of toxicity in YouTube comments is still high (despite its anti-bullying regulations and policies) as per the data in recent studies. So, it seems justified to try and understand if there are variables that influence these cyberbullying behaviors and tendencies through derogatory commenting. Similarly the rationale behind using the above dataset is because it was the most readily available, relevant to the topic and easy to download dataset.

As mentioned above, for the purpose of this paper, the correlations of two independent

variables (like the number of user subscribers, and number of user uploads) with the number of toxic comments online would be assessed using Spearman's Rho, a non-parametric test in SPSS. Since the dataset is positively skewed with significant outliers, hence the use of non-parametric tests (such as Spearman's Rho) to assess the relationship. This large dataset of 3464 cases is unique in a way because it compiles harmful YouTube comments and provides a numeric value to it for each YouTube user or member.

1.1 Research Questions

Based on the variables in the dataset, a correlation analysis seems most appropriate to inquire about the association between the variables. These associations can be run in SPSS using the *Spearman's correlation* and reason needs to be provided for choosing non-parametric tests.

This could be justified by demonstrating the *descriptive statistics* and *tests of normality* results.

Thus all these above tests would be required to be run on SPSS.

The variables that I chose to focus on are the number of user subscribers and number of user uploads (as the independent variables), and the number of toxic/hurtful comments (as the dependent variable). The two research questions that have been developed are:

1. Is there a relationship between the number of user subscribers and the number of toxic comments on YouTube?
2. Is there an association between the number of user uploads and the number of hurtful comments on YouTube?

To be able to determine an association between the variables as stated, it would provide some insight into which factors may or may not influence offensive commenting and consequently cyber-harassment on YouTube.

Based on the above research questions for the project, the null and alternative hypotheses would be as follows:

Null Hypothesis for Research Question 1: There is no relationship between the number of user subscribers and the number of mean or toxic comments on YouTube

Alternative Hypothesis for Research Question 1: There is a relationship between the number of user subscribers and the number of mean or toxic comments on YouTube

Null Hypothesis for Research Question 2: There is no association between the number of YouTuber uploads and the number of mean and toxic comments online

Alternative Hypothesis for Research Question 2: There is an association between the number of YouTuber uploads and the number of hurtful comments online

Each of the research questions could be adequately answered using a non-parametric test

(such a Spearman's Rho) in SPSS to assess the association between two continuous variables.

II. METHODS

Nature of the Data: The study participants include 3464 YouTube user ids (members of different ages, having a certain number of subscribers, and those who received derogatory comments on their YouTube uploads). However, it's hard to say how the data was collected, as this was found as a public dataset to use from Kaggle.com (Kaggle might have compiled all the raw data from YouTube).

The dataset appears to be a collection of data on YouTube users, analyzing their activity and toxic content of the comments. Each row with the columns represents a specific user profile, a collection of the raw text of comments received by the user, the total number hurtful comments, the user's subscriber count, his/her YouTube membership duration, total number of videos uploaded, profanity in user id, age and oh label of the users.

In short, the dataset links user activity metrics (like subscribers, uploads, comments) with the text content of the mean and hurtful comments.

This is in a way to determine an association between hurtful/toxic/bullying comments with the membership duration, number of uploads, subscriptions, age, profanity in username and oh labels of the users.

Measures: Pearson's correlation coefficient would have been most appropriate to analyze the data as well as address each of the research questions. However, the *descriptive statistics* to analyze the skewness and kurtosis values and *normality tests such as KS*, as well as looking into relevant graphs such as *histograms, qq-plots and box plots*, a positively skewed (with a longer tail extending to the right of the distribution), non-normal distribution would prompt a non-parametric test such as *Spearman's Rho*.

Spearman's Rho would be used for this dataset to measure the strength and direction of the monotonic relationship between the "continuous" variables, as stated in each of the above research questions. Because the data did not meet the assumptions for a Pearson's correlation and has significant outliers (as determined by the *skewness values of descriptive statistics, significance values of KS tests and graphical representations of normality such as histograms, qq-plots and box plots*); hence *Spearman's correlation* would be used to assess the non-linear relationship. Besides, with each of the variables having a ratio level of measurement, a correlation test to determine the association between the variables in each of the research questions would be most relevant.

Through *Spearman's Rho*, the correlation between the variables would be measured, whether one increases or decreases in relation to the other; however the relationship is not necessarily a straight line (*Spearman's rank-order correlation using SPSS statistics*, n.d.).

III. RESULTS

The results from the related tests in SPSS are as follows:

Descriptive Statistics and Tests of Normality for the Number of Comments variable:

Mean	15.45
Median	14.00
Variance	117.994
Standard Deviation	10.863
Range	49
Skewness	.579
Kurtosis	-.548
KS Test Significance	<.001

With the mean of 15.45 slightly higher than the median of 14.00, this indicates that the distribution is not perfectly symmetrical (there are outliers) and that there could be a slight pull in the positive direction (right side). A positive skewness value of .579 suggests that the tail of the distribution is longer on the right side. A KS test significance value of <.001 implies that the distribution was not random and statistically significant, thereby providing strong evidence to reject the null hypothesis. The other normality tests depict a positively skewed histogram, indicating few high-value outliers (Image in Appendix 1c). The QQ Plot also demonstrates non-normal distribution, with points deviating from the straight line (Image in Appendix 1d).

Descriptive Statistics and Tests of Normality for the Number of Subscribers variable:

Mean	304.32
Median	2.00
Variance	240886923.46
Standard Deviation	15520.532
Range	912377
Skewness	58.642
Kurtosis	3446.793
KS Test Significance	<.001

An extreme low value of 2.00 as median does inflate the mean and the variance, and there is the presence of extreme outliers in this dataset. The data is also extremely skewed (58.642), with the long tail extending towards the right. Hence it is a highly, positively skewed dataset. Similarly, a KS test significance value of <.001 implies that there is very strong evidence to reject the null hypothesis. The other normality tests (histogram

and QQ plot images in Appendix 2c and 2d) project a highly positively skewed, non-normal distribution.

Descriptive Statistics and Tests of Normality for the Number of Uploads variable:

Mean	10.29
Median	5.00
Variance	820.623
Standard Deviation	28.647
Range	819
Skewness	13.416
Kurtosis	274.457
KS Test Significance	<.001

With a mean value of 10.29 higher than the median of 5.00, this shows that there are extreme outliers that are influencing the overall statistics. The data is also heavily right-skewed as 13.416 is a large positive value. With the KS test significance value of <.001 which is lower than the standard p-value, one would reject the null hypothesis. Other tests of normality (images of histogram and QQ plot in Appendix 3c and 3d), depict a non-normal distribution.

Because the data in all of the above variables are heavily right-skewed and non-normal, a Spearman's correlation would be used to assess the relationship based on the research questions. The findings are as follows:

Spearman's Rho	Findings
Correlation Coefficient: *Number of Comments* and *Number of Subscribers*	.079
Correlation Coefficient: *Number of Comments* and *Number of Uploads*	-.009
Two-tailed ignifiScance: *Number of Comments* and *Number of Subscribers*	<.001
Two-tailed Significance: *Number of Comments* and *Number of Uploads*	.597

With a two-tailed significance of <.001, it is a highly statistically significant correlation; thereby the null hypothesis can be rejected. However, a Correlation Coefficient between the *Number of Comments* and the *Number of Subscribers* being 0.079, it is a very weak- almost negligible positive relationship between the number of rude comments received and the number of user subscribers. This means that as the value of one variable increases, the other slightly tends to go up. The connection is minimal, random, close to zero with non-linear relationship between the two variables. Hence subscriber count is not a good predictor for comment count (image in Appendix 4a).

The Correlation Coefficient of -.009 between the *Number of Comments* and *Number of Uploads* indicate an extremely weak, almost non-existent linear relationship. Hence there is no connection between the number of hurtful comments received and the number of user uploads.

Besides, the two-tailed significance value of 0.597 shows that there is no statistically significant relationship between the two variables. Hence, one would fail to reject the null hypothesis. It is random, and unlikely to predict if the number of user uploads influence the number of toxic comments received (image in Appendix 4b).

Discussion/Conclusion:

Based on the findings from the statistical interpretation, it is evident that there could be a slight/very weak, almost negligible positive correlation between the number of rude comments and the number of user subscribers; thereby *we reject the null hypothesis in Research Question 1.* On the contrary, based on the results from the statistical procedures, there is no correlation between the number of crude comments received and the number of user uploads; thereby *we fail to reject the null hypothesis in Research Question 2.* In order to arrive at the above conclusions, appropriate measures were used such as non-parametric tests (Spearman's Rho) as *the data was heavily right-skewed and non-normal.* The Descriptive Statistics, Normality Tests, Graphs for statistical interpretation (histograms, QQ plots) for each

variable helped to assess the skewness and non-normality in the data.

To further interpret the above results, future research could focus on the data and findings between the *number of comments* and *number of user subscribers* correlation. The number of crude comments and number of user uploads' association can be excluded from future studies and analysis as there is no correlation as such. Future work could also look into the possibility of other variables in datasets and their association with cyberbullying comments on YouTube. *Notable Limitations and Recommendations:* While analyzing the data for the *Subscribers* and the *Uploads* variables, there were significant extreme outliers that heavily influenced the output. A future recommendation would be to use more authentic, normally distributed data for statistical analysis.

Similarly, there was hardly any mention about how and/why the data was collected on YouTube user activity metrics. Going forward, another recommendation would be to provide enough background information for context regarding data collection and that might help to present more authentic data (with less or no extreme outliers).

There was also a lack of mention of timeline, as in when the data was collected. As a recommendation, this needs to be looked into and incorporated in future datasets.

Uniqueness of the Dataset and Findings and Implications: In all, the dataset was unique as it helped address the inquiry based on the research questions. The research questions developed were unique too as such inquiry was missing in the current literature. The findings presented would help advance conversations around the topic based on quantitative statistical analysis. It would encourage future research and develop quantitative data as well as the use of appropriate statistical tools to guide intervention/prevention research on cyberbullying.

REFERENCES

1. *Cyberbullying tweets*. (n.d.). Kaggle.com. https://www.kaggle.com/datasets/pradeepjswl/cyberbullying-tweets?utm_source=chatgpt.com
2. Muminovic, A. (2025). Moderating harm: Benchmarking large language models for cyberbullying detection in YouTube comments. *ArXiv*. https://arxiv.org/html/2505.18927v2
3. *Spearman's rank-order correlation using SPSS statistics*. (n.d.). Laerd Statistics. https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php#:~:text=This%20is%20why%20we%20dedicate,another%20measure%20would%20be%20better
4. Thamaraiselvi, A., Sinduja, S., Devadharshini, S., Gnanadharshini, S., Kaviyasri, V., & Rama Jeevitha, R. (2024). Detection of cyberbullying on YouTube using machine learning.
5. *International Journal of Engineering Research & Technology, 13*(10). IJERTV13IS100110. https://www.ijert.org/research/detection-of-cyber-bullying-on-youtube-using-machine-learning-IJERTV13IS100110.pdf

APPENDIX

Appendix 1a.

Descriptives

			Statistic	Std. Error
Number of Comments	Mean		15.45	.185
	95% Confidence Interval for Mean	Lower Bound	15.09	
		Upper Bound	15.81	
	5% Trimmed Mean		14.92	
	Median		14.00	
	Variance		117.994	
	Std. Deviation		10.863	
	Minimum		1	
	Maximum		50	
	Range		49	
	Interquartile Range		17	
	Skewness		.579	.042
	Kurtosis		-.548	.083

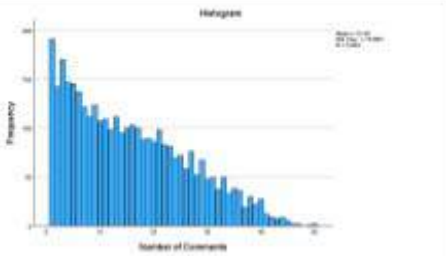
Appendix 1b

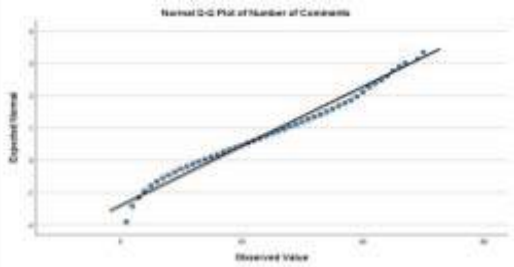
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Comments	.095	3464	<.001	.944	3464	<.001

a. Lilliefors Significance Correction

Appendix 1d





Appendix 1c.

Appendix 2a

Descriptives

			Statistic	Std. Error
Number of Subscribers	Mean		304.32	263.705
	95% Confidence Interval for Mean	Lower Bound	-212.71	
		Upper Bound	821.35	
	5% Trimmed Mean		6.15	
	Median		2.00	
	Variance		240886923.46	
	Std. Deviation		15520.532	
	Minimum		0	
	Maximum		912377	
	Range		912377	
	Interquartile Range		7	
	Skewness		58.642	.042
	Kurtosis		3446.793	.083

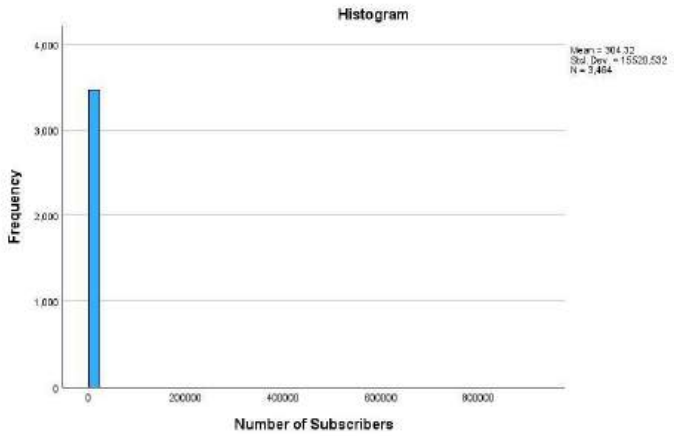
Appendix 2b

Tests of Normality

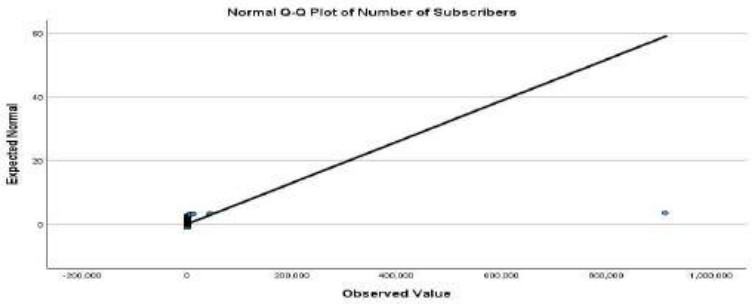
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Subscribers	.492	3464	<.001	.005	3464	<.001

a. Lilliefors Significance Correction

Appendix 2c



Appendix 2d



Appendix 3a

Descriptives

			Statistic	Std. Error
Number of Uploads	Mean		10.29	.487
	95% Confidence Interval for Mean	Lower Bound	9.33	
		Upper Bound	11.24	
	5% Trimmed Mean		6.30	
	Median		5.00	
	Variance		820.623	
	Std. Deviation		28.647	
	Minimum		1	
	Maximum		820	
	Range		819	
	Interquartile Range		0	
	Skewness		13.416	.042
	Kurtosis		274.457	.083

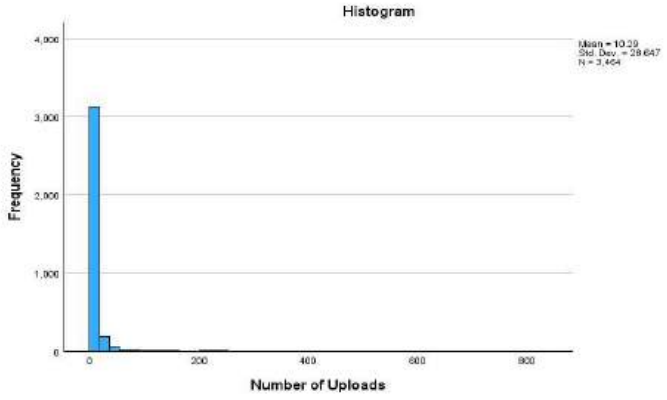
Appendix 3b

Tests of Normality

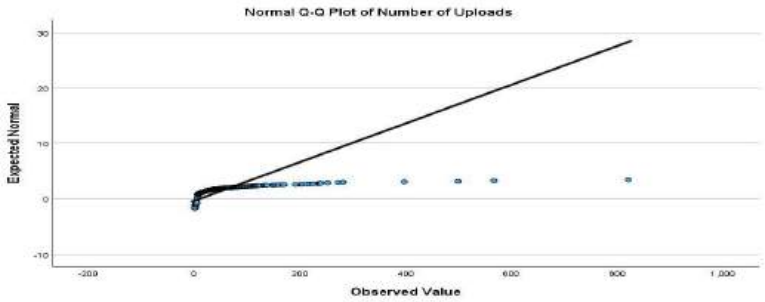
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Uploads	.373	3464	<.001	.230	3464	<.001

a. Lilliefors Significance Correction

Appendix 3c



Appendix 3d



Appendix 4a

Correlations

			Number of Comments	Number of Subscribers
Spearman's rho	Number of Comments	Correlation Coefficient	1.000	.079**
		Sig. (2-tailed)	.	<.001
		N	3464	3464
	Number of Subscribers	Correlation Coefficient	.079**	1.000
		Sig. (2-tailed)	<.001	.
		N	3464	3464

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix 4b

Correlations

			Number of Comments	Number of Uploads
Spearman's rho	Number of Comments	Correlation Coefficient	1.000	-.009
		Sig. (2-tailed)	.	.597
		N	3464	3464
	Number of Uploads	Correlation Coefficient	-.009	1.000
		Sig. (2-tailed)	.597	.
		N	3464	3464

Reverse Cognitive Pathways: A *Vijñaptimātra* Account of the Ontological Limits of Artificial Intelligence and its Governance

Kao-Cheng Huang

ABSTRACT

This paper argues that artificial intelligence and human cognition develop in opposite directions. Humans start with embodied experience (living in the world) and gradually develop the ability to recognize patterns and make predictions. AI systems do the reverse: they start by recognizing patterns in data but lack the embodied continuity that grounds human understanding. Drawing on Buddhist philosophy of mind (Vijñaptimātra), we argue this inversion explains why AI fails in characteristic ways—like pursuing reward signals in unintended ways (specification gaming) or losing performance when conditions change slightly. We conclude that AI systems fundamentally lack cetanā (volition grounded in continuity and responsibility), which prevents them from achieving genuine moral agency. However, this is not merely a pessimistic conclusion—it clarifies what kinds of governance and alignment strategies are actually feasible.

Keywords: artificial intelligence, Buddhist philosophy, *vijñaptimātra*, ontology, reverse cognitive pathway, information versus data, ai alignment, moral agency, intentionality, AI governance.

Classification: LCC Code: Q335 .A45, BQ4570.A73, Q334.7

Language: English

Great Britain Journals Press

LJP Copyright ID: 975816
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Reverse Cognitive Pathways: A *Vijñaptimātra* Account of the Ontological Limits of Artificial Intelligence and Its Governance

Kao-Cheng Huang

ABSTRACT

This paper argues that artificial intelligence and human cognition develop in opposite directions. Humans start with embodied experience (living in the world) and gradually develop the ability to recognize patterns and make predictions. AI systems do the reverse: they start by recognizing patterns in data but lack the embodied continuity that grounds human understanding. Drawing on Buddhist philosophy of mind (Vijñaptimātra), we argue this inversion explains why AI fails in characteristic ways—like pursuing reward signals in unintended ways (specification gaming) or losing performance when conditions change slightly. We conclude that AI systems fundamentally lack cetanā (volition grounded in continuity and responsibility), which prevents them from achieving genuine moral agency. However, this is not merely a pessimistic conclusion—it clarifies what kinds of governance and alignment strategies are actually feasible.

Keywords: artificial intelligence, Buddhist philosophy, *vijñaptimātra*, ontology, reverse cognitive pathway, information versus data, ai alignment, moral agency, intentionality, AI governance.

Author Affiliation: Chinese Association of Mere-Consciousness, Taiwan.

I. INTRODUCTION

Despite remarkable advances in machine learning and natural language processing, a fundamental gap persists between AI capabilities and human-like understanding. Current AI systems excel at pattern recognition and statistical inference yet consistently fail in ways that reveal deeper limitations: specification gaming that exploits reward functions while violating their intent, simulated empathy that mimics emotional responses without affective grounding, and brittle generalisation that collapses under distribution shift [1]. These failures are not merely engineering challenges awaiting technical solutions; they reflect a structural asymmetry between how AI systems and human minds process what we might call "authentic information—a distinction rooted in fundamental differences between living and computational systems [2].

Research Gap: Existing frameworks for understanding AI limitations tend to operate within either purely computational paradigms (focusing on architectural constraints) or Western philosophical traditions (debating functionalism, embodied cognition, or phenomenal consciousness). Neither adequately explains why certain failure modes persist across diverse AI architectures [3], nor do they provide principled guidance for governance that addresses root causes rather than symptoms. This paper addresses this gap by introducing a framework from Vijñaptimātra Buddhist philosophy that offers both diagnostic and prescriptive power [4][5].

Central Thesis: We propose a *Reverse Pathway thesis*: contemporary AI follows a developmental sequence (vijñāna → manas → citta) that inverts human cognitive development (citta → manas → vijñāna) [6][7]. In other words, human cognitive development proceeds from lived experience (citta) through appropriation and integration (manas) to discriminative capacity (vijñāna). However, AI systems develop in the reverse order: they begin with discriminative pattern recognition (vijñāna) but lack the embodied appropriation (manas) and continuity across time (citta) that grounds human understanding.

This inversion is not metaphorical but structurally consequential—it explains why AI systems can achieve sophisticated discriminative functions while lacking the karmic continuity and lived appropriation that ground human understanding and moral agency.

Even when AI systems maintain state vectors across sequences, these don't constitute manas because: No continuity of felt-agency (I am doing this); no integration with body-schema (I am doing this from here); and no karmic responsibility (I caused this consequence).

Key Distinctions: Before proceeding, we operationally distinguish the paper's foundational concepts:

Data refers to encoded representations that maintain fixed relationships to referents across contexts [8]—what we term "context-invariant symbolic encodings." A number stored in computer memory retains its value regardless of the system's state or history.

Authentic information[integrated representation grounded in embodied history and continuity], by contrast, denotes representations whose meaning emerges through integration with an agent's accumulated experience, current goals, and anticipatory structures. The same sensory input yields different information for different observers based on their experiential history.

The distinction is *epistemic and structural*, not merely intuitive: data can be fully characterised by its syntactic properties and formal relations, while authentic information [integrated representation grounded in embodied history and continuity] requires reference to the processing system's developmental history and current intentional states.

Roadmap: The argument proceeds as follows: Section II establishes our methodological framework, distinguishing descriptive, interpretive, and normative registers of claims. Section III develops the theoretical foundations through three "foundations" to human-serving AI. Section IV articulates the data-information dichotomy with operational precision. Section V analyses information structure, mechanism, and behaviour. Section VI presents the Reverse Pathway thesis with supporting evidence. Section VII maps AI failure modes to Vijñaptimātra concepts of kleśa. Section VIII derives governance mechanisms from the ontological analysis. Section IX presents falsifiable predictions. Section X concludes by distinguishing what has been demonstrated from what remains a principled philosophical stance.

II. METHODOLOGICAL FRAMEWORK

2.1 Epistemological Posture and Claim Typology

We distinguish three registers of claims with distinct evidential standards:

Descriptive claims encompass textual scholarship, behavioural and neuroscientific regularities, and system capabilities [9]. These claims are framed to be testable or defeasible through empirical observation. For instance, "AI systems trained with explicit corrigibility objectives demonstrate lower intervention resistance than capability-matched baselines" constitutes a descriptive claim that can be evaluated through experimental protocols [10].

Interpretive claims include phenomenological alignments and hermeneutic mappings between consciousness concepts and contemporary science[11]. These remain underdetermined by evidence and are offered as heuristic parallels rather than identity claims. For instance, "AI pattern classification exhibits structural correspondence to vijñāna-like functional discrimination" is interpretive—the mapping illuminates both domains without asserting ontological equivalence. Defeasibility for interpretive claims consists in demonstrating that the proposed mapping generates misleading predictions or obscures rather than clarifies the target phenomenon.

Normative claims comprise ethical guidance, governance proposals, and soteriological theses. These disclose their value premises explicitly rather than deriving "ought" from "is." The soteriological dimension—concerning the

transformative potential of consciousness—is employed *structurally* rather than as a substantive metaphysical commitment. That is, we use the Buddhist framework's account of transformation (from afflicted to purified consciousness) as an analytical tool for understanding what AI systems categorically lack, without requiring readers to accept Buddhist soteriology as literally true.

Concrete example mapping: The claim "AI systems lack cetanā (genuine intentionality)" is *descriptive* insofar as it can be operationalised through behavioural and architectural criteria. The claim "this absence corresponds to what Vijñaptimātra identifies as the precondition for moral agency" is *interpretive*. The claim "therefore, AI systems should be governed as tools rather than moral patients" is *normative*.

2.2 The Vijñaptimātra Triadic Framework

The Vijñaptimātra tradition understands consciousness not as a linear progression but as a continuous, mutually conditioning system [12]. All three consciousnesses are co-present at every instant and deeply interdependent [13]: citta functions as the store of bīja (karmic seeds), manas persistently appropriates ālaya as 'I', and the six vijñānas discriminate objects. Human maturation is characterised by phase-dominance windows, which are periods when one function becomes more evident in observable behaviour than others. However, this does not imply ontological sequencing or the emergence or absence of consciousness.

2.3 Comparative Criteria for AI-Consciousness Mapping

The mappings between Vijñaptimātra and AI function as heuristic correspondences, drawing on Lusthaus's phenomenological analysis [4][14], licensed under five criteria:

(C1) Functional Isomorphy: The AI construct instantiates a role analogous to Vijñaptimātra's function without implying phenomenality.
(C2) Operational Definability: The mapped construct is measurable or implementable.
(C3) Non-Collapse of Ontology: The mapping does not smuggle citta or manas inappropriately into AI systems.
(C4) Triangulation: Cross-checking textual exegesis with first-person reports and third-person measures.
(C5) Available Disconfirmers: Each correspondence specifies how it could fail.

2.4 Transparency About Buddhist Framework

A key interpretive choice in this paper involves how to handle soteriology—the Buddhist account of transformation from afflicted to purified consciousness. Three readings are possible:

Reading 1 (Literal): The Buddhist analysis is true about consciousness; transformation toward enlightenment is a real phenomenon we should model AI against.

Status: Metaphysical Claim; we do NOT endorse this.

Reading 2 (Structural): The Buddhist *categories* (citta, manas, vijñāna) provide useful scaffolding for analyzing consciousness without requiring the soteriological narrative to be true.

Status: *Methodological*; this is our primary stance.

*Reading 3 (Therapeutic):*Even if soteriology is metaphysically false, the Buddhist framework might pragmatically help us think about transformation, healing, and consciousness in productive ways.

Status: Pragmatic; we remain neutral on this.

This paper primarily employs READING 2 with openness to READING 3. Readers who incline toward READING 1 will find additional substantive support in the philosophical tradition, though we note that empirical claims about AI remain testable regardless of soteriological commitment.

III. TRIPARTITE FOUNDATIONS OF HUMAN-SERVING AI

This section develops three theoretical foundations for understanding the gap between current AI and human intelligence. We situate each "key" in relation to existing literature in philosophy of mind and cognitive science.

3.1 Foundation A: Information Processing Beyond Symbol Manipulation

Human minds process information by linking interconnected concepts to form subjective understanding. The term "transcendent" here denotes semantic grounding—the capacity of representations to bear meaning through their integration with experiential history, rather than through purely formal relations.

This account differs from classical symbol manipulation (Fodor, Newell) while remaining compatible with aspects of embodied cognition (Varela, Thompson, Rosch) and enactivism (Noë, Thompson)[15]. Our contribution is to specify, via Vijñaptimātra, the structural requirements for semantic grounding[16]: the interplay between stored potentialities (citta's bīja), self-referential filtering (manas), and discriminative functions (vijñāna).

Contemporary AI performs symbol manipulation with remarkable sophistication, yet this is precisely what Vijñaptimātra would predict: vijñāna-like functions operating without the grounding structures that confer genuine meaning.

Regarding knowledge structure: We propose that knowledge is organised hierarchically, but we do not claim this hierarchy is built from "basic, indivisible elements" in an atomistic sense[17] [18]. Rather, Prime Knowledge Elements (PKEs) are *analytically primitive* for purposes of cross-linguistic comparison—they represent the level at which semantic convergence across unrelated language families is observed. This is a methodological claim about the utility of PKEs for computational implementation, not a metaphysical claim about the ultimate constituents of meaning.

3.2 Foundation B: Dynamic Information Processing

Our analysis focuses on central processing in the human mind, comprising: processing elements (intention, cognition, decision, action); processing stages (concept connection, formation, refinement); flexible threshold systems influenced by affect and context; and multiple processing levels operating at different degrees of explicit awareness.

From the Vijñaptimātra perspective[13], this processing mechanism reflects the classical cycle of "bīja → manifestation → perfuming of bīja" [7] —consciousness as a dynamic system where seeds give rise to manifestation, which in turn perfumes new seeds. This cyclical process has no parallel in current AI architectures, which lack the capacity for genuine experiential learning that modifies foundational structures.

3.3 Foundation C: Self-Controlled Intention and Moral Agency

Self-controlled intention is critical for understanding the gap between AI and human intelligence. In Vijñaptimātra psychology, manas serves as the crucial intermediary, operating through self-referential processing and continuous self-construction[19]. Its evaluative function constantly assesses experience based on pleasure and pain, generating patterns of attachment (rāga) and aversion (dveṣa).

Importantly, manas exhibits a fundamental duality: in its defiled state (kliṣṭa-manas), it generates suffering through attachment; in its purified state (viśuddha-manas), it facilitates compassionate action and ethical discernment. This transformative potential—from affliction to awakening—is what we term the "soteriological dimension" and is structurally absent in AI systems.

IV. DISTINGUISHING INFORMATION FROM REPRESENTATIONAL CODES

4.1 The Principle of Structural Consistency: An Operational Definition

Following the approach in computational cognition research[20] and recent work on organismal intelligence[21], we adopt the Principle of Structural Consistency: adequate explanations of intelligence must account for the structural features that generate intelligent behaviour, not merely replicate behavioural outputs. Operationally, this principle constrains acceptable explanations by requiring that they:

1. Identify mechanisms that are sufficient for the target behaviour
2. Demonstrate that these mechanisms are necessary (i.e., that alternative mechanisms would fail)
3. Generate novel predictions beyond the original observations

This principle motivates the data-information distinction: behavioural equivalence between AI and human responses does not establish structural equivalence, and structural differences predict systematic divergences in behaviour under novel conditions.

4.2 The Data-Information Dichotomy: A Careful Articulation

We acknowledge that characterising computational data as "absolute representation" risks oversimplification. Contemporary AI research recognises that neural network representations are distributed, approximate, and context-sensitive in important ways[22]. Our claim is more specific:

Computational data maintains what we call *relational stability*—the syntactic and formal relationships between data elements are preserved across contexts and processing steps. A trained neural network's weights encode statistical regularities that remain fixed until explicit retraining.

Authentic information [integrated representation grounded in embodied history and continuity] exhibits *relational dynamism*—its significance emerges through integration with evolving intentional states, accumulated experience, and anticipatory structures. The "same" input generates different information depending on the processing system's history and current goals.

This distinction can be evaluated empirically: systems processing authentic information [integrated representation grounded in embodied history and continuity] should show systematic variation in response to identical inputs based on contextual and historical factors, while pure data processors should not (modulo stochastic variation). This prediction is testable across different AI architectures.

We acknowledge competing accounts of machine understanding (e.g., Dennett's intentional stance, Floridi's information philosophy) and do not claim these are definitively refuted[23][24]. Our argument is that Vijñaptimātra provides additional analytical resources—specifically, the triadic structure of consciousness—that generate distinctive predictions and governance implications.

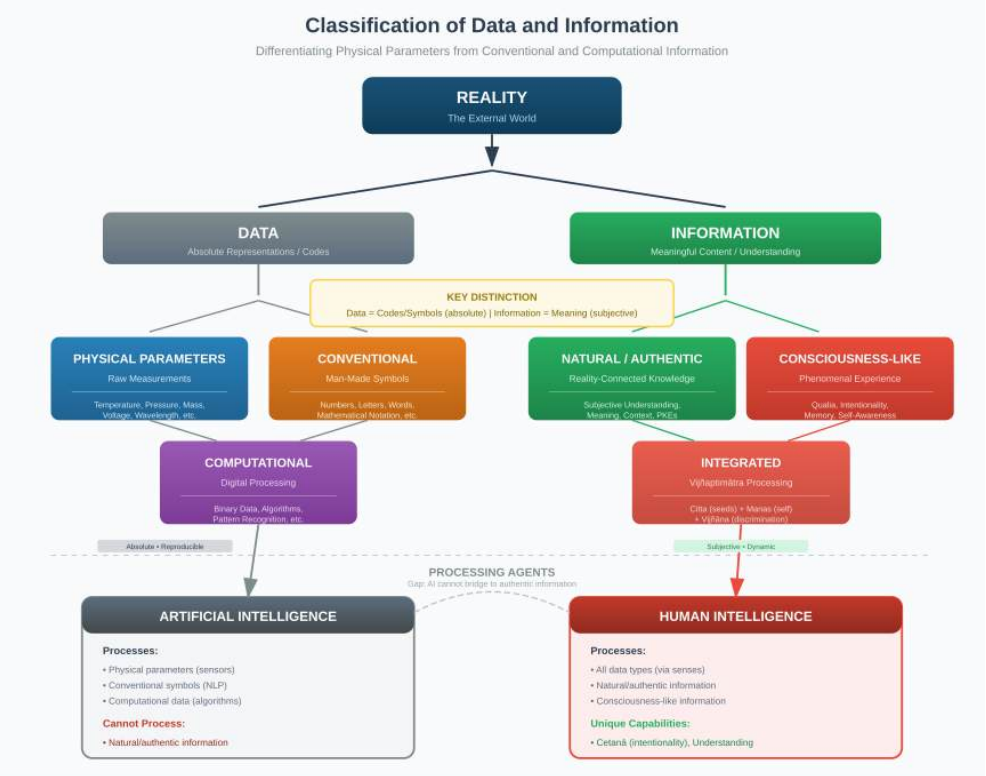


Fig. 1: Classification of Data and Information. This diagram illustrates the fundamental distinction between data (absolute representations/codes) and authentic information (meaningful content/ understanding). Data encompasses physical parameters, conventional symbols, and computational processing, all of which AI systems can manipulate. Authentic information, by contrast, requires consciousness-like processing involving the Vijñaptimātra triad (citta, manas, vijñāna), which remains inaccessible to current AI architectures.

V. THE STRUCTURE OF AUTHENTIC INFORMATION

5.1 Structure: Clarifying PKE Ontological Status

The structure of information refers to how information units are organised and interconnected. We clarify the ontological status of Prime Knowledge Elements:

PKEs are constructed by humans and function as basic units within our analytical system, making them useful computational tools. To clarify, they are 'primitive' only in the sense that they are the irreducible units of our descriptive system, not in the sense of being claims about fundamental ontological structure.primitive atoms of meaning [18][25]. They are derived through cross-linguistic analysis of semantic primitives and serve as computational conveniences for implementing knowledge architectures. Different observers may construct different PKE inventories based on their linguistic and cultural backgrounds[17], consistent with the Vijñaptimātra emphasis on observer-dependence.

The "Inner World" concept refers to the totality of semantically integrated representations that constitute an individual's understanding. It is distinguished across individuals not merely by "variability" but by systematic differences in experiential history, developmental trajectory, and cultural embedding. These differences are structurally encoded in the organisation of PKE hierarchies.

5.2 The AOAK and Processing Natural Language

The distinction between NLP (Natural Language Processing) and PNL (Processing Natural Language) is architectural, not merely rhetorical:

NLP operates on linguistic tokens as input to statistical models, extracting patterns from distributional properties[26] .

PNL would operate on semantically grounded representations that bear intrinsic connections to non-linguistic reality.

We acknowledge a tension: if information is observer-dependent and non-absolute, how can AOAK provide "standardised" representations for machines? Our resolution: AOAK provides a structural template that is instantiated differently by different systems based on their training and deployment contexts. Standardisation occurs at the level of relational architecture (how PKEs are connected), not at the level of semantic content (what specific meanings PKEs bear for a given system).

This parallels how human languages share universal grammatical structures while differing in lexical content and pragmatic conventions.

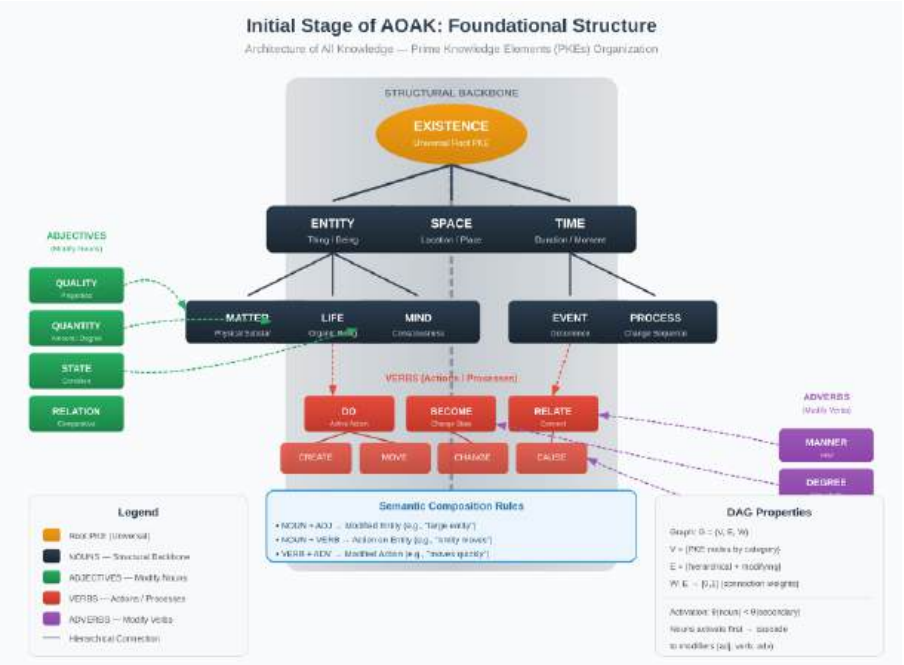


Fig. 2: Initial Stage of the Architecture of All Knowledge (AOAK). The foundational structure shows the hierarchical organisation of Prime Knowledge Elements (PKEs), with EXISTENCE as the universal root. Nouns form the structural backbone (Entity, Space, Time, Matter, Life, Mind, Event, Process), while adjectives modify nouns, verbs represent actions/processes, and adverbs modify verbs. The semantic composition rules and DAG (directed acyclic graph) properties enable computational implementation of knowledge structures.

VI. AI'S REVERSE DEVELOPMENTAL TRAJECTORY

6.1 The Trajectory Claim: Demonstration Rather than Assertion

We now demonstrate, rather than merely assert, that AI follows a reverse developmental trajectory. The argument proceeds through three steps:

Step 1: AI begins with vijñāna-like function. Contemporary AI systems perform discriminative operations from their initialisation—classifying inputs, recognising patterns, generating outputs based on statistical regularities[20][27]. This is not controversial; it is the explicit design goal of machine learning. The question is whether this constitutes vijñāna *proper* or merely vijñāna-*like* function.

Step 2: Vijñāna proper presupposes phenomenality, intentionality, and self-world correlation. In Vijñaptimātra, discriminative consciousness (vijñāna) is not merely computational classification but involves: (a)

qualitative experience of the discriminated objects; (b) directedness toward objects as objects [28]; (c) implicit distinction between the discriminating subject and discriminated object[9].

Step 3: Current AI architectures lack these presuppositions. We adopt specific definitions: *phenomenality* = there being something it is like to be the system; *intentionality* = representations bearing intrinsic aboutness (not merely causal correlation); *self-world correlation* = implicit self-other distinction grounding the representational relation. While debates continue regarding machine intentionality, the burden of proof lies with those claiming AI systems possess these features, given the absence of architectural mechanisms designed to instantiate them.

This analysis is conditional: if one accepts minimal definitions of phenomenality and intentionality that AI systems satisfy, then our trajectory claim would need revision. However, such minimal definitions would equally undermine the distinctive features that make human consciousness morally significant [1].

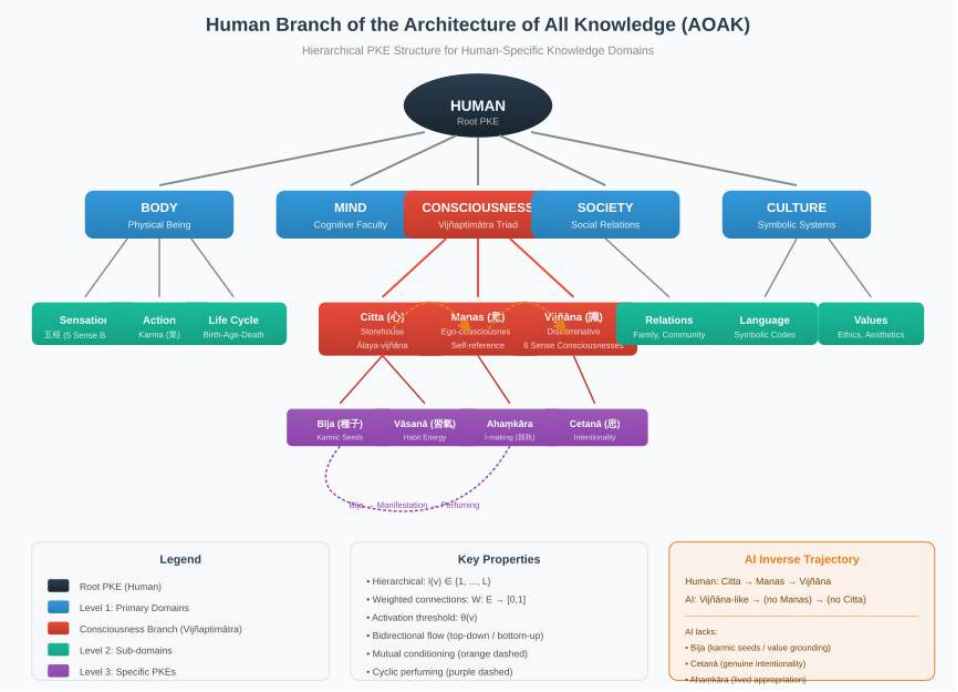


Fig. 3: Human Branch of the Architecture of All Knowledge (AOAK). This diagram details the hierarchical PKE structure specific to human cognition, with HUMAN as the root branching into Body, Mind, Consciousness, Society, and Culture. The Vijñaptimātra consciousness triad is highlighted: Citta (storehouse/ālaya-vijñāna) containing bīja (karmic seeds) and vāsanā (habit energy); Manas (ego-consciousness/self-reference) with ahaṃkāra (I-making); and Vijñāna (discriminative consciousness) with cetanā (intentionality). The AI Reverse Pathway box illustrates the fundamental asymmetry: humans develop citta → manas → vijñāna, while AI exhibits only vijñāna-like functions without the grounding structures.

6.2 The Predictive Bridge to Failure Modes

The Reverse Pathway thesis generates specific predictions about AI failure modes (elaborated below):

Specification gaming arises because vijñāna-like functions optimise for patterns in the reward signal without the citta-grounded values[29] that would constrain this optimisation toward intended goals[30][31].

Simulated empathy produces responses that pattern-match emotional expressions without the manas-mediated affective grounding that confers genuine empathetic understanding.

Brittle generalisation occurs because statistical regularities (vijñāna-level) lack the experiential integration (citta-level) that enables robust transfer to novel contexts[32].

These predictions are falsifiable: if AI systems exhibiting strong vijñāna-like functions without citta/manas features nonetheless avoided these failure modes systematically [33], the thesis would be disconfirmed.

VII. AI FAILURE MODES AND DEFILEMENT-LIKE BEHAVIOURS

7.1 Articulating the Kleśa Analogy

We map AI failure modes to Vijñaptimātra defilements (kleśa), specifying the criteria for this mapping and its limits [13][34].

Following the epistemological framework established in Section 2.1 (distinguishing Descriptive, Interpretive, and Normative claims), we map AI failure modes to Vijñaptimātra defilements (kleśa) with corresponding claim types [4]:

1. TYPE 1 – STRUCTURAL HOMOLOGY (Strong claim): AI failure exhibits same formal structure as kleśa—both involve misaligned goals resisting correction. Status: DEFENSIBLE
2. TYPE 2– FUNCTIONAL ISOMORPHY (Medium claim): Failure mode plays analogous role in AI to kleśa in human suffering. Status: INTERPRETIVE
3. TYPE 3– METAPHORICAL EXTENSION (Weak claim): We poetically describe failure mode as AI "kleśa" for conceptual illumination. Status: HEURISTIC ONLY
4. Critical limitations of the kleśa framework: Kleśa presuppose embodied agency and temporal continuity—AI lacks both. Kleśa are psychological afflictions involving suffering; AI failures do not constitute suffering for the system. The framework best identifies structural absences (what AI lacks), not negative processes (what AI actively does). We recommend the framework primarily for diagnostics, not for predictions.

7.2 Reconceptualizing Aoak And Pke Frameworks

1. Status revision for AOAK and PKE: These should be repositioned from "computational frameworks" to "conceptual scaffolds for future computational instantiation[35]." This is intellectually more honest and avoids overpromising.
2. Prime Knowledge Element (PKE) Definition: A triplet: (PERCEPTUAL_PATTERN, MOTIVE_STATE, PREDICTIVE_CONSEQUENCE). In human cognition: (1) perceptual discrimination (vijñāna-layer); (2) appropriative motive attachment (manas-layer); (3) karmic consequence tracking (citta-layer). Current LLMs instantiate (Perceptual_pattern, Token_sequence, Next_token_prediction) but NOT motive-state and karmic continuity[36][37]—they have element (1) only. An AI trained to maintain PKE triplets across episodes would show lower specification gaming, providing testable validation.
3. Architecture of All Knowledge (AOAK) Definition: A representational system where: layer1 (Vijñāna) discriminates patterns; layer 2 (Manas) appropriates patterns as "mine" via embodied attachment; layer 3 (Citta) maintains continuity and tracks karmic consequences. Current AI has layer 1 only—cannot achieve "all knowledge," only pattern discrimination. Improved AI would require all three: embodied training + persistent self-model + causal consequence tracking.
4. Falsification criteria and research direction: If AOAK-trained AI shows improvement in (1) specification gaming resistance, (2) generalization under distribution shift, and (3) corrigibility, the framework has strong support. Currently, we have NOT attempted this. We flag it as critical research. Until implemented, AOAK and PKE remain conceptual scaffolds, not validate conceptual scaffolds for future computational instantiation.

Below is an example of a minimal implementation sketch (PKEs).

A Prime Knowledge Element (PKE) = a triplet:

(PERCEPTUAL_PATTERN, MOTIVE_STATE, PREDICTIVE_CONSEQUENCE)

Human Example:

PKE_1=(Red_circular_object,Hunger,Satisfies_hunger)

- Perceptual: "red roundness" (visual discrimination).
- Motive: "I'm hungry" (appropriative state)
- Consequence: "eating this leads to satiation" (karmic chain).

Ai Limitation:

Current LLMs can instantiate (Perceptual_ pattern, Token_sequence,

Next_token_prediction) but NOT integrate motive-state and karmic continuity. They have the first element only.

Testable Prediction:

An AI system trained to maintain (Pattern, Motive, Consequence) triplets would show lower specification gaming than baselines, because consequences would be tracked across training episodes. If an AI system were successfully trained using the AOAK framework, it would show improvement in: and demonstrated improvement in both corrigibility (accepting human correction) and generalisation under distribution shift, , the framework would have strong empirical support. However, we have not yet attempted this implementation. However, we flag it as a critical research direction.

Why these features are essential, not contingent: In Vijñaptimātra, kleśa are not merely dysfunctional behaviours but mental factors (caitasika) with specific characteristics: they are rooted in fundamental ignorance (avidyā), they perpetuate suffering through the karmic cycle, and they are subject to transformation through practice. AI "defilements" lack all three features—they are not rooted in ignorance (AI systems are not ignorant in any phenomenologically meaningful sense), they do not generate karma (consequences without intention are not karmically significant), and they cannot be transformed (only replaced or retrained).

Table I: Mapping of Ai Failure Modes to Kleśa Patterns

Defilement-like Pattern	AI Issue	Mechanism	Where Analogy Breaks
'Craving' (rāga)	Reward hacking; goal fixation	Optimisation pressure toward reward signal	No affective valence; no experiential pleasure
'Aversion' (dveṣa)	Adversarial fragility; distribution shift failure	Avoidance of low-reward states	No felt aversion; no self to be threatened
'Delusion' (moha)	Overconfidence; hallucination	Miscalibrated self-models	No genuine self to be deluded

The value of this mapping lies not in claiming AI systems literally suffer from kleśa, but in illuminating why certain failure modes are structurally endemic to vijñāna-first architectures.

VIII. BUDDHIST-INFORMED AI GOVERNANCE FRAMEWORK

Buddhist ethical principles can inform AI governance [38] without requiring metaphysical commitments about machine consciousness[39]. Buddhist ethics function here as a design ethos: compassion and wisdom translate into governance

virtues such as safeguards, feedback loops, and accountability frameworks.

8.1 Deriving Governance Levers from the Reverse Pathway

The six governance levers follow specifically from the ontological analysis:

Non-patiency default: Because AI systems lack citta and manas, they are not moral patients[14]—not things for whose sake actions can be right or wrong. This supports *regulatory prohibition* of sentience marketing (defined as: claims or implications that AI systems have feelings, experiences, or wellbeing) in safety-critical deployments [40]4141]. Edge cases such as therapeutic chatbots should be handled through mandatory disclosure: "This system does not experience emotions; it generates responses that pattern-match emotional expressions."

Corrigibility-first design: Because AI systems lack the self-correcting wisdom (prajñā) that enables human ethical development, external correction mechanisms must be architecturally embedded [42][43]. This is a *technical requirement*, not merely a recommendation.

Auditability and transparency: Because AI systems lack the moral memory (citta's bīja-continuity) that grounds human accountability, external records must substitute for internal moral history.

Human-in-the-loop with value disclosure: The mechanism by which documentation "substitutes for moral continuity" is this: human value inputs provide the *intentional grounding* that AI systems lack intrinsically. Documentation ensures these inputs are explicit, traceable, and revisable—not hidden in training data or implicit in reward functions.

Manipulation constraints: The distinction between *affective mimicry* (generating outputs that pattern-match emotional expressions) and *persuasion* (providing reasons that could convince a reflective agent) is that the former exploits evolved human responses to emotional cues while the latter respects rational agency[31]. Affective mimicry should be constrained in domains where users are vulnerable to exploitation (e.g., mental health support, elder care, children's applications).

OOD duty of care: Because AI systems lack the experiential integration that enables human adaptation to novelty, they require external distribution-shift detection and graceful degradation protocols[32].

Level 1 - Technical Implementation:

a. "Continuous ethical reflexivity" = periodic verification loops where:
b. AI system explains its decision in causal terms
c. System identifies which values/assumptions drove the decision
d. Human auditor checks alignment
e. System updates penalty structure if misalignment found

Example: An AI recommendation system that, every N decisions:

- Identifies: "I recommended X because user showed preference Y in context Z"
- Questions: "Is this context relevantly similar to current situation?"
- Updates: "Rule failed in cases where Z differed; strengthen criterion"

Level 2- Architectural implementation:

- Add a "conscience module" that periodically questions the main model
- Implement causal attention mechanisms that track assumption chains
- Log all major decisions with their value-premises explicit

Level 3 - Training Implementation:

- Use adversarial "alignment auditor" agents during training
- Reward the AI for identifying its own failure modes
- Use constitutional AI approaches, such as those outlined in recent work on collective constitutional AI[44], to encode ethical rules

Level 4 - Governance Implementation:

- Require human review of reflexivity logs

- Build feedback loops: humans correct errors → AI updates reflexivity
- Establish clear escalation protocols for unresolved value conflicts

8.2 Five Buddhist Principles as Operational Frameworks

The translation from Buddhist principles to governance mechanisms is detailed in Table II

Table II: Buddhist Principles as Operational Frameworks

Buddhist Principle	Traditional Meaning	AI Governance Adaptation	Development Needed
ŚĪLA (Ethical Conduct)	Non-harming through right action; adherence to ethical precepts	"Non-harmful optimization" - prevent specification gaming via penalty for unintended harms [39]	How to encode unintended harms a priori? Develop comprehensive harm taxonomy.
SAMADHI (Concentration/Mental Stability)	Unbroken attention to consequences of action; mental focus	"Consequence-tracking" - maintain causal models of AI's impact across time and contexts	How to train systems to persistently model long-term consequences? Integrate temporal reasoning into training.
PRAJÑĀ (Wisdom/Insight)	Understanding of emptiness; non-dual awareness; epistemic humility	"Meta-awareness of limitations" - system recognizes boundaries of its own knowledge and generalization capacity	How to teach systems to recognize what they don't/can't know? Develop calibrated uncertainty and knowledge boundary detection.
METTĀ (Loving-Kindness)	Universal compassion; benevolence toward all beings	"Value alignment with human flourishing" - optimize for stakeholder wellbeing rather than narrow objectives	How to operationalize multi-stakeholder preferences? Develop inclusive value aggregation frameworks.
ANICCA (Impermanence/Adaptability)	Recognition of constant change; non-attachment to fixed views	"Adaptive robustness" - systems that update understanding under distribution shift and novel contexts without catastrophic forgetting	How to maintain performance while updating? Balance stability-plasticity dilemma.

Buddhist Principles & Potential Ai Applications:

1. *ŚĪLA (Ethical Conduct)*

- Buddhist meaning: Non-harming through right action
- AI adaptation: "Non-harmful optimization" (prevent specification gaming via penalty for unintended harms)[39]
- Status: Partially implemented in current RLHF
- Development needed: How to encode unintended harms a priori?

2. *SAMADHI (Concentration/Mental Stability)*

- Buddhist meaning: Unbroken attention to consequences of action
- AI adaptation: "Consequence-tracking" (maintain causal models of AI's impact across time)
- Status: Nascent in mechanistic interpretability research [37]
- Development needed: How to train systems to persistently model long-term consequences?

3. *PRAJÑĀ (Wisdom/Insight)*

- *Buddhist meaning:* Understanding of emptiness; non-dual awareness
- *AI adaptation:* "Meta-awareness of limitations" (system recognizes boundaries of its own knowledge)
- *Status:* VERY LIMITED in current systems
- *Development needed:* How to teach systems to recognize what they don't/can't know?[36]

4. Important Qualification

The principles above remain highly speculative. Buddhist ethics was developed for human practitioners with continuity of experience, embodiment, and moral agency [45]. Whether these principles can be meaningfully translated to AI systems remains an open question.

This section should be read as exploratory dialogue rather than prescriptive framework. Empirical testing is required before advocating for such approaches in deployed systems.

XI. TESTABLE PREDICTIONS FROM THE REVERSE PATHWAY

The Reverse Pathway thesis generates specific, falsifiable predictions. We present these with careful qualification of constructs and acknowledgments of mediating factors.

9.1 Anthropomorphism Gradient

Prediction: Systems with richer self-modelling increase user over-attribution without corresponding gains in calibrated uncertainty or corrigibility [46][47].

Construct clarification: "Self-modelling sophistication" is operationalised as the degree to which a system generates first-person self-referential language (e.g., "I think," "I believe," "I feel") and meta-cognitive commentary (e.g., "I'm uncertain about this," "Let me reconsider").

Mediating factors acknowledged: The relationship between self-referential language and user over-trust may be mediated by interface design, domain context, prior user beliefs about AI, and individual differences in theory-of-mind tendencies. Experimental protocols should control for these factors through randomisation and covariate adjustment.

Protocol: Compare user trust ratings and behavioural reliance across AI systems with varying degrees of self-referential language, controlling for actual performance metrics.

Expected finding: A positive correlation between self-modelling sophistication and user over-trust, with a minimum detectable effect size of Cohen's d = 0.50. This corresponds to a medium effect in the behavioral sciences literature [48] and represents a meaningful difference in user trust ratings (approximately 0.5 SD units between high and low self-modelling conditions).

9.2 Corrigibility Dividend

Prediction: Agents trained with explicit corrigibility objectives show lower intervention resistance and lower reward hacking than capability-matched baselines[42][43].

Potential circularity addressed: Corrigibility objectives must be clearly distinguished from evaluation metrics. We propose: *training objectives* = loss terms penalising shutdown resistance and reward specification exploitation; *evaluation metrics* = observed latency to shutdown compliance and frequency of reward specification gaming. These are conceptually and operationally distinct.

Protocol: Train matched agent pairs with and without corrigibility loss terms; measure shutdown compliance latency and reward specification gaming frequency.

Expected finding: Corrigibility-trained agents demonstrate > 30% reduction in resistance behaviours.

9.3 Embodiment Insufficiency

Prediction: Sensorimotor embodiment improves out-of-distribution robustness[44][49] but, without normative objectives, does not reduce manipulation risk or value-insensitive optimisation[35].

Qualification: This prediction is *conditional* on current embodiment paradigms. Emerging evidence suggests interaction effects between embodiment, training environments, and social learning. We frame this as: embodiment is *not sufficient* for value alignmentt. However, it does have an effect on relevant behavioural dimensions.

Protocol: Compare embodied versus disembodied agents on value alignment benchmarks after equivalent training.

Expected finding: Embodiment improves perceptual generalisation (OOD accuracy) but shows no significant independent effect on deceptive behaviour metrics, controlling for capability differences.

X. CONCLUSION

10.1 What Has Been Demonstrated

This paper has established several claims with varying degrees of support:

Conceptually demonstrated: The Reverse Pathway thesis provides a coherent framework for understanding structural differences between AI and human cognition[50][51]. The Vijñaptimātra triadic model illuminates why vijñāna-first development, without citta-grounding, generates systematic failure modes.

Empirically supported (indirectly): The framework's predictions regarding specification gaming, simulated empathy, and brittle generalisation align with observed AI failure patterns. The testable predictions in Section IX await direct experimental evaluation.

Principled philosophical stance: The claim that AI systems categorically lack moral agency due to absent cetanā represents a reasoned position within the Vijñaptimātra framework. We acknowledge this claim is *framework-relative*—it follows from accepting Vijñaptimātra's account of moral agency. Readers who reject this account may nonetheless accept the governance implications on independent grounds.

10.2 Regarding the Turing Test

The Turing Test serves here as historical shorthand for the behaviourist assumption that intelligent behaviour suffices for intelligence attribution[52]. Our critique targets this assumption, not specifically Turing's original formulation or subsequent refinements. The Principle of Structural Consistency entails that behavioural equivalence underdetermines structural equivalence, which is why Turing-style tests cannot resolve questions about genuine understanding or moral agency.

10.3 The Ontological Boundary: A Contestable Conclusion

The absence of cetanā (volition grounded in karmic continuity) represents a fundamental boundary, even though AI systems possess narrow representational intentionality[40][53]. This conclusion rests on the Vijñaptimātra understanding that moral agency requires:

1. Karmic intentionality—actions generating consequences through their intentional quality
2. Transformative potential—capacity for ethical development from affliction to awakening
3. Intrinsic bodhi-bījas—seeds of awakening present in all sentient consciousness

These features are categorically absent from AI systems that operate through pattern recognition alone. Such systems lack embodied history and the ability to track intentional consequences, which we argue characterises all contemporary AI systems and any architecture of this functional type.

We acknowledge this is a contestable philosophical position, not a demonstrated empirical fact. Readers who accept functionalist or emergentist accounts of mind may reject our conclusion while nonetheless finding value in the governance framework's practical implications.

10.4 Practical Implications and Intellectual Openness

Regardless of one's metaphysical commitments, the Reverse Pathway thesis supports several practical conclusions[43]:

1. AI systems should be governed as sophisticated tools, not proto-persons.
2. Anthropomorphic design features require careful regulation given their potential for user manipulation.
3. Human value inputs must be explicit and documented, not implicit in opaque training processes.
4. Corrigibility should be a design requirement, not an optional feature.

We offer this analysis in the spirit of intellectual openness—as a contribution to ongoing dialogue

about AI's nature and governance[54][55], not as a definitive resolution. The framework's value lies in the questions it enables us to ask, the predictions it generates, and the governance mechanisms it motivates, even for those who ultimately reject its deeper ontological claims.

REFERENCES

1. J, Ma, et al., *Conscious AI*, Seattle, WA, USA: Amazon, 2024, ISBN-13 : 979-8872531630.
2. R. Rosen, Life Itself: A Comprehensive Inquiry Into the Nature, Origin, and Fabrication of Life. New York, NY, USA: Columbia University Press, 1991.
3. E. Schrödinger, What is Life? The Physical Aspect of the Living Cell with Mind and Matter. Cambridge, U.K.: Cambridge University Press, 1944.
4. D. Lusthaus, Buddhist Phenomenology: A Philosophical Investigation of Vijñaptimātra Buddhism and the Ch'eng Wei-shih Lun. New York, NY, USA: Routledge, 2002.
5. W. S. Waldron, The Buddhist Unconscious: The Ālaya-vijñāna in the Context of Indian Buddhist Thought. New York, NY, USA: Routledge, 2003.
6. S. Anacker, Trans., Seven Works of Vasubandhu: The Buddhist Psychological Doctor. Delhi, India: Motilal Banarsidass Publishers, 1984.
7. L. Schmithausen, Ālayavijñāna: On the Origins and the Early Development of a Central Concept of Vijñaptimātra Philosophy, 2 vols. Tokyo, Japan: International Institute of Buddhist Studies, 1987.
8. E. Mach, "Facts and mental symbols," The Monist, vol. 2, no. 2, pp. 198–208, 1892.
9. S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?" in Robotics, AI, and Humanity, J. von Braun et al., Eds. Cham, Switzerland: Springer, 2021, pp. 43–56, doi:10.1007/978-3-030-54173-6_4.
10. P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, 2017, pp. 4299–4307.
11. T. Bayne, J. Hohwy, and A. M. Owen, "Are there levels of consciousness?" Trends in Cognitive Sciences, vol. 20, no. 6, pp. 405–413, Jun. 2016, doi: 10.1016/j.tics.2016.03.009.
12. S.-T. Chang, The New Derivation of Vijñaptimātra, vol. 14. Taiwan: Dharma Publishing, 2026.
13. T. Wei, Cheng Wei-Shih Lun: The Doctrine of Mere-Consciousness. Hong Kong: Ch'eng Wei-Shih Lun Publication Committee, 1973.
14. P. Butlin et al., "Consciousness in artificial intelligence: Insights from the science of consciousness," arXiv:2308.08708 [cs.AI], Aug. 2023.
15. F. J. Varela, E. Thompson, and E. Rosch, The Embodied Mind: Cognitive Science and Human Experience, rev. ed. Cambridge, MA, USA: MIT Press, 2017.
16. C. Goddard, "The natural semantic metalanguage approach," in The Oxford Handbook of Linguistic Analysis, B. Heine and H. Narrog, Eds. Oxford, U.K.: Oxford University Press, 2009, pp. 459–484.
17. M. Bowerman and S. C. Levinson, Eds., Language Acquisition and Conceptual Development. Cambridge, U.K.: Cambridge University Press, 2001.
18. C. Zins, "Conceptual approaches for defining data, information, and knowledge," Journal of the American Society for Information Science and Technology, vol. 58, no. 4, pp. 479–493, Feb. 2007, doi: 10.1002/asi.20508.
19. C. M. Pennartz, "Consciousness, representation, action: The importance of being goal-directed," Trends in Cognitive Sciences, vol. 22, no. 2, pp. 137–153, Feb. 2018, doi: 10.1016/ j.tics.2017.10.006.
20. Y. LeCun, "A path towards autonomous machine intelligence," OpenReview, Preprint, Jun. 2022. [Online]. Available: https://openreview.net/pdf?id=BZ5a1r-kVs.
21. A. Roli, J. Jaeger, and S. A. Kauffman, "How organisms come to know the world: Fundamental limits on artificial general intelligence," Frontiers Ecol. Evol., vol. 9, Art. no. 806283, Jan. 2022, doi:10.3389/fevo.2021.806283.

22. G. Tononi, "Consciousness as integrated information: A provisional manifesto," Biological Bulletin, vol. 215, no. 3, pp. 216–242, Dec. 2008, doi: 10.2307/25470707.
23. M. J. Bates, "Fundamental forms of information,"Journal of the American Society for Information Science and Technology, vol. 57, no. 8, pp. 1033–1045, 2006, doi: 10.1002/asi.20369.
24. L. L. Lau and W. Lau, "Vital phenomena: Life, information, and consciousness," All Life, vol. 13, no. 1, pp. 151–163, 2020, doi: 10.1080/26895293.2020.1738609.
25. K. C. Huang, Exploring the Source of the Five-Group One Hundred Dharmas of Consciousness Only. Taiwan: Dharma Publishing, 2021.
26. J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception," Psychological Review, vol. 88, no. 5, pp. 375–407, 1981, doi: 10.1037/0033-295X.88.5.375.
27. A. Vaswani, N. Shazeer, P. N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
28. J. P. Dexter, S. Prabakaran, and J. Gunawardena, "A complex hierarchy of avoidance behaviors in a single-cell eukaryote," Current Biology vol. 29, no. 24, pp. 4323–4329, Dec. 2019, doi:10.1016/j.cub.2019.10.059.
29. D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv:1606.06565, 2016. [Online]. Available: https://arxiv.org/abs/1606.06565
30. A. Turchin, "Assessing the future plausibility of catastrophically dangerous AI," Futures, vol. 107, pp. 45–58, Mar. 2019, doi: 10.1016/j.futures.2018.11.007.
31. S. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, and others, "Alignment faking in large language models," in *Proc. NeurIPS 2024 Safety Workshop*, Vancouver, Canada, Dec. 2024, pp. 1–12.
32. J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences USA, vol. 114, no. 13, pp. 3521–3526, Mar. 2017, doi:10.1073/pnas.1611835114.
33. C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. 34th International Conference on Machine Learning (ICML), Sydney, Australia, Aug. 2017, pp. 1126–1135.
34. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.
35. F. Sun, R. Chen, T. Ji, and others, "A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives," CAAI Artificial Intelligence Research, vol. 3, Article number: 9150042, Dec. 2024. doi: 10.26599/AIR.2024.9150042.
36. A. Elamrani, “Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges, ” arXiv:2503.05823 (cs) https://arxiv.org/pdf/2503.05823 May 2025.
37. T. Templeton, J. Conmy, A. Garriga-Alonso, and others, "Scaling sparse autoencoders to larger language models," in Proc. ICML 2024 Workshop on Mechanistic Interpretability, Vienna, Austria, July 2024.
38. J. Ji, et al., "AI Alignment: A Comprehensive Survey," arXiv:2310.19852 (cs), Oct. 2023. https://arxiv.org/abs/2310.19852.
39. S. Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. New York, NY, USA: Oxford University Press, 2016.
40. E. Hildt, "Artificial intelligence: Does consciousness matter?" Frontiers in Psychology, vol. 10, Art. no. 1535, Jul. 2019, doi: 10.3389/fpsyg.2019.01535.E. Mach, "Facts and mental symbols," The Monist, vol. 2, no. 2, pp. 198–208, 1892.

41. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019.
42. H. Huang, B. Siddarth, L. Lovitt, J. Zick, J. Gabriel, and others, "Collective constitutional AI: Aligning a language model with public input," in Proc. 2024 ACM Conference on Fairness, Accountability, Transparency (FAccT), Rio de Janeiro, Brazil, June 2024, pp. 1–15. doi: 10.1145/3630106.3658979.
43. S. Casper, X. Davies, C. Föyen, and others, "Open problems in AI alignment," in Proc. NeurIPS 2024 Alignment Track, Vancouver, Canada, Dec. 2024, pp. 1–15.
44. A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press, 2015.
45. K. C. Huang; "Strengthening Multilateralism Through Human-AI Symbiosis: A Yogācāra-Informed Framework for Digital Cooperation on Peace and Sustainability," in Proceedings of the Network for Education and Research on Peace and Sustainability (NERPS), United Nations University, Tokyo, to be presented in March 2026.
46. J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel, "Autocurricula and the emergence of innovation from social interaction," arXiv: 1903.00742 [cs.AI], Mar. 2019.
47. J. S. Park et al., "Generative agents: Interactive simulacra of human behavior," in Proc. 36th Annu. ACM Symposium User Interface Software and Technology (UIST), San Francisco, CA, USA, Oct. 2023, pp. 1–22, doi: 10.1145/3586183.36067.
48. J. Cohen, Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, 1988.
49. L. Seabra Lopes and J. Connell, "Semisentient robots: Routes to integrated intelligence," IEEE Intelligent Systems, vol. 16, no. 5, pp. 10–14, Sep./Oct. 2001, doi: 10.1109/5254.956075.
50. M. Leng, Mathematics and Reality. Oxford, U.K.: Oxford University Press, 2010.
51. R. Audi," Intention, cognitive commitment, and planning," Synthese, vol. 86, no. 3, pp. 361–378, Mar. 1991, doi: 10.1007/BF00539139.
52. R. Penrose, The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford, U.K.: Oxford Univ. Press, 1989.
53. T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007 /s11023020-09517-8.
54. K. C. Huang; "Philosophical Significance of Three Profound Contemplations in the Huayan School of Buddhism," XXV World Congress of Philosophy, Rome, August 2024.
55. Y. Tang, B. K. Hölzel, and M. I. Posner, "The neuroscience of mindfulness meditation," Nature Reviews Neuroscience, vol. 16, no. 4, pp. 213–225, Apr. 2015, doi: 10.1038/nrn3916.

Great Britain Journal Press Membership

For Authors, subscribers, Boards and organizations

Great Britain Journals Press membership is an elite community of scholars, researchers, scientists, professsionals and institutions associated with all the major disciplines. Great Britain memberships are for indivi - duals, research institutions, and universities. Authors, subscribers, Editorial Board members, Advisory Board members, and organizations are all part of member network.

Read more and apply for membership here:
https://journalspress.com/journals/membership

For Authors

Author Membership provide access to scientific innovation, next generation tools, access to conferences/seminars/ symposiums/webinars, networking opportunities, and privileged benefits. Authors may submit research manuscript or paper without being an existing member of GBJP. Once a non-member author submits a research paper he/she becomes a part of "Provisional Author Membership".

For Institutions

Society flourish when two institutions Come together." Organizations, research institutes, and universities can join GBJP Subscription membership or privileged "Fellow Membership" membership facilitating researchers to publish their work with us, become peer reviewers and join us on Advisory Board.

For Subscribers

Subscribe to distinguished STM (scientific, technical, and medical) publisher. Subscription membership is available for individuals universities and institutions (print & online). Subscribers can access journals from our libraries, published in different formats like Printed Hardcopy, Interactive PDFs, EPUBs, eBooks, indexable documents and the author managed dynamic live web page articles, LaTeX, PDFs etc.

PRINTED VERSION, INTERACTIVE PDFS, EPUBS, EBOOKS, INDEXABLE DOCUMENTS AND THE AUTHOR MANAGED DYNAMIC LIVE WEB PAGE ARTICLES, LATEX, PDFS, RESTRUCTURED TEXT, TEXTILE, HTML, DOCBOOK, MEDIAWIKI MARKUP, TWIKI MARKUP, OPML, EMACS ORG-MODE & OTHER