EDITORIAL NOTE

It is with great pride and pleasure that we present the inaugural volume of the Islington Journal of Multidisciplinary Research (IJMR), the official peer-reviewed research journal of Islington College, Kathmandu, Nepal. This first volume marks a significant milestone in our commitment to promoting research, innovation, and academic excellence while fostering a vibrant culture of scholarly inquiry.

The primary objective of IJMR is to provide a credible platform for researchers, academicians, professionals, and students to disseminate original, high-quality, and impactful multidisciplinary research. Through this journal, we seek to encourage intellectual exchange, interdisciplinary collaboration, and the advancement of knowledge that addresses contemporary challenges at local, national, and global levels.

The inaugural volume features six peer-reviewed research articles spanning diverse fields, including artificial intelligence, healthcare, finance, behavioral finance, banking, public policy, and digital marketing. Collectively, these contributions demonstrate the value of multidisciplinary research in addressing real-world challenges and advancing evidence-based knowledge across different domains.

The Editorial Board extends its sincere appreciation to all authors for their valuable contributions and to our reviewers for their dedication in maintaining the quality and integrity of the publication through a rigorous peer-review process. We also express our heartfelt gratitude to the management and academic community of Islington College for their continued support in making the launch of this journal possible.

We hope this inaugural issue provides valuable insights to researchers, educators, practitioners, policymakers, and students, while inspiring further research, innovation, and interdisciplinary collaboration. We warmly invite scholars from Nepal and around the world to contribute to future volumes of IJMR as we continue to build a journal that promotes excellence in research and meaningful academic engagement.

Editorial Board
Islington Journal of Multidisciplinary Research (IJMR)

CONTENTS

Islington Journal of Multidisciplinary Research (IJMR)

Volume 1, Number 1, 2026

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 1–11

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/e5vr3575

AI Fraud Detection and Financial Trust in Nepals SME Payment Ecosystem: A Readiness Framework

Yogesh Pant[1,*], Aditya Pudasaini[1], Roshan Shrestha[1], Alish K.C.[1]

[1]BSc (Hons) Computing, Islington College, Kathmandu, Nepal
[*]*Correspondence: yogeshpant911@gmail.com*

ARTICLE HISTORY

Received: 5 March 2026 Revised: 8 April 2026
Accepted: 25 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Pant et al. (2026). 'AI fraud detection and financial trust in Nepals SME payment ecosystem: A readiness framework', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 1–11. Available at: https://doi.org/10.67556/e5vr3575

ABSTRACT

Nepals rapid transition toward digital payments has outpaced the development of cybersecurity and data-governance infrastructure, leaving small and medium enterprises (SMEs) exposed to rising cyber-enabled fraud and an accompanying deficit of financial trust. Artificial intelligence (AI)-based fraud detection has matured into a standard layer of defense in data-rich financial markets, but these systems rest on assumptions, namely high data maturity, cloud infrastructure, and developed explainability regimes, that do not hold in Nepals fragmented, infrastructure-constrained payment ecosystem. This study conducts a structured integrative review of literature spanning AI-based fraud detection, SME digital trust, and Nepals payment landscape to examine the resulting transferability gap between global AI capabilities and local readiness. Thematic synthesis of the reviewed sources shows that supervised models, behavioral analytics, and graph-based detection cannot be directly transplanted into Nepals context because of data fragmentation, reactive fraud reporting, weak enforcement of Explainable AI (XAI) norms, and acute psychological resistance among merchants. To address this mismatch, the study proposes a five-stage conceptual readiness framework, derived through an explicit gap-to-stage mapping process and assessed through theoretical grounding and comparative benchmarking against existing AI-adoption and maturity models, that sequences data interoperability, unsupervised anomaly detection, economic triage, and XAI compliance ahead of advanced model deployment. The framework offers policymakers, fintech developers, and financial institutions a context-sensitive, trust-centered roadmap for responsible AI-enabled fraud management in Nepal and comparable emerging economies.

Keywords: AI fraud detection, digital payments, financial trust, SMEs

Introduction

Background and Motivation

Over the last five years, Nepals financial system has shifted rapidly from a cash-based to a digital-payment-based economy. The shift accelerated following the COVID-19 pandemic and a series of supportive policies from Nepal Rastra Bank (NRB), the countrys central bank. By mid-2024, mobile banking had reached 24.6 million users, and digital wallets had become commonplace among urban and semi-urban SMEs (Nepal Rastra Bank 2024). For Nepali SMEs, QR-code and mobile-wallet adoption has moved from a competitive advantage to an operational necessity for remaining competitive and accessing credit facilities (Tamang, Bhaskar and Chatterjee 2021).

Globally, the same period has seen AI-based fraud detection mature into a standard layer of financial-crime defense. Supervised classification, unsupervised anomaly detection, behavioral analytics, and graph-based detection are now embedded across financial institutions in data-rich markets (Hilal, Gadsden and Yawney 2022; Hernandez Aros et al. 2024; Nicholls, Kuppa and Le-Khac 2021; Vanini et al. 2023), and supervisory authorities in other emerging markets are actively building internal capacity to monitor AI-related risks, even as that capacity remains uneven across jurisdictions (Crisanto et al. 2024).

Nepals pace of digitalization, however, has consistently outpaced its security and data-governance infrastructure, leaving SMEs exposed to fraud risks they are ill-equipped to absorb. The Financial Intelligence Unit (FIU) has reported that fake payment screenshots, OTP hijacking, and the use of students as money mules account for over 63% of all suspicious transaction reports (Financial Intelligence Unit, Nepal Rastra Bank 2024), and fraud involving digital applications is reported to have increased by 1,225% between 2020 and

2025 (Dhital 2025). For SMEs operating with little margin for error, these schemes are not only a direct financial loss but also erode trust in digital systems, pushing some merchants back toward untraceable cash transactions.

Although major financial institutions worldwide use AI and machine learning to track transactions and flag anomalies in real time, transplanting these solutions into Nepal is not straightforward. AI-based fraud detection of this kind depends on cleanly labeled datasets, cloud computing, and mature regulatory frameworks for explainability, conditions that are presently underdeveloped in Nepal. Deploying such systems prematurely risks generating high false-positive rates that would deepen, rather than resolve, the existing trust deficit (Widayani, Fiernaningsih and Herijanto 2022).

NRB has begun to respond. It has amended its payment-system directives to require incident logging and multi-year business-continuity planning by payment service providers (The Kathmandu Post 2025b), drafted AI-specific guidelines covering fraud detection, credit scoring, and risk management for licensed financial institutions (The Kathmandu Post 2025a), and finalized a Regulatory Sandbox through which new technologies, including AI-based services, can be tested under live supervision (Nepal News 2026). These measures suggest that the regulatory groundwork for AI-enabled fraud detection is beginning to take shape, but they remain early-stage and largely procedural, and do not yet resolve the underlying data fragmentation, infrastructure gaps, and explainability deficits that determine whether AI-based detection can be deployed responsibly.

Scope, Objectives and Research Questions

This paper examines how AI-based fraud detection can be made to work for, rather than against, financial trust and resilience in the digital payments of Nepali SMEs. The scope covers three areas: trends in digital-payment adoption among Nepali SMEs; technical approaches to digital-payment fraud detection available in the literature; and the infrastructural, regulatory, economic, and behavioral factors that determine AI readiness in Nepals payment ecosystem. The study does not involve primary data collection, model building, or model testing; it is bounded to a structured review and conceptual synthesis of existing literature. The empirical scope is geographically limited to Nepal, though international literature is used throughout for comparison and theoretical grounding.

The study is guided by three objectives, each paired with a corresponding research question:

- **Objective 1:** To review the literature on digital payments, fraud detection, and SME financial trust in relation to AI.
  RQ1: What AI-based fraud detection capabilities have been developed in data-mature financial markets, and what assumptions about data, infrastructure, and regulation underlie their effectiveness?
- **Objective 2:** To identify the infrastructural, regulatory, economic, and behavioral readiness challenges affecting AI adoption in Nepals SME payment ecosystem.
  RQ2: What infrastructural, regulatory, economic, and behavioral barriers constrain the transfer of global AI fraud-detection capabilities into Nepals SME digital-payment ecosystem?
- **Objective 3:** To propose a conceptual readiness framework that supports policymakers, fintechs, and financial institutions in moving toward secure, AI-enabled transactions.
  RQ3: How can a phased readiness framework be derived from the identified barriers and validated, within the bounds of a conceptual study, so that it sequences AI adoption in a way that is both technically feasible and trust-preserving for Nepals SMEs?

Research Contribution

This paper makes four primary contributions: (i) It synthesizes global AI-based fraud detection capabilities and evaluates their applicability within Nepals financial ecosystem, (ii) It formally defines and examines the transferability gap associated with deploying AI fraud detection systems in resource-constrained environments, (iii) It proposes a five-stage conceptual readiness framework tailored to low-data, high-risk settings, and (iv) It advances a trust-centered perspective on AI adoption, emphasizing Explainable AI (XAI) and economic feasibility alongside algorithmic performance. (v) It demonstrates a replicable conceptual validation methodology for framework papers in low-resource information-systems contexts, applying theoretical grounding, gap-traceability mapping, and comparative benchmarking as an alternative to empirical validation where primary data collection is outside the study's scope.

Theoretical and Conceptual Foundations

Financial Trust in Digital Ecosystem

Financial trust constitutes the cornerstone of digital payment adoption, reflecting institutional integrity, transaction security, and perceived consumer protection (Rubio and Tulcanaza-Prieto 2025). In developing countries, trust is frequently undermined by historical financial instability, opaque dispute resolution mechanisms, and recurrent cyber incidents. SMEs exhibit heightened trust sensitivity due to thin profit margins, limited risk-bearing capacity, and direct exposure to transaction fraud. Innovation Resistance Theory suggests that psychological barriers including fear, distrust, and perceived loss of control exert a stronger influence on adoption intentions than functional or cost-related factors (Widayani, Fiernaningsih and Herijanto 2022). As a result, technological solutions that appear unclear or make decisions without sufficient transparency may discourage adoption and lead users to revert to cash-based transactions.

AI-Based Fraud Detection Systems

In recent years, AI fraud detection has evolved from traditional rule-based systems to machine learning models capable of analyzing complex patterns within large volumes of transactional data. Supervised models, including Random Forests and Support Vector Machines, rely on extensively labeled historical datasets to classify fraudulent behavior (Hernandez Aros et al. 2024; Dahal, Bhattarai and Karki 2021). However, class imbalance and evolving fraud tactics have spurred interest in unsupervised anomaly detection techniques such as Autoencoders and Isolation Forests, which identify deviations without pre-labeled fraud instances (Hilal, Gadsden and Yawney 2022). Advanced architectures like Graph Neural Networks (GNNs) map relational patterns to detect organized fraud syndicates (Nicholls, Kuppa and Le-Khac 2021). Despite algorithmic sophistication, operational deployment requires deep behavioral session tracking, low-latency API integration, and economic triage models to balance detection accuracy with investigation costs (Vanini et al. 2023; Karki 2012). Importantly, the opacity of deep learning models is inconsistent with regulatory demands for transparency and user trust. Therefore, this necessitates Explainable AI (XAI) frameworks such as LIME and SHAP (Gupta 2024; Hilal, Gadsden and Yawney 2022).

Technology Readiness and Socio-Technical System

The adoption of AI is not just about technical feasibility since it occurs within socio-technical systems characterized by interrelated elements such as the infrastructure, regulations, human resources, and organizational cultures. The theory of socio-technical systems suggests that technical success requires matching the capabilities with the social and organizational context. Readiness assessments include the analysis of maturity in data, interoperability requirements, computing capabilities, and enforcement (Hilal, Gadsden and Yawney 2022; Vanini et al. 2023). Emerging markets lack adequate AI governance, as they have no mandates in place for issues like algorithmic auditability, privacy, and consumer redress programs (Crisanto et al. 2024; Nicholls, Kuppa and Le-Khac 2021). The innovation resistance theory additionally highlights the role of the fear of complexity and insecurity, as well as institutional lack of accountability, in constraining innovation (Widayani, Fiernaningsih and Herijanto 2022; Rubio and Tulcanaza-Prieto 2025).

Review Methodology

Research Design and Rationale for an Integrative Review

This study adopts a structured integrative literature review, a design developed for synthesizing theoretical, empirical, and institutional sources into new conceptual knowledge, rather than aggregating quantitative findings from a methodologically homogeneous body of studies (Torraco 2016; Whittemore and Knafl 2005). An integrative review is appropriate here for two reasons. First, the literature spanning AI-based fraud detection, SME digital trust, and Nepals payment ecosystem is disciplinarily heterogeneous, spanning machine-learning evaluation studies, behavioral-adoption surveys, and institutional or regulatory reports, a mix that does not lend itself to the structured quantitative aggregation used in systematic or meta-analytic reviews. Second, the studys objective is theory-building rather than effect-size estimation: the aim is to construct a new conceptual framework by mapping global AI capabilities against Nepals local realities, which is precisely the kind of generative, cross-domain synthesis the integrative review method is designed to support (Torraco 2016). This contrasts with closely related systematic reviews in similar contexts, such as Ampumuza, Katushabe and Tamale (2026), which are designed primarily to catalogue existing detection methods. Because our objective is to construct a novel conceptual framework rather than merely inventory current tools, we require the broader interpretive synthesis that an integrative design affords.

Data Source and Search Strategy

Literature was drawn from international academic databases (IEEE Xplore, ScienceDirect, SpringerLink, and MDPI), global institutional repositories (including the World Bank Global Findex Database), and localized Nepali institutional sources (Nepal Rastra Bank publications, FIU-Nepal strategic reports, and Nepal Journals Online). During revision, this base was supplemented with recent regulatory and methodological sources, including Bank for International Settlements FSI Insights papers and Nepal Rastra Bank circulars and directives reported through 2026, to capture developments that postdate the original search and to strengthen the methodological grounding requested in review. Search terms combined the core constructs under study (for example, AI fraud detection, SME digital trust, Nepal digital payment, Explainable AI finance, AI readiness framework), adapted to each databases syntax.

Scope and Analytical Lens

Sources were included if they were peer-reviewed or institutionally authoritative (a central bank, financial intelligence unit, or international financial-sector body), published between 2018 and 2026, and substantively addressed fraud, financial trust, AI/ML-based detection, or SME digital-payment adoption. Sources were excluded if they addressed AI fraud detection in technical isolation from any deployment or adoption context, or if they fell outside the 20182026 window without continued citation relevance. The original draft synthesized fifteen core sources across these domains. However, to strengthen the paper and address requests for a broader literature base, we significantly expanded this pool during the revision process. We added twelve new sources, incorporating recent global AI-adoption frameworks, low-resource fraud models, and the latest regulatory updates from Nepal, alongside key methodology papers to justify our review design. The final analysis is now supported by a robust, well-rounded pool of twenty-seven sources.

Analytical Process: From Literature to Framework

The 5-Stage Phased Readiness Framework presented in Section 5 was not asserted a priori, it was derived from the literature through a three-step analytical process.

First, open coding was applied to the retained sources to identify recurring constructs across three domains: (i) the technical capabilities and data/infrastructure assumptions of global AI fraud-detection systems; (ii) the structural and regulatory characteristics of Nepals payment ecosystem; and (iii) the behavioral and psychological determinants of SME trust in digital systems.

Second, axial coding grouped these open codes into five recurring barrier categories that appeared independently across the technical, local-context, and adoption literatures: data maturity and interoperability; infrastructure and connectivity; regulatory frameworks and enforcement (including XAI); human capital and digital literacy; and economic viability and cost optimization. These five categories structure the readiness analysis presented in Section 4.3 and are each separately traceable to specific cited sources (see Table 3 in Section 6.3).

Third, each barrier category was translated into a corresponding intervention stage, and the five resulting stages were sequenced according to technical and institutional dependency rather than by the order in which the barriers were identified in the literature. Basic digitization and cyber-safety (Stage 1) were placed first because the literature indicates that without baseline platform reliability and merchant awareness, neither systematic data collection nor trust-building can proceed (Widayani, Fiernaningsih and Herijanto 2022; Gupta 2024). Data readiness and interoperability (Stage 2) precede algorithmic deployment because supervised and even most unsupervised techniques require some minimum of structured, aggregated transaction data, which the literature shows Nepal currently lacks (Pathak 2024; Financial Intelligence Unit, Nepal Rastra Bank 2024). Unsupervised detection and economic triage (Stage 3) is positioned before advanced model deployment because the absence of labeled fraud data forecloses supervised learning in the near term, while economic triage is required to keep false-positive costs proportionate to SME transaction values (Hilal, Gadsden and Yawney 2022; Vanini et al. 2023). AI implementation and XAI compliance (Stage 4) follows because the regulatory-readiness literature indicates that explainability cannot be retrofitted onto opaque models after deployment without undermining trust (Nicholls, Kuppa and Le-Khac 2021; Dhital 2025). Trust-building and ecosystem scaling (Stage 5) is positioned last because it is the outcome the preceding four stages are designed to produce, rather than a precondition for them, consistent with Inno-

vation Resistance Theorys account of trust as a cumulative product of repeated, transparent, low-harm interactions (Widayani, Fiernaningsih and Herijanto 2022; Rubio and Tulcanaza-Prieto 2025).

This derivation logic, mapping each stage to a literature-identified barrier and ordering stages by dependency rather than by narrative convenience, is the basis for the framework-validation procedures described in Section 6.

The literature evidence supporting each of these sequencing decisions is presented in full in Section 4; the present section establishes the process by which they were generated. Readers encountering the barrier citations above before reading Section 4 may treat them as forward references they are intended as a preview of the evidential basis that Section 4 elaborates.

Literature Review and Related Work

Nepals SME Digital Payment Ecosystem

Nepals transition to digital payments accelerated through pandemic-era avoidance of physical currency and NRB policies promoting QR-code and wallet interoperability (Tamang, Bhaskar and Chatterjee 2021). Adoption among individual users has been strong, but SMEs continue to face structural inefficiencies. The absence of a universal switch connecting financial institutions and payment service providers (PSPs) prevents data interoperability across the ecosystem (Pathak 2024), and some merchants continue to rely on physical cash because of unreliable power and internet connectivity, particularly outside major urban centers (Anakpo, Xhate and Mishi 2023; Klapper et al. 2025).

Global AI Fraud Detection Capabilities and the Transferability Gap

International financial institutions rely on a layered set of AI techniques: supervised models trained on historical fraud labels, unsupervised anomaly detectors such as Autoencoders and Isolation Forests for previously unseen fraud patterns, Graph Neural Networks for mapping fraud-ring relationships, behavioral analytics drawing on device and session data, and economic triage models that adjust detection thresholds to transaction value (Hilal, Gadsden and Yawney 2022; Hernandez Aros et al. 2024; Nicholls, Kuppa and Le-Khac 2021; Vanini et al. 2023). Increasingly, these systems are paired with Explainable AI methods, such as LIME and SHAP, to satisfy regulatory demands for transparency (Gupta 2024; Hilal, Gadsden and Yawney 2022).

These capabilities, however, assume conditions, mature labeled data, low-latency cloud infrastructure, advanced XAI enforcement, and a trained data-science workforce, that do not hold in Nepal. The transferability gap, defined here as the difference between the technical assumptions of global AI fraud detection and the infrastructural, regulatory, and behavioral realities of developing-country payment systems, manifests along five dimensions: data cohesion; connectivity and computing capacity; regulatory enforcement of explainability; cybersecurity awareness; and tolerance for false-positive cost. Table 1 maps each of these dimensions against Nepals current reality.

Readiness Analysis: Mapping AI Capabilities Against Nepals Reality

Table 1 summarizes the mismatch between global AI requirements and Nepals payment-ecosystem realities across the five barrier dimensions identified through the analytical process described in Section 3.4.

Table 1: Readiness Gap Analysis

Dimensions	Global AI Requirement	Nepal's Current Reality	Key Sources
Data Maturity & Interoperability	More than 100 behavioral features per session, centrally aggregated, well-labeled transaction histories	Fragmented data across non-interoperable PSPs; reactive detection triggered only by victim reports; no centralized record of recurring fraud typologies such as Re. 1 test-debit probing	Vanini et al. (2023); Pathak (2024); Financial Intelligence Unit, Nepal Rastra Bank (2024)
Infrastructure & Connectivity	Low-latency cloud computing and API integration for real-time scoring	Urban connectivity has improved, but semi-urban and rural areas lack reliable power and internet, limiting real-time deployment	Hilal et al. (2022); Tamang et al. (2021); Klapper et al. (2025)
Regulatory Frameworks & Enforcement	Enforceable XAI mandates, algorithmic auditability, privacy-by-design rules	Strong legislative intent (regulatory sandbox, draft AI guideline, directive amendments) but limited enforcement capacity for XAI and KYC violations	Nicholls et al. (2021); Dhital (2025); The Kathmandu Post (2025; 2025b); Nepal News (2026)
Human Capital & Digital Literacy	Skilled data-science talent; forensic capacity to validate flagged cases	Workforce skill gaps compounded by widespread cybersecurity-awareness gaps among merchants, raising susceptibility to social engineering	Vanini et al. (2023); Gupta (2024); Widayani et al. (2022)
Economic Viability & Cost Optimization	Cost-optimized thresholds balancing investigation expense against fraud loss, sustainable at enterprise scale	Thin SME margins make high false-positive rates economically punishing and likely to drive reversion to cash	Vanini et al. (2023); Irianto & Chanvarasuth (2025); Klapper et al. (2025); Anakpo et al. (2023)

Two implications follow from this mapping. First, the dimensions are not independent: data fragmentation (dimension 1) is itself partly a consequence of weak enforcement (dimension 3), since no directive currently compels PSPs to share standardized transaction logs. Second, the dimension most resistant to short-term intervention, human capital and digital literacy, is also the one most directly responsible for the social-engineering fraud patterns described in Section 4.1, which suggests that technical fixes alone, however sophisticated, cannot resolve the trust deficit without parallel investment in merchant education.

Related Work: Positioning This Study Among Existing Literature Streams

Table 2 positions this study against five distinguishable literature streams that, between them, cover most of the ground this paper draws on.

Table 2: Positioning Relative to Related Work

Literature Stream	Representative Studies	Focus	Limitation Relative to This Study
Global technical AI fraud-detection literature	Hilal et al. (2022); Hernandez Aros et al. (2024); Nicholls et al. (2021); Vanini et al. (2023)	Algorithmic capability and accuracy (supervised, unsupervised, GNN, XAI) in data-rich settings	Assumes data and infrastructure maturity; does not address transferability to low-resource markets
Nepal-specific cyberfraud and cybersecurity literature	Dhital (2025); Financial Intelligence Unit, Nepal Rastra Bank (2024); Gupta (2024)	Documents the scale and typology of cyber-enabled fraud in Nepal	Recommends reactive and awareness-based measures; does not propose an algorithmic or phased adoption pathway
SME digital-trust and adoption-barrier literature	Widayani et al. (2022); Rubio & Tulcanaza-Prieto (2025); Irianto & Chanvarasuth (2025)	Psychological and behavioral barriers to digital-payment adoption	Does not examine how algorithmic opacity specifically compounds these barriers
TOE-based AI-adoption literature	Badghish & Soomro (2024)	Identifies organizational, technological, and environmental determinants of general AI adoption by SMEs	Explains why SMEs adopt AI broadly, but offers no sequencing logic for fraud-specific, trust-sensitive deployment
Low-resource hybrid fraud-detection frameworks	Ampumuza et al. (2026)	Systematic review proposing a hybrid AI–audit model for fraud detection in resource-constrained cooperative finance (Uganda SACCOs)	Proposes a technical hybrid model but does not stage deployment by infrastructural/regulatory readiness or treat XAI as a gating compliance mechanism

Taken together, these five literature streams converge on, but do not resolve, the problem motivating this study. They establish that the technical capability exists, that Nepals fraud problem exists, that trust barriers exist, and that general AI adoption can be explained through TOE-type determinants; none of them, however, proposes a sequencing logic that connects these strands into a deployment pathway oriented specifically toward fraud detection and toward defusing, rather than risking, SME trust through staged use of explainability and economic triage.

The Research Gap

Synthesizing Sections 4.1 to 4.4 surfaces a gap that no single literature stream addresses on its own. The global technical literature (Hilal, Gadsden and Yawney 2022; Vanini et al. 2023; Hernandez Aros et al. 2024) consistently recommends complex supervised algorithms that require large volumes of labeled data and substantial cloud infrastructure, conditions absent in developing countries such as Nepal; several of these studies explicitly call for further research into unsupervised algorithms suited to data-scarce environments. The Nepal-specific literature (Dhital 2025; Financial Intelligence Unit, Nepal Rastra Bank 2024; Gupta 2024) documents the scale of the problem convincingly but stops at reactive recommendations, faster reporting and awareness campaigns, without exploring proactive, algorithmic alternatives. The SME trust and adoption literature (Widayani, Fiernaningsih and Herijanto 2022; Rubio and Tulcanaza-Prieto 2025; Irianto and Chanvarasuth 2025) identifies real psychological barriers to digital-payment adoption but does not examine how introducing opaque, black-box AI models could deepen exactly the trust deficit these studies describe. Closer to a technical solution, TOE-based adoption studies (Badghish and Soomro 2024) and low-resource hybrid detection frameworks (Ampumuza, Katushabe and Tamale 2026) come nearer to this studys territory, the first by modeling why SMEs adopt AI, the second by proposing a technical model for fraud detection under resource constraints, but neither sequences deployment according to infrastructural and regulatory readiness, nor treats explainability and trust as the organizing logic of that sequence.

The gap that remains, once all five streams are read together, is the absence of any framework connecting global AI capability to Nepals infrastructural, regulatory, and behavioral reality in a way that gives policymakers, fintechs, and financial institutions a concrete, ordered pathway from reactive fraud management to proactive, trust-preserving AI adoption. This paper addresses that gap through the 5-Stage Phased Readiness Framework. Following Hilal et al.s (2022) call to study unsupervised techniques in previously data-scarce settings, the framework offers a definite roadmap that begins with data preparedness and culminates in XAI-compliant deployment, tailored specifically to Nepals digital-payment context.

Proposed Conceptual Readiness Framework

The readiness gap synthesized in Section 4.5 reveals that Nepal's digital payment ecosystem cannot support the direct deployment of the AI fraud detection capabilities described in Section 4.2. The five barrier dimensions identified in Table 1 data fragmentation, infrastructure constraints, weak XAI enforcement, digital-literacy deficits, and thin SME margins are not isolated problems; they are interdependent preconditions that must be addressed in a specific order before advanced algorithmic systems can function without generating the false positives and trust erosion that would undermine adoption. The following framework proposes that sequence.

The 5-Stage Phased Readiness Model

Addressing the gap mentioned above involves abandoning a "plug and play" paradigm in AI implementation. Rather, this paper proposes a 5-stage phased readiness model specific to the digital payment environment in Nepal. The framework represents the key theoretical contribution of this research paper, as each stage builds upon the previous one (Figure 1).

- **Stage 1: Basic Digitization & Cyber-Safety**
  Before leveraging AI, it is imperative to lay the foundation and ensure that the baseline is secured. This will involve ensuring increased tech support in rural regions and the creation of platform-specific awareness campaigns on social engineering and QR code manipulation. At this step, trust will be gained via the basics such as reliability and training.
- **Stage 2: Data Readiness & Interoperability**
  National data formatting requirements must be agreed upon by the regulators and PSPs. Data interoperability will lead to the creation of a platform where localized fraud typologies can be aggregated to be used for training supervised models.
- **Stage 3: Unsupervised Detection & Economic Triage**
  Due to the current lack of labels, PSPs should utilize an unsupervised anomaly detection system such as Isolation Forests. By adjusting the detection threshold according to the value of the transaction, unnecessary costs related to the processing of false positives in SME transactions can be avoided.
- **Stage 4: AI Implementation & XAI Compliance:**
  With adequate levels of data maturity attained, the organizations can experiment with cutting-edge technologies, like Graph Neural Networks for identifying money mule associations. It is essential that these experiments be carried out in the NRB's regulatory sandbox and in accordance with Explainable AI (XAI) methodologies. In the case that the transaction of an SME is rejected, an explanation must be provided to avoid causing psychological harm due to a "black box" outcome.
- **Stage 5: Building Trust & Scaling the Ecosystem**
  The last stage shifts from protecting from fraud to actively building trust. Given that the use of AI technologies proves effective in reducing financial loss and speeding dispute resolution, confidence in these technologies grows. Increased trust in turn leads to financial inclusivity and transition from the cash-based informal economy.

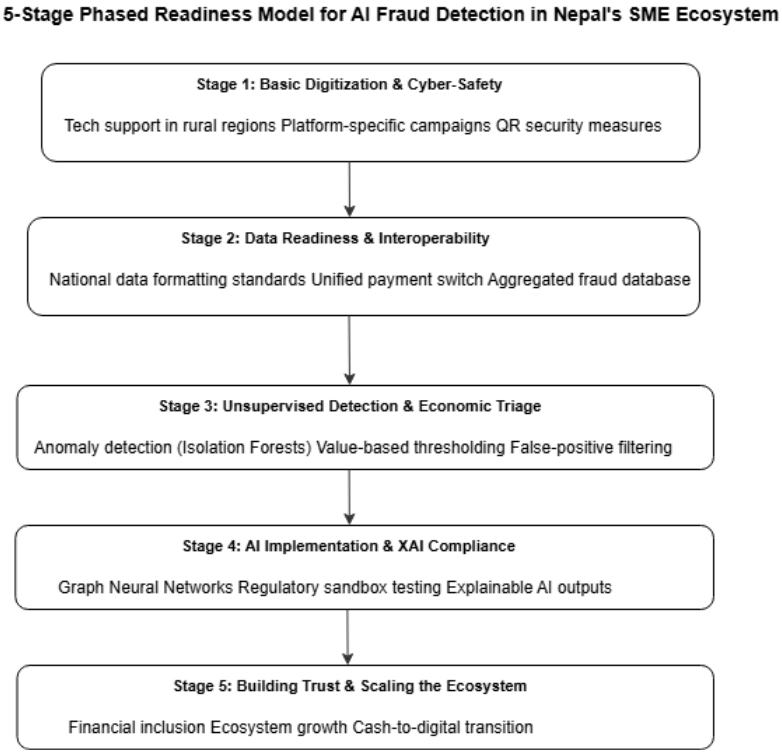

Figure 1: 5-Stage Phased Readiness Model for AI Fraud Detection in Nepal's SME Digital-Payment Ecosystem

Framework Validation

Validation Approach for Conceptual Frameworks

Frameworks of this kind are most rigorously validated through structured expert-consensus methods such as the Delphi technique, in which a panel of subject-matter experts iteratively reviews and refines a proposed model until consensus is reached (J. Skulmoski, T. Hartman and Krahn 2007). Recent examples in adjacent conceptual-framework research illustrate the approach: Manzini et al. (2026) validated a systems-based organizational-resilience framework through a two-round Delphi process with a multidisciplinary expert panel, reaching high consensus on the frameworks structure before recommending it for practical use. Because the present study is explicitly conceptual, built on literature synthesis rather than primary data collection, a full Delphi panel falls outside its current scope; conducting one is identified in Section 8 as the appropriate next methodological step before any pilot implementation. Within the bounds of a conceptual study, however, validity can and should still be demonstrated through methods that interrogate the frameworks grounding and positioning rather than its real-world performance. This study uses three such methods: theoretical grounding, gap-traceability mapping, and comparative benchmarking against existing frameworks.

Theoretical Grounding

The framework's internal validity is grounded in the two theoretical lenses underpinning this study. Rather than applying competing theories to cross-check findings, which triangulation proper requires, this study uses IRT and STST as complementary analytical anchors that together account for all five stages at different levels of analysis. Innovation Resistance Theory accounts for Stage 1 (basic digitization and cyber-safety) and Stage 5 (trust-building), since the theory holds that psychological barriers, fear, distrust, and perceived loss of control, are addressed first by establishing reliability and only later resolved through demonstrated, repeated benefit (Widayani, Fiernaningsih and Herijanto 2022). Socio-Technical Systems Theory accounts for Stages 2 through 4, since it holds that technical capability (data infrastructure, algorithms, XAI tooling) must be matched to social and organizational context, here, regulatory enforcement capacity and merchant-level digital literacy, before deployment can succeed. No stage in the framework relies on a theoretical assumption that is not already established in Section 2, supporting the frameworks internal theoretical coherence.

Gap-Traceability Validation

A second test of validity is whether each stage can be traced transparently back to a specific, cited deficiency identified in the literature review, rather than being introduced without evidentiary grounding. Table 3 demonstrates this traceability.

Table 3: Stage-to-Gap Traceability

Framework Stage	Literature-Identified Barrier Addressed	Supporting Sources
Stage 1: Basic Digitization & Cyber-Safety	Human-capital and digital-literacy gap; social-engineering susceptibility	Widayani et al. (2022); Gupta (2024)
Stage 2: Data Readiness & Interoperability	Data fragmentation; absence of a universal payment switch	Pathak (2024); Financial Intelligence Unit, Nepal Rastra Bank (2024)
Stage 3: Unsupervised Detection & Economic Triage	Absence of labeled fraud data; thin SME margins and false-positive intolerance	Hilal et al. (2022); Vanini et al. (2023); Irianto & Chanvarasuth (2025)
Stage 4: AI Implementation & XAI Compliance	Weak enforcement of explainability and auditability mandates	Nicholls et al. (2021); Dhital (2025); The Kathmandu Post (2025; 2025b)
Stage 5: Building Trust & Scaling	Cumulative trust deficit; reversion to cash	Rubio & Tulcanaza-Prieto (2025); Widayani et al. (2022)

Each stage maps to at least one literature-identified barrier, with no stage left unsupported by the review and no major barrier identified in Section 4.3 left unaddressed by any stage, indicating that the frameworks scope is neither under- nor over-specified relative to the reviewed evidence.

Comparative Benchmarking Against Existing Frameworks

A third validation step benchmarks the framework's structure against three classes of existing models: generic AI-maturity models used in enterprise settings, TOE-based AI-adoption models, and low-resource hybrid fraud-detection frameworks. This comparison evaluates whether the proposed framework occupies a distinct, non-redundant position in the landscape of existing readiness and maturity models, a recognized technique for establishing the validity of a new conceptual contribution (Torraco 2016).

Generic AI maturity models, such as those published by Gartner (2026) and MIT Sloan Management Review (2026), are designed to benchmark organization-wide AI capability across a firm's entire technology portfolio. They are sector-agnostic, assume an enterprise IT baseline that is largely absent in Nepal's SME sector, and treat governance as one maturity dimension among several rather than as a sequenced precondition. They do not address infrastructure-poor deployment contexts, nor do they model trust or explainability as terminal outcomes.

TOE-based AI adoption models, represented here by Badghish and Soomro (2024), explain the organizational, technological, and environmental determinants of AI adoption decisions by SMEs. Their contribution is explanatory rather than prescriptive they identify why adoption occurs but offer no sequencing logic for the order in which technical preconditions should be satisfied, particularly in fraud-sensitive, trust-critical deployment scenarios.

Low-resource hybrid fraud-detection frameworks, such as Ampumuza, Katushabe and Tamale (2026), address resource-constrained contexts directly by proposing a hybrid AIaudit model for fraud detection in cooperative finance. However, this framework proposes a technical architecture without staging its deployment according to infrastructural or regulatory readiness, and does not treat XAI compliance as a gating mechanism that must precede advanced model deployment. Applying such a framework to Nepal's context without first completing Stages 1 and 2 of the proposed model basic digitization and data interoperability would result in deploying detection algorithms against fragmented, non-standardized transaction data, producing the high false-positive rates that Section 4.3 identifies as the primary trust-destruction mechanism for thin-margin SMEs.

Taken together, this comparison confirms that the proposed 5-Stage Phased Readiness Framework occupies a distinct position in the existing landscape: it is the only model among those reviewed that simultaneously sequences deployment by infrastructural and regulatory readiness, treats economic triage and XAI compliance as ordered preconditions rather than optional features, and positions trust as the terminal output of the entire deployment sequence rather than a background assumption. A full comparative analysis across these dimensions is presented in Table 4 in Section 7.1.

Results and Discussion

Comparative Analysis with Existing Frameworks

Table 4 compares the proposed 5-Stage Readiness Framework with three classes of existing models that address related, but distinct, problems.

Table 4: Comparative Analysis of Readiness and Adoption Frameworks

Feature	**This Study (5-Stage Readiness Framework)**	**Generic AI Maturity Models** *(Gartner, 2026; MIT Sloan Management Review, 2026)*	**TOE-Based AI Adoption Models** *(Badghish & Soomro, 2024)*	**Low-Resource Hybrid Fraud Frameworks** *(Ampumuza et al., 2026)*
Primary purpose	Sequence fraud-detection deployment by infrastructural and regulatory readiness	Benchmark organization-wide AI capability	Explain determinants of AI-adoption decisions	Propose a technical hybrid detection model
Sector specificity	Fraud detection in SME digital payments	Sector-agnostic	Sector-agnostic (general AI)	Cooperative and low-resource finance
Sensitivity to infrastructure-poor settings	Explicit; the central organizing logic	Largely absent; assumes an enterprise IT baseline	Partial; environment is one of three TOE pillars, not central	Present, but framed around audit/technical hybridity rather than staged readiness
Treatment of trust and explainability	XAI compliance is a sequenced gate; trust is the terminal outcome	Governance is one maturity dimension among several	Not directly modeled	Not addressed
Validation status	Theoretical grounding, gap-traceability mapping, and comparative benchmarking (this paper); Delphi panel recommended as next step	Industry-developed, practitioner-validated	Empirically tested via SME survey data	Systematic synthesis

Novelty and Advantages

The proposed frameworks distinguishing contribution is that it treats readiness as a sequence rather than a score, and treats trust as the frameworks terminal output rather than a background condition. Unlike generic AI maturity models, which assume an enterprise IT baseline largely absent in Nepals SME sector, or TOE-based adoption models, which explain why adoption happens but not in what technical order it should happen, the 5-Stage framework provides ordered, actionable guidance for a fraud-detection use case in which false positives carry outsized reputational and trust costs for thin-margin SMEs. Relative to low-resource hybrid detection frameworks such as Ampumuza, Katushabe and Tamale (2026), the present frameworks distinguishing feature is its explicit staging logic, technical capability is deliberately withheld until the infrastructural and regulatory preconditions for its responsible use are in place, rather than proposed independently of readiness.

Implications

Theoretical Implications

This study extends Innovation Resistance Theory and Socio-Technical Systems Theory by demonstrating how algorithmic opacity and infrastructural fragmentation jointly suppress AI adoption in emerging markets. It introduces the transferability gap as a conceptual lens for evaluating technology-deployment mismatches, and demonstrates, through Sections 6 and 7.1, how a conceptual framework of this kind can be validated and positioned without primary data collection.

Physical Implications

Policymakers can use the framework to prioritize data standardization, sandbox testing, and XAI mandates before incentivizing advanced AI procurement. Fintech developers should focus on lightweight unsupervised models and explainability modules rather than resource-intensive deep-learning architectures. Financial institutions can implement economic triage to optimize fraud-investigation costs while preserving SME operational continuity.

Policy Implications

NRB and other regulatory bodies should institutionalize data-sharing standards, mandate algorithmic transparency for financial AI, and expand cybersecurity-literacy programs targeted at merchant populations. International development partners can support connectivity expansion and cloud-infrastructure subsidies to reduce the digital divide. Notably, NRBs recent Regulatory Sandbox and draft AI guideline (Nepal News 2026; The Kathmandu Post 2025a) provide exactly the kind of supervised-testing mechanism Stage 4 anticipates, suggesting that the institutional groundwork for the frameworks later stages is beginning to emerge even as this framework was being developed and revised.

Limitations and Future Research

This study has several limitations that should be acknowledged.

- **Conceptual Nature:** The research relies on integrative literature synthesis rather than empirical data collection, algorithmic testing, or pilot implementation. Consequently, the proposed 5-Stage Phased Readiness Framework has not

undergone real-world validation within Nepals SME digital-payment ecosystem.

- **Validation Gap:** Section 6 validates the framework conceptually, through theoretical grounding, gap-traceability mapping, and comparative benchmarking, but a structured Delphi panel of regulators, fintech practitioners, and SME representatives (J. Skulmoski, T. Hartman and Krahn 2007) remains the recommended next step to establish expert consensus on the frameworks content validity before any pilot implementation.
- **Context Specificity:** The framework is tailored to Nepals particular infrastructural, regulatory, and behavioral context; direct transferability to other emerging economies requires further comparative analysis.
- **Data Constraints:** The analysis depends on publicly available institutional reports and peer-reviewed literature; proprietary transaction data from Nepali PSPs or banks was not accessible for model-training simulations.
- **Future Research Directions:** To address these limitations, future work should prioritize the following.
  - **Expert Panel Validation:** Convene a structured Delphi panel of NRB representatives, fintech developers, and SME merchants to formally validate the frameworks content and structure ahead of pilot work.
  - **Pilot Testing:** Collaborate with three to five Nepali payment service providers to evaluate Stage 3 unsupervised-detection performance on anonymized real-world transaction data.
  - **Stakeholder Validation:** Conduct semi-structured interviews with policymakers, fintech developers, and SME merchants to assess perceived usability, trust implications, and adoption barriers for each framework stage.
  - **Simulation Modeling:** Develop agent-based or economic simulations to quantify the impact of false-positive reduction via value-based thresholding (Stage 3) and XAI-compliant explanations (Stage 4).
  - **Cross-Country Comparison:** Extend the transferability-gap analysis to similar emerging economies (for example, Bangladesh, Sri Lanka, Kenya) to refine the frameworks generalizability.

Conclusion

The rapid digitization of Nepals financial sector has outpaced the development of defensive security infrastructure, creating a pronounced trust deficit among SMEs exposed to localized cyber-enabled fraud. While global financial institutions successfully leverage AI for real-time fraud detection, direct transplantation into Nepals fragmented, infrastructure-constrained, and trust-sensitive environment remains unviable.

Based on a structured integrative literature review, this study identifies a critical transferability gap and proposes a five-stage conceptual readiness framework, derived transparently from literature-identified barriers and validated through theoretical grounding, gap-traceability mapping, and comparative benchmarking against existing AI-adoption and maturity models. Rather than prioritizing technological implementation, the model emphasizes ecosystem maturation. The findings demonstrate that data interoperability, economic triage, and Explainable AI (XAI) compliance must precede advanced machine learning deployment to preserve merchant confidence and prevent operational disruption.

Returning to the three research questions that guided this study: RQ1 is addressed in Section 4.2, which establishes that global AI fraud detection relies on supervised models, unsupervised anomaly detectors, GNNs, and XAI frameworks, all of which assume data maturity, cloud infrastructure, and regulatory enforcement absent in Nepal. RQ2 is addressed through the five barrier dimensions in Table 1, which show that data fragmentation, infrastructure gaps, weak XAI enforcement, digital-literacy deficits, and thin SME margins collectively prevent direct transplantation of these capabilities. RQ3 is addressed by the derivation process in Section 3.4 and the three-method validation in Section 6, which together demonstrate that a phased sequence beginning with data interoperability and culminating in XAI-compliant deployment is both technically feasible and trust-preserving for Nepal's SME context.

These insights carry direct implications for stakeholders. For the Nepal Rastra Bank and regulatory bodies, the immediate priority should be institutionalizing data-sharing standards and expanding regulatory sandboxes for XAI-compliant model testing. FinTech developers and commercial banks should focus on lightweight unsupervised anomaly detection and value-based transaction thresholding, rather than deploying resource-intensive algorithms that exceed current infrastructural capacity.

Ultimately, artificial intelligence cannot function as a standalone solution to Nepals digital fraud challenge. Its effectiveness depends entirely upon sequential infrastructure development, regulatory clarity, and sustained institutional trust.

Disclosure Statement

The authors have no financial or non-financial disclosures to share for this article.

Ethical Approval: Not applicable.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Data Availability Statement: This research used already existing data from different sources. All of them have been cited as well.

Use of Artificial Intelligence (AI) Tools: This research utilized AI to improve the language as well as finding minor grammatical errors. The ideas, analysis, and argument are the authors' own.

Authors' Contributions: Yogesh Pant contributed to conceptualization, literature synthesis, framework development, manuscript drafting, and data analysis. Aditya Pudasaini contributed to methodology design, literature review, critical revision, and formatting compliance. Roshan Shrestha contributed to academic supervision, scope refinement, and editorial guidance. Alish K.C. contributed to research validation, theoretical framing, and final manuscript review. All authors read and approved the final version of the manuscript.

Funding: No funding was received.

Conflict of Interest: All authors have no competing interests.

Acknowledgement

The authors thank Islington College for institutional support and academic resources provided during the preparation of this manuscript.

References

Ampumuza, Dalton, Calorine Katushabe and Micheal Tamale (Jan. 2026). "A systematic review and future directions for AI-driven detection of fraud patterns in SACCO transactions." In: *Frontiers in Artificial Intelligence* 8. ISSN: 2624-8212. 10.3389/frai.2025.1690482. http://dx.doi.org/10.3389/frai.2025.1690482.

Anakpo, Godfred, Zizipho Xhate and Syden Mishi (2023). "The Policies, Practices, and Challenges of Digital Financial Inclusion for Sustainable Development: The Case of the Developing Economy." In: *FinTech* 2.2, pp. 327–343. ISSN: 2674-1032. 10.3390/fintech2020019. http://dx.doi.org/10.3390/fintech2020019.

Badghish, Saeed and Yasir Ali Soomro (Feb. 2024). "Artificial Intelligence Adoption by SMEs to Achieve Sustainable Business Performance: Application of TechnologyOrganizationEnvironment Framework." In: *Sustainability* 16.5, p. 1864. ISSN: 2071-1050. 10.3390/su16051864. http://dx.doi.org/10.3390/su16051864.

Crisanto, Juan Carlos, Cris Benson Leuterio, Jermy Prenio and Jeffery Yong (Dec. 2024). *Regulating AI in the Financial Sector: Recent Developments and Main Challenges*. FSI Insights on Policy Implementation 63. Basel: Bank for International Settlements, Financial Stability Institute. https://www.bis.org/fsi/publ/insights63.pdf.

Dahal, Rewan Kumar, Ganesh Bhattarai and Dipendra Karki (Mar. 2021). "Management Accounting Practices on Organizational Performance Mediated by Rationalized Managerial Decisions." In: *International Research Journal of Management Science* 5.1, pp. 148–167. ISSN: 2542-2510. 10.3126/irjms.v5i1.35870. http://dx.doi.org/10.3126/irjms.v5i1.35870.

Dhital, Hemant (2025). "The evolving landscape of Cybercrime in Nepal: A multi-Year Analysis of Platform Specific Trends and Victim Demographics (2077-2082 B.S. / 2020-2025 A.D.)" In: *Applied Data Science and Analysis* 2025, pp. 165–177. ISSN: 3005-317X. 10.58496/adsa/2025/014. http://dx.doi.org/10.58496/ADSA/2025/014.

Financial Intelligence Unit, Nepal Rastra Bank (2024). *Strategic Analysis Report 2024: Cyber Enabled Frauds*. Tech. rep. Kathmandu: Nepal Rastra Bank. https://www.nrb.org.np/contents/uploads/2024/11/FIU-Nepal-Strategic-Analysis-Report-2024.pdf.pdf.

Gartner (2026). *AI Maturity Model and AI Roadmap Toolkit*. Online. Accessed 28 June 2026. https://www.gartner.com/en/chief-information-officer/research/ai-maturity-model-toolkit.

Gupta, Lokesh (2024). "Issues of Cyber security and its solutions in Nepalese Context." In: *NPRC Journal of Multidisciplinary Research* 1.2, pp. 122–127. ISSN: 3059-9148. 10.3126/nprcjmr.v1i2.69333. http://dx.doi.org/10.3126/nprcjmr.v1i2.69333.

Hernandez Aros, Ludivia, Luisa Ximena Bustamante Molano, Fernando Gutierrez-Portela, John Johver Moreno Hernandez and Mario Samuel Rodríguez Barrero (2024). "Financial fraud detection through the application of machine learning techniques: a literature review." In: *Humanities and Social Sciences Communications* 11.1. ISSN: 2662-9992. 10.1057/s41599-024-03606-0. http://dx.doi.org/10.1057/s41599-024-03606-0.

Hilal, Waleed, S. Andrew Gadsden and John Yawney (May 2022). "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances." In: *Expert Systems with Applications* 193, p. 116429. ISSN: 0957-4174. 10.1016/j.eswa.2021.116429. http://dx.doi.org/10.1016/j.eswa.2021.116429.

Irianto, Aloysius Bagas Pradipta and Pisit Chanvarasuth (May 2025). "Drivers and Barriers of Mobile Payment Adoption Among MSMEs: Insights from Indonesia." In: *Journal of Risk and Financial Management* 18.5, p. 251. ISSN: 1911-8074. 10.3390/jrfm18050251. http://dx.doi.org/10.3390/jrfm18050251.

J. Skulmoski, Gregory, Francis T. Hartman and Jennifer Krahn (2007). "The Delphi Method for Graduate Research." In: *Journal of Information Technology Education: Research* 6, pp. 001–021. ISSN: 1539-3585. 10.28945/199. http://dx.doi.org/10.28945/199.

Karki, Dipendra (2012). *Economic impact of tourism in Nepal's economy using cointegration and error correction model*. en. 10.13140/RG.2.1.4839.5684. https://www.researchgate.net/doi/10.13140/RG.2.1.4839.5684.

Klapper, Leora, Dorothe Singer, Laura Starita and Alexandra Norris (2025). *The Global Findex Database 2025: Connectivity and Financial Inclusion in the Digital Economy*. Washington, DC: World Bank. ISBN: 9781464822049. 10.1596/978-1-4648-2204-9. http://dx.doi.org/10.1596/978-1-4648-2204-9.

Manzini, Dumisani, Rudolph Oosthuizen, Hilda Chikwanda and Rina Peach (Apr. 2026). "SystemsBased Organisational Resilience Framework: A Delphi StudyBased Validation and Verification." In: *Systems Research and Behavioral Science* 43.4, pp. 1615–1644. ISSN: 1099-1743. 10.1002/sres.70061. http://dx.doi.org/10.1002/sres.70061.

MIT Sloan Management Review (2026). *What's Your Company's AI Maturity Level?* Online. Accessed 28 June 2026. https://mitsloan.mit.edu/ideas-made-to-matter/whats-your-companys-ai-maturity-level.

Nepal News (2026). *Everything You Need to Know about NRB's New Regulatory Sandbox*. Online. Accessed 28 June 2026. https://english.nepalnews.com/s/business/everything-you-need-to-know-about-nrbs-new-regulatory-sandbox/.

Nepal Rastra Bank (2024). *Annual Report Fiscal Year 2023/24*. Tech. rep. Kathmandu: Nepal Rastra Bank. https://www.nrb.org.np/contents/uploads/2025/07/Annual-Report-2023-24-English.pdf.

Nicholls, Jack, Aditya Kuppa and Nhien-An Le-Khac (2021). "Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape." In: *IEEE Access* 9, pp. 163965–163986. ISSN: 2169-3536. 10.1109/access.2021.3134076. http://dx.doi.org/10.1109/ACCESS.2021.3134076.

Pathak, Arbind (Feb. 2024). "Digital Payment In Nepal: An Overview And Recommendations." In: *Rupandehi Campus Journal* 4.1, pp. 25–34. ISSN: 2976-1158. 10.3126/rcj.v4i1.62916. http://dx.doi.org/10.3126/rcj.v4i1.62916.

Rubio, Jeniffer and Ana Belén Tulcanaza-Prieto (May 2025). "Digital Payments Trust in Latin America and the Caribbean." In: *Economies* 13.5, p. 140. ISSN: 2227-7099. 10.3390/economies13050140.

Tamang, Abinash, Prem Kumar Bhaskar and Jyotir Moy Chatterjee (2021). "Acceleration of Digital Payment Adoption during COVID-19 Pandemic: A Case Study of Nepal." In: *LBEF Research Journal of Science, Technology and Management* 3.2, pp. 1–13.

The Kathmandu Post (Dec. 2025a). *Nepal Rastra Bank Drafts AI Guideline for Banks and Digital Payment Firms*. Online. Accessed 28 June 2026. https://kathmandupost.com/money/2025/12/12/nepal-rastra-bank-drafts-ai-guideline-for-banks-and-digital-payment-firms.

– (Mar. 2025b). *Nepal's Central Bank Amends Online Payment System Directives to Curb Fraud*. Online. Accessed 28 June 2026. https://kathmandupost.com/money/2025/03/07/nepal-s-central-bank-amends-online-payment-system-directives-to-curb-fraud.

Torraco, Richard J. (Oct. 2016). "Writing Integrative Literature Reviews: Using the Past and Present to Explore the Future." In: *Human Resource Development Review* 15.4, pp. 404–428. ISSN: 1552-6712. 10.1177/1534484316671606. http://dx.doi.org/10.1177/1534484316671606.

Vanini, Paolo, Sebastiano Rossi, Ermin Zvizdic and Thomas Domenig (Mar. 2023). "Online payment fraud: from anomaly detection to risk management." In: *Financial Innovation* 9.1. ISSN: 2199-4730. 10.1186/s40854-023-00470-w. http://dx.doi.org/10.1186/s40854-023-00470-w.

Whittemore, Robin and Kathleen Knafl (Nov. 2005). "The integrative review: updated methodology." In: *Journal of Advanced Nursing* 52.5, pp. 546–553. ISSN: 1365-2648. 10.1111/j.1365-2648.2005.03621.x. http://dx.doi.org/10.1111/j.1365-2648.2005.03621.x.

Widayani, Anna, Nilawati Fiernaningsih and Pudji Herijanto (Dec. 2022). "Barriers to digital payment adoption: micro, small and medium enterprises." In: *Management & Marketing* 17.4, pp. 528–542. ISSN: 2069-8887. 10.2478/mmcks-2022-0029. http://dx.doi.org/10.2478/mmcks-2022-0029.

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 12–22

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/hq2s5x30

Machine Learning Based Postpartum Depression Risk Prediction: A Case Study on a Bangladeshi Dataset with Transferability Discussions for Nepal

Aditi Adhikari[1,*], Nirupan Karki[2],

[1]Islington College, Kathmandu, Nepal
[2]Pulchowk Campus, Institute of Engineering, Tribhuvan University, Kathmandu, Nepal
[*]*Correspondence: adhikarisharmaaditi@gmail.com*

ARTICLE HISTORY

Received: 21 March 2026 Revised: 24 April 2026
Accepted: 23 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Adhikari, A. and Karki, N. (2026). 'Machine learning based postpartum depression risk prediction: A case study on a Bangladeshi dataset with transferability discussions for Nepal', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 12–22. Available at: https://doi.org/10.67556/hq2s5x30

ABSTRACT

Postpartum depression (PPD) is a common perinatal mood disorder that often goes undetected, affecting roughly 10-15 percent of mothers, with higher rates across the Middle East and Asia and about half of cases never reported. Awareness remains low in countries such as Nepal, making early detection difficult. This study used the dataset Data for Postpartum Depression Prediction in Bangladesh (Raisa and Kaiser 2025) to test whether supervised machine learning can sort mothers into three EPDS-based risk categories (Low, Medium, High), and how much of that signal depends on the dataset's own depression-screening features rather than on sociodemographic and psychosocial variables. Four classifiers (Logistic Regression, Random Forest, XGBoost, and a Soft Voting Ensemble) were compared using a leakage-free pipeline, repeated stratified 5-fold cross-validation, paired McNemar and DeLong significance tests, bootstrap confidence intervals, and permutation/SHAP-based interpretability analysis. Under cross-validation, the Voting Ensemble and XGBoost were statistically indistinguishable (mean accuracy 76.5% versus 77.4%, macro-average ROC-AUC 0.91-0.92 for both), both modestly ahead of Logistic Regression and Random Forest, with no significant pairwise difference except a single class-level AUC comparison. Feature-importance and SHAP analysis showed that the antenatal PHQ-9 score dominates every model's predictions; removing all PHQ-2/PHQ-9 features cost every model 17 to 19 accuracy points (Random Forest fell from 76% to 59%), pointing to substantial redundancy between the EPDS-based target and the PHQ-based predictors, both collected in the same interview. Mutual-information feature selection showed that 10-30 of the 87 encoded features recover most of the full model's performance, and SMOTE gave a small, consistent improvement on the minority Medium class. All models separated High- and Low-risk classes well (AUC > 0.93) but struggled with the Medium class (AUC 0.81-0.85), whose EPDS band overlaps the Low/High boundary. Gradient boosting and a soft-voting ensemble are competitive, well-calibrated screening aids under default hyperparameters, but a meaningful share of their accuracy reflects redundancy between two depression screens rather than novel signal from contextual risk factors, and transfer of these results to Nepal remains a hypothesis pending local validation.

Keywords: Postpartum depression, machine learning, ensemble learning, digital health, maternal mental health, measurement circularity, model interpretability

Introduction

Machine learning (ML) is the branch of artificial intelligence (AI) concerned with building systems that improve their performance on a task from data and experience, without being explicitly programmed for every case (Jordan and Mitchell 2015). ML is the cognitive engine of industry 4.0, whose recent growth has been driven both by new learning algorithms and by the increasing availability of data and low-cost computation (Jordan and Mitchell 2015; Rajbhandari et al. 2020).

ML is usually divided into supervised, unsupervised and reinforcement learning; in supervised learning, the kind used in this study, a model learns from labelled examples and is then used either for classification, where the output is a category, or for regression, where the output is a value (Russell and Norvig 2021; Karki et al. 2023).

Healthcare has become one of the largest areas of ML application, ranging from diagnosis and image segmentation to disease-risk prediction and clinical decision support (Habehh and Gohel 2021). It is worth asking directly why an ML approach is preferable here to a traditional statistical risk model, such as a single multinomial logistic regression fitted with clinical judgement about which interactions to include. The answer is nuanced: Logistic Regression remains a strong, interpretable baseline in this dataset, and under cross-validation it is not dramatically worse than the tree-based models. The case for the ML approaches used here is threefold: (i) Random Forest and XGBoost can capture non-linear interactions between risk factors without the analyst specifying those interactions in advance, which matters because no single demographic feature separates the risk classes well, so any separating signal likely comes from combinations of features; (ii) tree ensembles provide a natural, model-based feature-importance ranking, supporting this study's aim of an interpretable screening aid; and (iii) a soft-voting ensemble can average out the distinct error patterns of structurally different models. This is not a claim that ML is categorically superior to statistical modelling here. Once the results are in hand, since the conclusion turns out to be more measured than the headline numbers alone would suggest.

The same techniques have been applied to mental health, where models trained on patient data and behavioural information can flag conditions such as depression. Postpartum depression (PPD) is one condition where this kind of early, data-driven screening could make a difference. PPD typically emerges in the weeks after childbirth and is distinct from the brief, milder baby blues many new mothers experience; it is characterised by persistent low mood, anhedonia, and disturbances in sleep and appetite beyond what is typical for the postpartum period, and carries real risks for maternal wellbeing and child development if untreated (OHara and McCabe 2013).

The aim of this study is to predict PPD risk using the dataset Data for Postpartum Depression Prediction in Bangladesh, framed as a multi-class classification task. The motivation is local: a recent mixed-methods study at a Kathmandu maternity hospital reported a PPD prevalence of 12.24 percent, linked to weak family and spousal support, difficult physical recovery and cultural beliefs (Neupane et al. 2024). Many women in Nepal remain unaware of PPD, so cases are missed. By learning from sociodemographic, psychosocial, pregnancy-related and newborn-related factors, an ML model may help tell apart women at higher and lower risk and support earlier screening, with the important caveat, that part of that separating signal comes from a second depression-screening instrument bundled into the same dataset, not from contextual risk factors alone.

No single algorithm is best for every dataset. Linear models such as Logistic Regression are transparent and quick to train but can miss complex interactions, while tree-based methods such as Random Forest and boosting methods such as XGBoost capture non-linear patterns at some cost to interpretability. Ensemble methods that combine multiple classifiers' probability outputs can push performance further, though the cross-validated results show that gain to be smaller and less certain than a single train/test split alone would suggest. Comparing individual classifiers alongside a combined ensemble, under a validation procedure that can detect whether the differences are real, is the core of this work. The rest of the paper reviews related work, sets out the data and methods, reports and discusses the results, and concludes.

Literature and Related Work

Motherhood is one of the most significant transitions in a woman's life. It is physical, as the body recovers and changes, but it is also psychological and social, as a woman's sense of self shifts toward being a mother (Hwang, Choi and An 2022). Childbirth brings a mix of emotions, from joy to fear and anxiety. Many new mothers go through a short period of baby blues, but for some the symptoms are more severe and last longer, which is when PPD sets in (OHara and McCabe 2013). A large share of these cases never reaches a clinician. Fewer than half of mothers with depressive symptoms disclose them, so up to half of all PPD cases go unreported (Gopalakrishnan et al. 2022). PPD is a perinatal form of major depressive disorder; it affects roughly 10 to 15 percent of women each year in the United States, around 500,000 women (GUINTIVANO, MANUCK and MELTZER-BRODY 2018), and about 17 percent of healthy mothers worldwide, with the highest rates seen in the Middle East and Asia (Shorey et al. 2018). Closer to the present context, Neupane et al. (2024) found a prevalence of 12.24 percent among mothers at a Kathmandu maternity hospital and pointed to inadequate support and cultural pressures as contributing factors, which shows the problem is real and under-recognised in Nepal.

Machine Learning Approaches to PPD Prediction

Saqib, Khan and Butt (2021) carried out a scoping review of how well ML predicts PPD across data sources including clinical records, electronic health records and social media, covering both classification and regression. Across the many algorithms assessed (Logistic Regression, SVM, Decision Trees, Random Forests, Naive Bayes, Neural Networks, K-Nearest Neighbors and XGBoost), Random Forest came out strongest with an AUROC of up to 0.884, although SVM and boosting methods also did well.

Qi et al. (2025) developed several ML models to predict PPD risk in perinatal women from demographic, psychosocial, mental-health-history and physiological variables, using the EPDS result as the label. Logistic Regression and an artificial neural network gave the best results, with AUC values of roughly 0.801-0.858. Huang et al. (2025) took an explainability-focused approach and found XGBoost the best of eight algorithms, with an AUC of 0.955. Natarajan et al. (2017) worked with survey data combining demo-

graphic details and self-reported emotional symptoms, and found Functional Gradient Boosting and Soft-Margin Boosting performed best.

A few patterns run through this literature. Simpler linear models such as Logistic Regression and SVM tend to be interpretable but more modest in accuracy, whereas ensemble methods usually score higher. Random Forest is reliable across several studies, and boosting methods, especially XGBoost, often reach the top scores. The strongest predictors are usually psychosocial and prior-mental-health variables rather than demographic details on their own, a pattern this study's own feature-importance analysis reproduces closely. though with the added finding that when the prior-mental-health variable is itself another depression-screening score measured at the same time as the target, the resulting accuracy gain is partly an artefact of measurement redundancy rather than purely a clinical insight. Performance also depends heavily on data quality and class balance, since maternal datasets are often skewed toward lower-risk cases and contain a good deal of missing data, which tends to hurt recall on the minority classes. These observations shaped the choices made later in this study.

The Research Gap

PPD is not only a clinical issue but an economic one. When it goes untreated it lowers maternal productivity, raises healthcare costs and can affect a child's development, all of which weigh on the human capital that economies depend on (Karki 2012). In settings where specialist mental-health services are thin on the ground, putting low-cost, data-driven screening tools into routine maternal care is a practical way to widen coverage without a matching rise in cost. Seen this way, automated PPD risk prediction is an example of technology being used to strengthen social and economic resilience. Most existing ML studies, though, draw on cohorts from high-income or East Asian settings, and very little work targets South Asian populations where awareness and reporting are lowest.

Two points of scope deserve to be stated precisely. First, soft-voting ensembles that combine linear, bagging, and boosting base learners are a well-established technique in the wider PPD-prediction literature reviewed above and elsewhere in clinical ML more broadly; nothing about the ensembling method itself is new in this paper. Second, the dataset used here is drawn from a single country (Bangladesh) and a single data-collection effort, so the results describe a single Bangladeshi maternal cohort rather than South Asian women as a broader population. With those two points in mind, this study's contribution is to apply a rigorous model comparison (evaluated under repeated cross-validation with paired significance testing rather than a single split) to this dataset, with an explicit check for circularity between the target and the PHQ-based predictors, and to discuss, cautiously, what the results might imply for maternal screening in Nepal. This study addresses that gap: training and rigorously comparing three widely used classifiers together with a Voting Ensemble on a Bangladeshi maternal dataset, checking how much of their performance depends on redundant depression-screening features, and discussing what that might imply for a future Nepal-specific screening tool.

Methodology

Dataset Description

The dataset is Data for Postpartum Depression Prediction in Bangladesh by Raisa and Kaiser (2025), obtained from Mendeley Data. It was chosen because Bangladesh sits in South Asia and shares many social and cultural traits with Nepal, where awareness of PPD is still limited; it has been discussed why this similarity should not be taken for granted. The data is tabular, with 800 records and 51 columns spanning sociodemographic, economic, medical and obstetric, psychosocial, lifestyle, relationship, newborn-care and mental-health factors. No duplicate rows were found in the raw file.

Depression is measured in the dataset through both PHQ and EPDS scores and results. PHQ-2 and PHQ-9 are general depression-screening tools and were not designed for the postpartum period, whereas the Edinburgh Postnatal Depression Scale (EPDS) was built specifically to screen for emotional distress during and after pregnancy and has been validated across many cultures (Cox, Holden and Sagovsky 1987). Since the goal here is to judge whether a woman is at Low, Medium or High risk of PPD, the EPDS Result column was used as the target, making this a supervised multi-class classification task.

The EPDS Result thresholds are not documented in the dataset's variable descriptions, so they were recovered directly from the data by cross-tabulating EPDS Score against EPDS Result:

Table 1: EPDS Result Thresholds, Recovered Directly from the Data

EPDS Result	Min score	Max score	n (of 800)
Low	0	9	260
Medium	8	12	190
High	12	30	350

Note. The Low and Medium bands overlap at scores 8–9, and the Medium and High bands overlap at score 12.

Two things are worth flagging about these thresholds. First, the Low and Medium bands overlap at score 8-9, and the Medium and High bands overlap at score 12. The three-way split is not a clean partition of the raw score, and this likely contributes to every model's difficulty with the Medium class. Second, the dataset documentation (Raisa and Kaiser 2025) describes a single structured questionnaire, administered face-to-face or online, that collects socio-demographic information together with the PHQ-2, PHQ-9 and EPDS instruments in one sitting. The three depression-screening instruments are therefore not staged at meaningfully different points in the perinatal timeline; PHQ-9 and EPDS are, in effect, two contemporaneous self-report measures of overlapping constructs answered in the same interview. This motivates the circularity check: a model that uses PHQ-9 to predict EPDS risk band is,

to a substantial extent, using one depression screen to predict another, rather than inferring risk from independent contextual factors.

Algorithms

Four classifiers were selected. Logistic Regression was included for its simplicity and interpretability and because it suits a categorical target, returning probabilities that help flag higher-risk women. Random Forest was included because it copes well with complex interactions between features and resists overfitting while ranking feature importance. XGBoost was included for its accuracy and its ability to model non-linear relationships among features such as received support, worry about the newborn, and antenatal depression (PHQ-2). Finally, a Soft Voting Ensemble pooled the class probabilities of all three base learners, aiming to reduce the variance of any single model and improve classification of the borderline Medium risk class. All four models were run with default-adjacent hyperparameters; no tuning was performed, so any claim about which model is best in this paper should be read as best under default settings, not best achievable.

Tools

The work was done in Python within a Jupyter Notebook environment for its cell-by-cell execution and immediate output. The main libraries were Pandas and NumPy for data handling, scikit-learn for preprocessing, the Logistic Regression, Random Forest and Voting Classifier models, cross-validation, and evaluation, and XGBoost for gradient boosting on the multi-class target. Class-imbalance handling used imbalanced-learn (SMOTE) inside a leakage-safe pipeline, McNemar's test used statsmodels, the DeLong test's p-values used SciPy, and model interpretability used SHAP.

Data Preprocessing

Duplicate rows were checked first: none were found among the 800 records, so no rows were dropped on that basis. Numeric outliers were inspected during exploratory analysis but not removed, since the extreme values (e.g. maternal age in the low 40s, PHQ-9 scores at the ceiling of the scale) are clinically plausible rather than data-entry errors, and removing them would bias the sample against exactly the higher-risk mothers a screening tool needs to detect.

A case-normalisation check on the categorical columns found that several categories were duplicated purely by inconsistent capitalisation. For example, Education Level contained both Primary School and Primary school as distinct string values, as did Husband's education level, and Total children contained both More than two and More than Two. Left uncorrected, one-hot encoding would have split each of these into two separate dummy columns, diluting the signal from what is really a single category and inflating the feature count. All categorical text was normalised to a single canonical spelling per category before any further preprocessing.

Missing values were dealt with next. Columns with very high proportions of missing data were dropped. This choice is practical but not free of cost, and that cost should be stated plainly rather than left implicit: Current monthly income, Monthly income before latest pregnancy, Addiction, Disease before pregnancy, and Diseases during pregnancy are all plausible, literature-supported predictors of PPD risk. Dropping them because more than half of respondents left them blank, rather than because they are uninformative, likely removes real signal from the model and is a genuine limitation of this analysis, not a neutral cleaning step.

Table 2: Columns Dropped for High Missingness (all >20% Missing), Plus the Identifier Column

Column	Missing (n)	Missing (%)
Addiction	789	98.6
History of pregnancy loss	613	76.6
Disease before pregnancy	588	73.5
Current monthly income	525	65.6
Age of immediate older children	517	64.6
Monthly income before latest pregnancy	437	54.6
Diseases during pregnancy	371	46.4
Feeling for regular activities	223	27.9
Need for Support	167	20.9

Columns with less than 5 percent missing values (Education Level, Husband's education level, Husband's monthly income, Abuse, and Trust and share feelings) were filled using the mode of each column. Mode imputation was kept as the simplest defensible choice here: all five columns are ordinal or nominal categorical variables with a small number of levels and a low missingness rate, so the risk of the mode systematically distorting the class distribution is small, and a more elaborate approach (e.g. k-nearest-neighbours imputation on the one-hot-encoded space) would add complexity for a handful of affected rows (at most 38 of 800, for Abuse) without a clear expected benefit. The imputation is fitted on the training data only.

Exploratory Analysis

1 shows the distribution of the target across its three classes: High 350 (43.8%), Low 260 (32.5%), and Medium 190 (23.8%). This is a moderate, not severe, imbalance (a 1.8:1 ratio between the largest and smallest class); it is handled explicitly with class weighting and SMOTE.

Age sits mostly between the mid-20s and early-30s, with a median near 27 and a few older outliers in the low-40s. The number of latest pregnancies is concentrated at one or two, with a small number of higher values reaching seven. PHQ-9 scores are fairly symmetric around a median of about 11, while EPDS scores have a noticeably wider spread. The demographic features turned out to carry little separating power on their own: age was almost identical across the Low, Medium and High groups. The clearest signal came from antenatal depression (PHQ-2): among the smaller group who screened positive, the High-risk class took up a much larger share while the Low-risk class shrank.

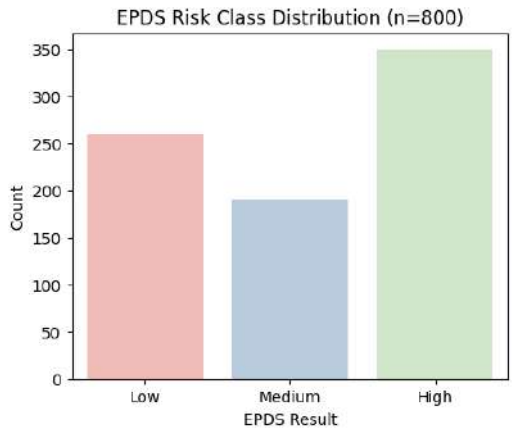

Figure 1: EPDS Result class distribution (n = 800

A correlation matrix of the four numeric features (Figure 2) makes the PHQ-9/EPDS relationship quantitatively visible before any modelling: PHQ-9 Score correlates with EPDS Score more strongly than any other numeric-feature pair, an early, purely descriptive signal of the circularity formally tested.

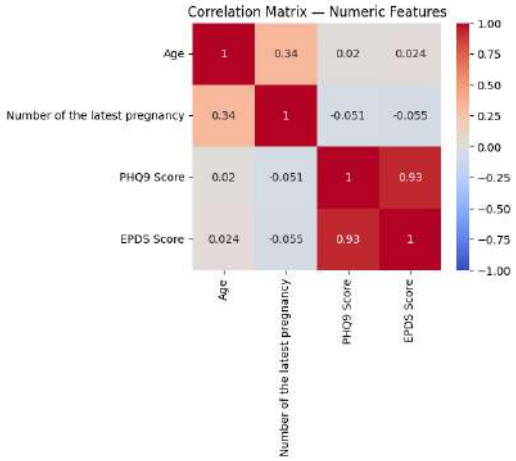

Figure 2: Correlation matrix, numeric features (Age, number of latest pregnancy, PHQ-9 Score, EPDS Score

Two things followed from this for the modelling stage. Because the demographic features did not separate the classes but the psychosocial and mental-health signal did, the predictive value was likely to come from interactions between features rather than any single variable, which favoured Random Forest and XGBoost. The uneven spread of records across the EPDS categories also hinted at class imbalance, which is why ROC-AUC, alongside macro-F1 and balanced accuracy, was used for comparing models rather than accuracy alone.

Train-Test Split, Imputation, and Encoding

An 80/20 split, stratified by EPDS Result with a fixed random_state of 42, is performed first. The mode for each of the five low-missingness categorical columns is then computed from the training rows only and applied to fill missing values in both sets, so that no information from the held-out cases influences the fill values. Numeric features are standardised with a StandardScaler fitted on the training data only. Categorical features are one-hot encoded with drop_first = True, and the test set is reindexed to the training columns with missing columns filled with zeros. A second, independent implementation of this logic, built as an sklearn ColumnTransformer/Pipeline that refits imputation, scaling and encoding on the training fold of every cross-validation split, is used for all cross-validated results, so no result in this paper depends on information leaking from a validation fold into its own training fold.

Stratifying the split (needed for reliable cross-validation and for McNemar's test to have a well-defined, class-balanced test set) gives a held-out test set of exactly 70 High, 52 Low, and 38 Medium cases (160 total), and every confusion matrix reported, sums exactly to 160 for every model.

Model Training

Logistic Regression was set with max_iter = 3000 so that it would converge given the many columns created by one-hot encoding, then fitted on the training data. Random Forest was set with 100 estimators, random_state = 42 and oob_score = True, which turns on out-of-bag validation using the samples left out of each bootstrap. XGBoost was set with 100 estimators, a learning rate of 0.1, a maximum depth of 3, the mlogloss evaluation metric and random_state = 42; a thin wrapper class label-encodes the target internally and decodes predictions back to High/Low/Medium, so that XGBoost can be used interchangeably with the other models inside scikit-learn's VotingClassifier and cross_validate utilities.

Evaluation Metrics

In addition to the confusion matrix, per-class precision/recall/F1, and the one-vs-rest ROC-AUC, this study reports: macro-averaged and weighted-averaged F1 (the former treats all three classes equally regardless of size, which matters given the class imbalance); balanced accuracy (the average of per-class recall, another imbalance-robust summary); the Matthews Correlation Coefficient (MCC), a single-number summary that is generally considered more informative than accuracy for imbalanced multi-class problems because it accounts for all four confusion-matrix quadrants simultaneously; and a macro-average one-vs-rest ROC-AUC (the unweighted mean of the three per-class AUCs). Ninety-five percent confidence intervals for accuracy, macro-F1, and balanced accuracy on the held-out test set were obtained by non-parametric bootstrap (2,000 resamples, sampling test-set rows with replacement and recomputing each metric per resample).

Ensemble Model

After evaluating the three base classifiers individually, a Soft Voting Ensemble was constructed using scikit-learn's VotingClassifier with voting = 'soft'. Each base learner contributes its predicted class probabilities, and the ensemble averages them before assigning the class with the highest averaged probability. Because the three base learners are structurally different (a linear model, a bagging ensemble, and a boosting

ensemble), their errors are partially independent, and averaging can cancel out some individual mistakes. Whether it does so to a statistically meaningful degree here is examined directly in later section.

Cross-Validation and Statistical Significance Testing

A single 80/20 split can make a small, noisy gap between two models look like a real effect. To guard against this, two complementary layers of validation are used. First, repeated stratified 5-fold cross-validation (5 folds, 10 repeats, 50 fits per algorithm) is run using the leakage-free pipeline, reporting mean and standard deviation of accuracy, macro-F1, and balanced accuracy. Second, on the held-out 160-case split, two paired tests compare the Voting Ensemble against each base learner: McNemar's test on the per-instance correct/incorrect indicator, and DeLong's test (a closed-form covariance estimator, validated against scikit-learn's roc_auc_score) comparing one-vs-rest ROC-AUC class by class.

Circularity Check: PHQ-Removed Comparison

To quantify how much of each model's performance depends on the PHQ-2/PHQ-9 features, the entire cross-validated pipeline was rerun with Depression before pregnancy (PHQ2), Depression during pregnancy (PHQ2), PHQ9 Score, and PHQ9 Result removed from the feature set before the train/test split, leaving the remaining sociodemographic, economic, obstetric, psychosocial, lifestyle, relationship, and newborn-care variables as the sole predictors. Results for the PHQ-included and PHQ-removed feature sets are reported side by side in later section.

Class Imbalance Handling

Given the moderate imbalance reported (43.8% High / 32.5% Low / 23.8% Medium), two standard balancing strategies were compared against the unweighted baseline using Random Forest as the representative model: (i) class_weight = 'balanced', which reweights the loss function inversely to class frequency; and (ii) SMOTE (Synthetic Minority Oversampling Technique), applied only to the training fold of each cross-validation split via an imbalanced-learn pipeline, so that synthetic points are never present in a validation fold.

Feature Selection

With 800 records and roughly 87 dimensions after one-hot encoding, the feature-to-row ratio invites overfitting. Mutual-information-based univariate feature selection (scikit-learn's SelectKBest with mutual_info_classif) was compared against the full encoded feature set for Random Forest, at k = 10, 20, and 30 selected features, again under repeated cross-validation with selection refit inside each training fold.

Interpretability

Three complementary views of feature importance were computed for Random Forest: (i) built-in impurity-based importance; (ii) permutation importance on the held-out test set (20 repeats, scored by the drop in macro-F1 when each feature is shuffled), which is less biased toward high-cardinality features; and (iii) SHAP values computed with a TreeExplainer, which attribute each prediction to individual feature contributions and allow both global and local interpretation.

Results and Discussion

Baseline Comparison on a Held-Out Split

Table 3 reports the extended metric set for all four models on the stratified 80/20 split (160 held-out cases: 70 High, 52 Low, 38 Medium).

Table 3: Held-Out Test Set (n = 160), Extended Metrics

Model	Accuracy	Macro F1	Weighted F1	Balanced Acc.	MCC	Macro ROC-AUC
Logistic Regression	0.788	0.742	0.777	0.742	0.669	0.903
Random Forest	0.756	0.683	0.730	0.695	0.623	0.900
XGBoost	0.781	0.749	0.778	0.745	0.660	0.923
Voting Ensemble	0.788	0.741	0.777	0.741	0.669	0.915

Note. Macro ROC-AUC is the unweighted mean of the three one-vs-rest AUCs.

Table 4: Non-Parametric Bootstrap 95% Confidence Intervals (2,000 Resamples), Held-Out Test Set

Model	Accuracy [95% CI]	Macro F1 [95% CI]	Balanced Acc. [95% CI]
Logistic Regression	0.787 [0.725, 0.850]	0.740 [0.668, 0.808]	0.742 [0.675, 0.806]
Random Forest	0.757 [0.688, 0.825]	0.681 [0.606, 0.756]	0.694 [0.628, 0.758]
XGBoost	0.782 [0.719, 0.844]	0.747 [0.680, 0.817]	0.746 [0.679, 0.816]
Voting Ensemble	0.788 [0.725, 0.850]	0.740 [0.666, 0.809]	0.741 [0.675, 0.806]

On this single split, XGBoost posts the highest macro-F1, balanced accuracy, and macro-average ROC-AUC of any individual model, and is essentially tied with Logistic Regression and the Voting Ensemble on accuracy. Random Forest is the weakest of the four here on every metric except MCC. It is shown with proper resampling, single-split rankings like this one are not especially stable, and the 95% bootstrap confidence intervals below (2,000 resamples of the 160-case test set) make the same point directly: all four models' intervals overlap almost completely.

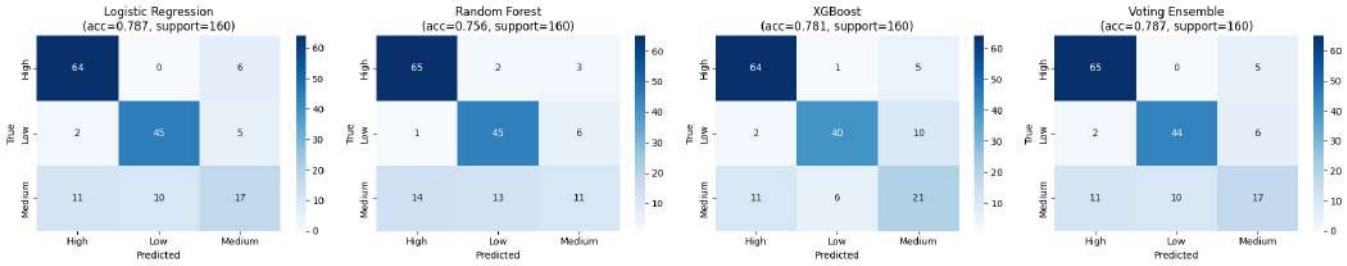

Figure 3: Confusion matrices, all four models, held-out test set. Each matrix sums exactly to 160

All four confusion matrices in Figure 3 sum exactly to the true test-set support of 160. The error-pattern differences between models are visible directly: Random Forest correctly identifies only 11 of 38 Medium cases (29% recall), noticeably worse than the other three models (which each correctly identify 17-21 of 38, 45-55% recall); XGBoost's largest single source of error is Low cases misclassified as Medium (10 of 52). Every model's dominant error pattern involves the Medium class, consistent with the overlapping EPDS score bands.

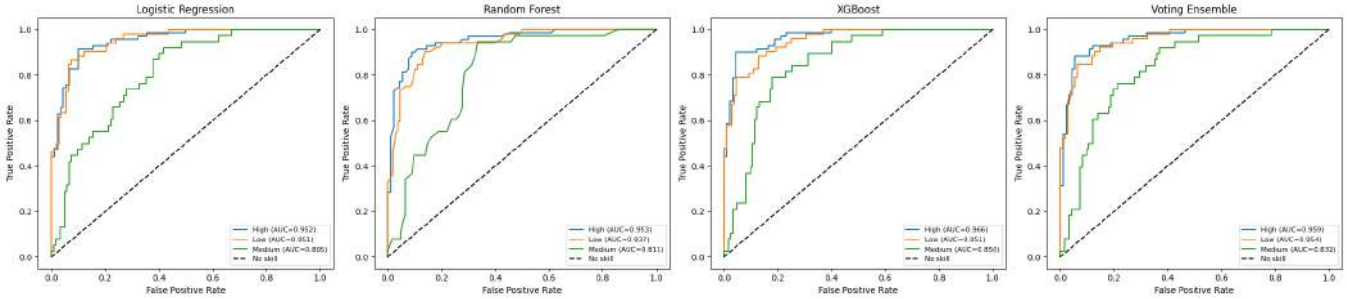

Figure 4: One-vs-rest ROC curves, all four models including the Voting Ensemble

The Medium-class ROC curve in every panel of Figure 4 climbs well above the no-skill diagonal, with AUC values between 0.805 (Logistic Regression) and 0.850 (XGBoost), a curve that discriminates meaningfully better than chance, just consistently less well than the High and Low curves (AUC 0.94-0.97 throughout).

Cross-Validated Comparison

Table 5 reports repeated stratified 5-fold cross-validation (10 repeats, 50 fits per model) over the full dataset, using the leakage-free ColumnTransformer pipeline.

Table 5: Repeated Stratified 5-Fold CV, 10 Repeats (50 Folds Total per Model)

Model	Accuracy (mean ± SD)	Macro F1 (mean ± SD)	Balanced Acc. (mean ± SD)
Logistic Regression	0.743 ± 0.030	0.700 ± 0.032	0.702 ± 0.032
Random Forest	0.762 ± 0.027	0.709 ± 0.034	0.713 ± 0.031
XGBoost	0.774 ± 0.027	0.744 ± 0.031	0.743 ± 0.032
Voting Ensemble	0.765 ± 0.028	0.725 ± 0.034	0.725 ± 0.032

Under this more stable estimate, XGBoost has the highest mean accuracy, macro-F1, and balanced accuracy of any model, with the Voting Ensemble second and Random Forest and Logistic Regression close behind. The gap between the Voting Ensemble and XGBoost (0.9 accuracy points) is smaller than a single fold-to-fold standard deviation (2.7-2.8 points) for either model, and Figure 5 shows the four boxplots overlapping substantially. Cross-validation therefore favours XGBoost as the strongest individual choice by a small margin, with the Voting Ensemble closely behind and effectively tied with it.

Statistical Significance Testing

McNemar's test compares the Voting Ensemble against each base learner on the held-out test set's per-case correct/incorrect outcomes:

Table 6: McNemar's Test (Continuity-Corrected), Voting Ensemble vs. Each Base Learner

Comparison	Ensemble-only correct	Base-only correct	Statistic	p-value	Significant?
Ensemble vs. Logistic Regression	3	3	3.00	1.000	No
Ensemble vs. Random Forest	10	5	5.00	0.302	No
Ensemble vs. XGBoost	7	6	6.00	1.000	No

DeLong's test compares the Voting Ensemble against Random Forest on one-vs-rest AUC, class by class:

Table 7: DeLong's Test, Voting Ensemble vs. Random Forest, per Class

Class	Ensemble AUC	Random Forest AUC	z	p-value	Significant?
High	0.959	0.953	1.43	0.152	No
Low	0.954	0.937	2.24	0.025	Yes
Medium	0.832	0.811	1.19	0.236	No

None of the three McNemar comparisons reach significance. Of the three DeLong comparisons, two do not reach significance, and one (the Low-class AUC) does (p = 0.025), a single class-level result out of nine paired comparisons across both tests on one 160-case split. Combining Table 5, Table 6, and Table 7, the fair summary is: the Voting Ensemble is competitive with, and on one class modestly better calibrated than, Random Forest specifically, but it is not reliably better than XGBoost. Given also that none of the four models were hyperparameter-tuned, any ranking among them in this paper should be read as best under default settings.

Circularity: Effect of Removing PHQ-2/PHQ-9 Features

Table 8 compares cross-validated performance with and without the four PHQ-derived features.

Table 8: Repeated 5-Fold CV (10 Repeats), Full Feature Set vs. PHQ-2/PHQ-9 Features Removed

Model	Accuracy (PHQ)	Accuracy (no PHQ)	Drop (pts)	Macro F1 (PHQ)	Macro F1 (no PHQ)
Logistic Regression	0.743	0.576	16.7	0.700	0.508
Random Forest	0.762	0.592	17.1	0.709	0.499
XGBoost	0.774	0.582	19.1	0.744	0.505
Voting Ensemble	0.765	0.592	17.2	0.725	0.509

Removing the PHQ-derived features costs every model 17 to 19 accuracy points and a comparable drop in macro-F1 a large, consistent effect across four structurally different algorithms. For context, a majority-class baseline (always predicting High) would score approximately 44% accuracy on this target; the PHQ-removed models (58-59% accuracy) clear that baseline by a modest but real margin, while the PHQ-included models (74-77%) do substantially better. This is the quantitative confirmation of the circularity risk: a meaningful share, though not all, of the model's headline accuracy is attributable to one depression-screening instrument (PHQ-9, administered in the same interview) predicting another (EPDS), rather than to the sociodemographic, obstetric, and relationship variables this study frames as its main contribution.

We think the most defensible way to present this in a screening context is to report both numbers rather than either alone: the PHQ-included accuracy answers does a combination of PHQ-9 and contextual factors predict EPDS risk band, a question with real clinical value (since PHQ-9 is itself a two-minute screening tool a clinic could administer directly), while the PHQ-removed accuracy answers the more novel and more difficult question this paper is centred on: whether risk can be flagged from contextual factors alone, without any depression screen already in hand. On that harder question, 58-59% accuracy on a three-class problem is a genuine, if modest, positive result, and a fairer basis for any claim about the study's contribution than the PHQ-included headline number.

Class Imbalance: Reporting and Handling

The class distribution (43.8% High, 32.5% Low, 23.8% Medium) is moderately imbalanced. Table 9 compares the unweighted Random Forest baseline against $class_weight = ' balanced'$ and SMOTE, both under repeated cross-validation.

Table 9: Class Imbalance Handling, Repeated 5-Fold CV (5 Repeats)

Configuration	Accuracy	Macro F1	Balanced Accuracy
Random Forest (baseline, unweighted)	0.763	0.711	0.714
Random Forest (class_weight = balanced)	0.763	0.711	0.715
Random Forest + SMOTE (train fold only)	0.772	0.734	0.736

class_weight = 'balanced' makes almost no difference here, but SMOTE gives a small, consistent improvement over the unweighted baseline (macro-F1 and balanced accuracy both up roughly 2 points), driven mainly by better recall on the minority Medium class. Given the imbalance is moderate rather than severe, this is a modest but

genuine gain; SMOTE (applied training-fold-only, as here, to avoid leaking synthetic minority examples into validation data) is the more defensible default for a future version of this pipeline.

Feature Selection

Table 10 compares the full 87-dimension encoded feature set against mutual-information-selected subsets of 10, 20, and 30 features, for Random Forest under repeated cross-validation.

Table 10: Mutual-Information Feature Selection, Repeated 5-Fold CV (5 Repeats), Random Forest

Feature Set	Accuracy	Macro F1	Balanced Accuracy
All features (∼87)	0.763	0.711	0.714
Top-10 mutual-information features	0.752	0.718	0.718
Top-20 mutual-information features	0.755	0.717	0.718
Top-30 mutual-information features	0.764	0.724	0.725

The top 10 mutual-information features alone recover essentially all of the full 87-feature model's accuracy, and slightly exceed it on macro-F1 and balanced accuracy; the top 30 features modestly outperform the full set on every metric. This indicates that most of the 87 one-hot-encoded dimensions contribute little beyond noise for this task and this sample size: a feature-selected model with a fraction of the dimensions is both more interpretable and, if anything, slightly more robust than the full-dimensional model used throughout the rest of this paper.

Interpretability: Feature Importance and SHAP

Three independent views of Random Forest's feature importance were computed on the held-out split: built-in impurity-based importance, permutation importance, and SHAP. All three agree closely on which feature matters most.

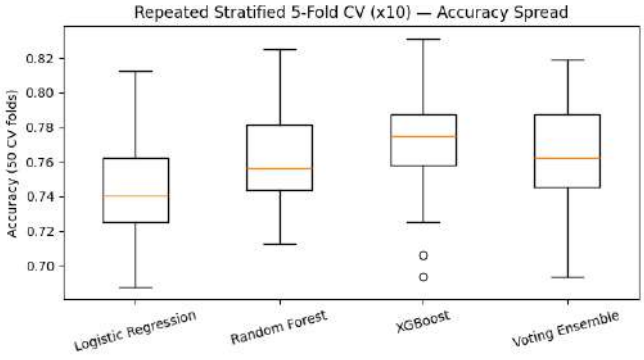

Figure 5: Random Forest built-in (impurity-based) feature importance, top 15 features

PHQ9 Score dominates every importance measure by a wide margin, roughly five times the impurity-based importance of the next-ranked feature (Figure 6), and by far the largest driver of macro-F1 when permuted. The SHAP summary plot (Figure 7) shows the same pattern for individual predictions: PHQ9 Score has both the largest average SHAP magnitude and a clean, monotonic relationship with predicted risk, while every other feature's contribution is visibly smaller and less consistent. This confirms that the model's skill is concentrated in a single depression-screening feature rather than distributed across the psychosocial and obstetric variables this study centres on. It also delivers a genuinely interpretable screening aid, though the interpretation it yields is a cautionary one about circularity rather than a list of actionable risk factors.

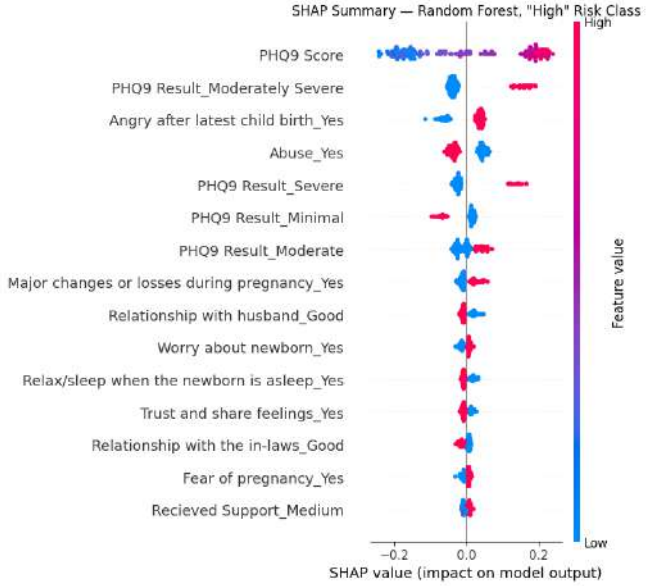

Figure 6: SHAP summary plot, Random Forest, High risk class

Why the Tree-Based Models and Ensemble Outperform Logistic Regression

Under the single 80/20 split (Table 3), Logistic Regression is competitive with or ahead of Random Forest and roughly tied with XGBoost and the Ensemble, a reminder that single-split comparisons are noisy. Under the more reliable cross-validated comparison (Table 5), a clearer pattern holds: Logistic Regression has the lowest mean accuracy, macro-F1, and balanced accuracy of the four models, 2-3 points behind XGBoost and 1-2 points behind Random Forest and the Ensemble. Three factors plausibly explain this gap. First, it was found that no single demographic feature separates the risk classes well, so the separating signal that does exist likely lives in interactions between features (for example, PHQ9 Score combined with a specific relationship-support or abuse category) rather than in additive main effects. A linear model like Logistic Regression can only capture such interactions if they are engineered as explicit terms, which was not done here, whereas Random Forest and XGBoost partition the feature space hierarchically and pick up conditional effects automatically. Second, one-hot encoding produces roughly 87 sparse, often correlated binary columns, including several PHQ9 Result category dummies that partially duplicate the continuous PHQ9 Score. Logistic Regression's coefficients spread importance across correlated dummies fairly evenly, while tree-based splits select the single most informative cut points directly, consistent with the finding that a handful of mutual-information-selected features recovers almost all of the full model's accuracy. Third, the Medium class sits, by construction, in a score band that overlaps both its neighbours (Table 1). This is inherently a fuzzy, nonlinear boundary problem, and axis-aligned tree partitioning is generally better suited to it than a single global linear decision surface per class. The Voting Ensemble's own edge over Logistic Regression follows from these same mechanisms, inherited from its two tree-based base learners, though averaging in a weaker linear model

pulls the ensemble toward, rather than above, its strongest individual member.

Limitations

Geographic and cultural transfer to Nepal is assumed, not demonstrated. The models were trained on data from Bangladesh. The claim that Bangladesh is culturally close to Nepal rests on broad regional similarity and is not validated against any Nepal-specific data. The two countries differ in healthcare-system structure, maternal-care access, and the socioeconomic correlates of PPD reported locally (Neupane et al. 2024). No external validation on a second dataset was performed, so any conclusion about usefulness for Nepal should be read as a hypothesis for future work, not a validated finding.

Dropped high-missingness columns may remove clinically important predictors. Income, addiction history, and disease history were dropped for missingness above 50%, not because they lack predictive value; all three are literature-supported PPD risk factors. Their absence likely removes real signal and reflects the dataset's completeness rather than the research question.

No hyperparameter tuning. All four models were run with default-adjacent settings. Every comparison in this paper should be read as default-configuration models, not each algorithm's best achievable performance, particularly given how close several of the cross-validated comparisons already are (Table 5).

Sample size and stability. The dataset is fairly small at 800 records, and the Medium class has few cases (190 of 800, only 38 in the test split). Per-class metrics for Medium carry wide uncertainty, visible in the bootstrap intervals in Table 4.

The EPDS label is a screening score, not a clinical diagnosis. This matters more, not less, once it is shown how much of the model's skill comes from another self-report screening instrument rather than independently observable risk factors.

Measurement circularity between predictors and target. This is the most consequential limitation: PHQ-9 and EPDS were collected in the same structured interview, and much of every model's apparent skill depends on this overlap. For a genuinely novel-information screening tool, the PHQ-removed results (Table 8, 58-59% accuracy) are the more honest benchmark.

For all these reasons the models in this paper should be treated as a decision-support aid whose current evidence base is modest, not a validated or deployable clinical screening tool.

Conclusion

This study applied supervised machine learning to predict postpartum depression risk using a Bangladeshi maternal dataset that covers sociodemographic, economic, obstetric, psychosocial, lifestyle, relationship, newborn-care and mental-health factors. Four classifiers were trained and compared under repeated cross-validation with paired significance testing: Logistic Regression, Random Forest, XGBoost and a Soft Voting Ensemble. Under this validation, XGBoost had the highest mean accuracy, macro-F1, and balanced accuracy of any individual model, with the Voting Ensemble a close second; McNemar's and DeLong's tests found no significant pairwise difference between the Ensemble and its base learners on the held-out split except a single class-level AUC comparison. The Ensemble and XGBoost are both competitive, well-calibrated choices under default hyperparameters, and combining structurally diverse classifiers remains a reasonable, low-effort way to approach this kind of problem, without a guarantee of outperforming a well-chosen single model.

The more consequential finding of this study is not about which model wins, but about what the winning models are actually learning. Removing the PHQ-2/PHQ-9 features cost every algorithm 17 to 19 accuracy points, and three independent interpretability methods (built-in importance, permutation importance, and SHAP) all identify PHQ-9 Score as the single dominant predictor, several times more influential than any other feature. Because PHQ-9 and EPDS were collected in the same structured interview, a substantial part of this study's headline accuracy reflects one depression-screening instrument predicting another, rather than novel signal from sociodemographic, obstetric, or relationship variables. This does not make the exercise worthless: a model that combines a brief screen like PHQ-9 with contextual factors to estimate EPDS risk band still has plausible clinical value, since it could let a service that already has PHQ-9 data flag likely EPDS risk without a second instrument, but that is a more modest and specific claim than predicting PPD risk from sociodemographic and psychosocial factors alone. This is why the title of this paper describes a single-country cohort study rather than claiming representativeness of South Asian women from a single-country dataset, and why ensembling is presented as a well-established, useful technique rather than a novel contribution in itself.

In practical terms, the work still speaks to a real healthcare problem: the early detection of a condition that often goes unreported. An affordable, data-driven tool of this kind could fit into a stretched health system if its evidence base were strengthened further. Future work should obtain or simulate a dataset in which PHQ and EPDS are administered at genuinely different time points, validate on a second, ideally Nepal-specific, maternal dataset, tune each base learner's hyperparameters, and test whether the PHQ-removed feature set combined with better handling of the currently dropped income and health-history variables can narrow the gap to the PHQ-included model. On the evidence gathered here, machine learning, and gradient boosting or a soft-voting ensemble in particular, retains real, if more modest than a headline accuracy number alone would suggest, potential as an early-screening aid for postpartum depression in maternal healthcare.

Disclosure Statement

The authors have no financial or non-financial disclosures to share for this article.

Ethical Approval: Not applicable. The study used a publicly available, de-identified secondary dataset and did not involve direct contact with human participants.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Data Availability Statement: The dataset analysed in this study is publicly available from Mendeley Data at https://data.mendeley.com/datasets/4nznnrk8cg/2 (Raisa and Kaiser 2025). An accompanying analysis notebook (PPD_Analysis_Revised.ipynb) reproduces every number, table, and figure in this paper.

Use of Artificial Intelligence (AI) Tools: AI-assisted tools were used to help implement the analysis pipeline (preprocessing, cross-validation, statistical testing, and interpretability code), and to help check language and formatting throughout. The authors are responsible for the accuracy, originality, and interpretation of the work.

Funding: This research received no specific grant from any funding agency.

Competing Interests: The authors declare no competing interests.

References

Cox, J. L., J. M. Holden and R. Sagovsky (1987). "Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale." In: *British Journal of Psychiatry* 150.6, pp. 782–786. ISSN: 1472-1465. 10.1192/bjp.150.6.782. http://dx.doi.org/10.1192/bjp.150.6.782.

Gopalakrishnan, Abinaya, Revathi Venkataraman, Raj Gururajan, Xujuan Zhou and Guohun Zhu (Dec. 2022). "Predicting Women with Postpartum Depression Symptoms Using Machine Learning Techniques." In: *Mathematics* 10.23, p. 4570. ISSN: 2227-7390. 10.3390/math10234570. http://dx.doi.org/10.3390/math10234570.

GUINTIVANO, JERRY, TRACY MANUCK and SAMANTHA MELTZER-BRODY (2018). "Predictors of Postpartum Depression: A Comprehensive Review of the Last Decade of Evidence." In: *Clinical Obstetrics & Gynecology* 61.3, pp. 591–603. ISSN: 0009-9201. 10.1097/grf.0000000000000368. http://dx.doi.org/10.1097/GRF.0000000000000368.

Habehh, Hafsa and Suril Gohel (Dec. 2021). "Machine Learning in Healthcare." In: *Current Genomics* 22.4, pp. 291–300. ISSN: 1389-2029. 10.2174/1389202922666210705124359. http://dx.doi.org/10.2174/1389202922666210705124359.

Huang, Xudong, Lifeng Zhang, Chenyang Zhang, Jing Li and Chenyang Li (Aug. 2025). "Postpartum depression risk prediction using explainable machine learning algorithms." In: *Frontiers in Medicine* 12. ISSN: 2296-858X. 10.3389/fmed.2025.1565374. http://dx.doi.org/10.3389/fmed.2025.1565374.

Hwang, W. Y., S. Y. Choi and H. J. An (2022). *Concept analysis of transition to motherhood*. Parsed from reference text – no usable DOI found.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." In: *Science* 349.6245, pp. 255–260. ISSN: 1095-9203. 10.1126/science.aaa8415. http://dx.doi.org/10.1126/science.aaa8415.

Karki, Dipendra (2012). *Economic impact of tourism in Nepal's economy using cointegration and error correction model*. en. 10.13140/RG.2.1.4839.5684. https://www.researchgate.net/doi/10.13140/RG.2.1.4839.5684.

Karki, Dipendra, Nirupan Karki, Rewan Kumar Dahal and Ganesh Bhattarai (Dec. 2023). "Future of Education in the Era of Artificial Intelligence." In: *Journal of Interdisciplinary Studies* 12.1, pp. 54–63. ISSN: 2392-4519. 10.3126/jis.v12i1.65448. http://dx.doi.org/10.3126/jis.v12i1.65448.

Natarajan, Sriraam, Annu Prabhakar, Nandini Ramanan, Anna Bagilone, Katie Siek and Kay Connelly (2017). "Boosting for Postpartum Depression Prediction." In: *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, pp. 232–240. 10.1109/chase.2017.82. http://dx.doi.org/10.1109/CHASE.2017.82.

Neupane, Maryada, Manita Bartaula, Simran Pradhan, Hom Prasad Adhikari, Lalita Shrestha, Puja Sharma and Nishchal Devkota (Aug. 2024). "Postpartum Depression among Mothers in a Maternity Hospital Kathmandu, Nepal: A Mixed Method Approach." In: *Journal of Nepal Medical Association* 62.277, pp. 575–581. ISSN: 0028-2715. 10.31729/jnma.8746. http://dx.doi.org/10.31729/jnma.8746.

OHara, Michael W. and Jennifer E. McCabe (Mar. 2013). "Postpartum Depression: Current Status and Future Directions." In: *Annual Review of Clinical Psychology* 9.1, pp. 379–407. ISSN: 1548-5951. 10.1146/annurev-clinpsy-050212-185612. http://dx.doi.org/10.1146/annurev-clinpsy-050212-185612.

Qi, Weijing, Yongjian Wang, Yipeng Wang, Sha Huang, Cong Li, Haoyu Jin, Jinfan Zuo, Xuefei Cui, Ziqi Wei, Qing Guo and Jie Hu (Mar. 2025). "Prediction of postpartum depression in women: development and validation of multiple machine learning models." In: *Journal of Translational Medicine* 23.1. ISSN: 1479-5876. 10.1186/s12967-025-06289-6. http://dx.doi.org/10.1186/s12967-025-06289-6.

Raisa, J. F. and M. S. Kaiser (2025). *Data for postpartum depression prediction in Bangladesh*. Parsed from reference text – no usable DOI found.

Rajbhandari, Sharad, Ghanashyam Khanal, Seeprata Parajuli and Dipendra Karki (Dec. 2020). "A Review on Potentiality of Industry 4.0 in Nepal: Does the Pandemic Play Catalyst Role?" In: *Quest Journal of Management and Social Sciences* 2.2, pp. 366–379. ISSN: 2705-4527. 10.3126/qjmss.v2i2.33307. http://dx.doi.org/10.3126/qjmss.v2i2.33307.

Russell, S. and P. Norvig (2021). *Artificial Intelligence: A Modern Approach (4th ed.)* Parsed from reference text – no usable DOI found. Pearson.

Saqib, Kiran, Amber Fozia Khan and Zahid Ahmad Butt (Nov. 2021). "Machine Learning Methods for Predicting Postpartum Depression: Scoping Review." In: *JMIR Mental Health* 8.11, e29838. ISSN: 2368-7959. 10.2196/29838. http://dx.doi.org/10.2196/29838.

Shorey, Shefaly, Cornelia Yin Ing Chee, Esperanza Debby Ng, Yiong Huak Chan, Wilson Wai San Tam and Yap Seng Chong (2018). "Prevalence and incidence of postpartum depression among healthy mothers: A systematic review and meta-analysis." In: *Journal of Psychiatric Research* 104, pp. 235–248. ISSN: 0022-3956. 10.1016/j.jpsychires.2018.08.001. http://dx.doi.org/10.1016/j.jpsychires.2018.08.001.

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 23–31

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/3fws8j75

Impact of Overconfidence Bias on Investment Decision Making: Mediating Role of Risk Tolerance among NEPSE Investors

Sahara Shrestha[1,*], Koshish Jung Lamichhane[1]

[1]BA (Hons) Accounting and Finance, Islington College, Kathmandu, Nepal
[*]*Correspondence: sahara.shtha@gmail.com*

ARTICLE HISTORY

Received: 11 April 2026 Revised: 14 May 2026
Accepted: 24 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Shrestha, S. and Lamichhane, K. J. (2026). 'Impact of overconfidence bias on investment decision making: Mediating role of risk tolerance among NEPSE investors', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 23–31. Available at: https://doi.org/10.67556/3fws8j75

ABSTRACT

Based on the behavioral finance theory, this study aims at investigating the role of overconfidence bias in the investment behavior of individual investors who transact on Nepal Stock Exchange (NEPSE) and whether risk tolerance acts as a mediator between the two. The research design adopted was quantitative, cross sectional and the primary data are collected using a structured questionnaire to the active retail investors of NEPSE. Correlation, multiple regression, and mediation analysis (Hayes' PROCESS macro) were used to analyze the data in SPSS. The results suggest that the relationship between overconfidence bias and investment decisions is direct, and that the relationship between overconfidence bias and investment decisions is partially mediated by risk tolerance. Finally, the study reveals that overconfidence is an important behavioral factor in the investment decision-making process in the Nepalese capital market, both directly and indirectly via the investors' willingness to take risks. The findings indicate that there is a need for a better targeted investor education and regulatory awareness program to promote more rational investment behavior of NEPSE investors.

Keywords: Overconfidence bias, risk tolerance, investment decision-making, behavioral finance, NEPSE

JEL Classification: G11 (Portfolio Choice); G40 (Behavioral Finance, General); G41 (Role and Effects of Psychological, Emotional, Social, and Cognitive Factors on Decision Making in Financial Markets)

Introduction

The concept of classical finance theory assumes rational behavior on the part of the investor, who understands and can efficiently process all the information available to him and who makes decisions that maximize his expected utility. Behavioral finance disproves this premise by demonstrating that psychological influences consistently bias financial decision making. Key concepts in this area of work include heuristic-driven decision making, whereby investors make decisions based on mental shortcuts instead of the full processing of information; the herding effect, which describes the tendency of investors to follow herd mentality and follow others as opposed to doing their own analysis; and prospect theory, which explains asymmetric risk preferences like loss aversion (M. A. Sattar, Toseef and M. F. Sattar 2020).All taken together, the results suggest that these behavioral factors can result in investors taking risks that a rational model would not suggest.

Overconfidence is an investor's tendency to underestimate the accuracy of his or her knowledge, skill, and judgement, as identified in this literature, and is one of the most well documented biases (Broekema and Kramer 2021). Previous international studies suggest that overconfidence is associated with excessive trading, with overtrading causing portfolio performance to suffer due to transaction costs and poor trading timing and with lack of diversification, since overconfident investors think their own ideas are superior to others (Statman, Thorley and Vorkink 2006; Barber and Odean 2001; Gervais and Odean 2001).

In Nepal, more access to online trading platforms, real-time market news, and sentiment on social media can result in a

stronger investor confidence than what available information might suggest. Furthermore, the Nepalese capital market is dominated by retail investors and the investors are young, with comparatively limited access to independent research and, in these circumstances, behavioral biases are more likely to affect their investment decisions than they would in a more mature market. Although there is increasing international evidence, the psychological mechanism by which overconfidence turns into investment decision has been studied relatively little among the Nepalese investors, specifically, the tolerance of risk of the investor in channeling this mechanism. It is important to understand this mechanism because it helps to elucidate the nature of the role played by overconfidence-whether it is a direct driver of investment decisions or whether it is a signal of risk appetite. Hence, in this study the researcher tried to examine the relation between overconfidence bias and investment decision making of NEPSE investors considering risk tolerance as a mediating variable. For this, the project aims to:

1. Examine the impact of over confidence bias in investment decision making of NEPSE investors.
2. Investigate the impact of overconfidence bias on the risk-taking capacity of the investors in the NEPSE.
3. To analyze risk tolerance impacts on investment decision making of the NEPSE investors.
4. Investigate the mediating effect of risk tolerance between overconfidence bias and investment decision-making of NEPSE investors.

Literature and Related Work

The impact of overconfidence bias on risk-taking and portfolio management is examined in this paper. A major component of behavioral finance is overconfidence, which causes investors to overestimate their skills, trade excessively, misjudge risks, and handle portfolios badly. The emphasis emphasizes how crucial it is to comprehend psychological aspects to make wiser financial judgements.

Theoretical Review

Behavioral finance examines the impact of psychological factors on financial decisions with the challenge that investors are not entirely rational. It combines the knowledge in psychology with conventional finance to describe why investors usually do not follow rational decision-making models. Key principles include Heuristic behavior, in which decisions are based on rules of thumb or trial and error, and the Herding Effect, in which investors follow others rather than performing their own study. Prospect Theory emphasizes biases such as loss aversion, which leads to risk-seeking in prospective losses and risk-avoidance in potential benefits (M. A. Sattar, Toseef and M. F. Sattar 2020).These findings demonstrate how behavioral factors might influence investors to accept more risks than sensible portfolio management would recommend.

The overconfidence bias has a significant impact on the investment decisions among NEPSE investors resulting in overtrading, poor diversification, and poor portfolio performance. This is more effective in the emerging capital market in Nepal where the access of investors to credible data is constrained. The extreme volatility of the price and the tendency to trade based on sentiment, further consolidate overconfidence as investors incorrectly attribute the short-term successes to their skills. Moreover, overconfident investors have a greater risk tolerance, which contributes to decision-making biases. To minimize such biases and encourage rational investment behavior within the NEPSE, enhanced investor education and awareness programs are necessary.

Empirical Studies

Developing Market Context

Studies have shown that overconfidence bias can lead to excessive trading, inefficient asset allocation, impulsive judgements, market timing errors, risk underestimating, and loss aversion (Awalakki 2023).Empirical study indicate that both overconfidence and herding are important influences on investment decisions, and these results are supported by earlier studies because of the tendency of investors, because of a lack of adequate knowledge, to make mistakes because of these biases, affecting financial advisers who try to help investors make right investment decisions (Madaan and Singh 2019).The findings indicate that overconfidence is a particularly dangerous bias in behaviors, since it causes investors to be guided by false confidence instead of judgement. Further studies have examined the determinants of overconfidence bias and how it influences investment performance through investors risk propensity. Additional research indicates that cognitive patterns like self-attribution, the illusion of control, and the conviction that one is superior to the typical investor are the root causes of overconfidence. Due to these psychological inclinations, investors in developing market are more likely to take bigger, sometimes unreasonable, risks (Dangol and Manandhar 2020). The results of the research above suggest that the underlying cognitive biases that cause overconfidence cause it to be an especially persuasive influence on investor behaviors, since they systematically promote larger and frequently irrational amounts of risk-taking. Subsequent studies show that overconfident investors often take unjustified risks and adopt unfavorable investment strategies because they believe they have superior information and forecasting skills. People often overestimate their skills and ignore crucial information, which makes them think they are superior to the accepted norm (Chaudhary 2025). This behavior shows that overconfidence always misleads the judgement of investors and drives them to take unnecessary risks.

Emerging Markets

Studies argue that overconfident investors are more likely to make quick decisions, rely too much on their own judgement, and be more eager to trade frequently. Other biases did exist, but their impact was insufficient to affect the results. This implies that overconfidence is the most significant behavioral element influencing investors' decision-making and investment

management (N. Aryal 2021). Specifically, three cognitive biasesillusion of control, self-attribution, and the better-than-average effect have been identified as key contributors to overconfidence. These biases not only directly affect investment decisions but also indirectly impact performance by shaping risk-taking behavior (Abdin et al. 2022). The most powerful bias in behavior is overconfidence, which influences the decision-making process of investors who tend to overtrade and depend too much on their judgment. Additional research reveals how behavioral biases are widespread in making investment choices across various markets and cultures. The findings demonstrate the pressing necessity of increasing investor awareness and education that could prevent the negative impact of those biases on investment outcomes. Combining the insights provided by the global and Nepalese perspectives, the literature gives a holistic view of how behavioral biasness promotes and influences investment behavior, and as such, it is relevant in both developed and emerging markets (Kandel, Basnet and M. Aryal 2024). Knowledge of behavioral biases is essential since a higher level of investor awareness and education can reduce their adverse impact and so this knowledge is useful in both developed and emerging markets. It is claimed that investors that are overconfident overestimate their skills and expertise, which causes them to trade and invest more since they think they can beat the market.

Nepalese Context

Overconfidence bias is not significantly related to investor experience, although it is related to the number of trades, suggesting that frequent investors tend to be more overconfident in their investment choices. Investors become more and more overconfident because they use mental shortcuts in their investment decisions, called heuristics, which involve overestimating the accuracy of their predictions and ignoring the potential risks, which reinforces the impact of psychological biases such as overconfidence on decision-making in a Nepalese investment context (Pandit 2021).Although this confidence might improve decisions, there are serious hazards associated with it since overconfident investors may ignore important information and misjudge market volatility, which could result in rash decisions and financial losses. To lessen its detrimental effects and encourage better investing decisions, overconfidence bias must be addressed through financial education and advice services. It is critical that investors, financial advisors, and legislators are aware of this bias (Gurung et al. 2024). Overconfidence may cause investors to overrate their capacities, risk too much and make poor choices concerning money, and in this regard the services of a financial educator and financial advisor can help in reducing the negative impacts of overconfidence.

Overall, overconfidence bias is a major source of irrational investment behavior, which makes investors overestimate their abilities, underestimate risks, and overtrade and in a poorly timed manner. This tendency is further supported by cognitive processes like the illusion of control, self-attribution bias, and better-than-average effect, leading to a higher risk-taking and a decrease in portfolio performance. It is important to note that overconfidence has a serval role in influencing investor behavior through risk tolerance. These findings imply that enhancing the financial education programs and investment advisory services can be used as effective interventions in reducing the negative effects of overconfidence bias and promote a more rational and informed investment decision-making.

Conceptual Framework

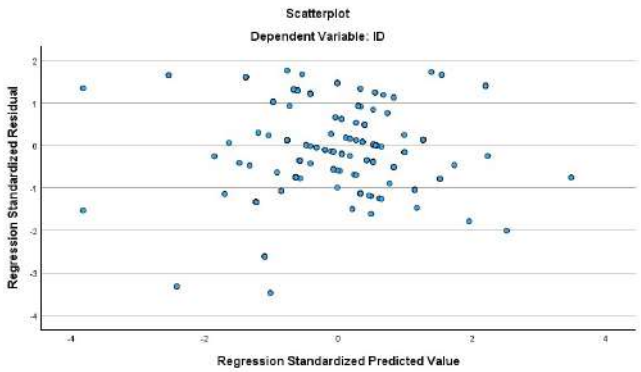

Figure 1: Conceptual Framework (Bist, 2025)

This conceptual framework states that the phenomenon of overconfidence bias affects the investment decisions of individual investors directly and indirectly through risk tolerance. Overconfidence also leads to greater risk taking where investors become more comfortable to take more risk and invest in risky markets . In that way, risk tolerance mediates the association between overconfidence bias and making investment decisions, which demonstrates the significance of the psychological aspect of investment decision-making in the developmental financial market in Nepal.

Research Hypotheses

H1: The overconfidence bias has a positive impact on investment decisions by Nepalese investors.

H2: The overconfidence bias has a positive effect on risk tolerance on Nepalese investors.

H3: The relationship between overconfidence bias and investment decisions of Nepalese investors is mediated by risk tolerance.

Methodology

Research Philosophy

The research philosophy that is adopted in this study is positivism. Positivism refers to a philosophical paradigm that places special focus on the establishment of laws that can be used in the explanation and prediction of both social and natural sciences. It is based on a hypothetico-deductive framework in which theories are developed, and hypothesis tested and the results of these tests are obtained based on empirical data (Park, et al., 2020).

This is suitable in the current study because it attempts to test

the causality of the relationship between overconfidence bias and investment decision-making and test the mediating influence of risk tolerance among NEPSE investors.

Research Approach

The research is deductive in nature and under this approach, a set of established premises are used to derive conclusions. It outlines the objectives to formulate accounting theory, based on assumptions to justify this method through inferences of these premises. The study looks at bias in investment decisions and evaluates the role of risk tolerance as a mediator between NEPSE investors, based on the behavioral finance theory (Zalaghi and Khazaei 2016).

Research Desgin

The research design is a quantitative, cross-sectional study that collects data through surveys among NEPSE investors. The research will examine the relationship between overconfidence bias and investment decision-making and use risk tolerance as a mediating variable. The structured questions are open to analysis using correlation, regression, and mediation methods, which results can be generalized to a larger population of NEPSE investors.

Population and Sampling Startegy

All the individual retail investors who are currently trading on the Nepal Stock Exchange and take their own investment decisions were included in the study population while institutional investors and non-active investors were excluded. Participants were selected using purposive sampling technique which is a non-probability sampling because it is used to reach participants which meet specific criteria in terms of age (18 years and older), actively trading, and responsible for their own investment decision. The minimum sample size required was computed using the formula for large or unknown populations used by Cochran (1977) for a 95% confidence level, a 5% margin of error and an assumed population proportion of 50%, yielding a minimum sample of around 384 respondents.

Data Collection Methods

The primary data was collected by using a structured questionnaire online (Google Forms) and investor networks and intermediaries. The questionnaire had four sections: 1) Demographic characteristics (age, gender, income, investment experience, and trading frequency); 2) An overconfidence bias scale adapted from existing behavioral finance questionnaires; 3) A risk tolerance scale which is based on the validated financial risk tolerance scale of Grable and Lytton (1999); 4) An investment decision making scale derived from previous research on investor behavior, information use and decision quality.

Variables and Measurements

Table 1: Held-Out Test Set (n = 160), Extended Metrics

Variable	Role	Basis	Dimensions	Measurement
Overconfidence Bias	Independent	Ahmad & Shah (2020)	Miscalibration, Overestimation, Perceived Knowledge	Likert-scale questions
Risk Tolerance	Mediating	Grable & Lytton (1999)	Willingness, Comfort, Psychological Preparedness	Multiple choice questions
Investment Decision-Making	Dependent	Ahmad & Shah (2020)	Information Evaluation, Investment Behavior Management	Likert-scale questions

Note. Variable roles and measurement instruments as applied in this study.

Data Analysis

Multiple regression was used to explore the relation between overconfidence bias and investment decision making, where investment decision making was the dependent variable and overconfidence bias and risk tolerance were the predictors. The mediating effect of risk tolerance between overconfidence bias and investment decision was tested using Hayes' PROCESS macro (Model 4), which is the product of path a (overconfidence bias risk tolerance) and path b (risk tolerance investment decisions, accounting for overconfidence bias). This indirect effect was evaluated using a percentile bootstrap confidence interval (with 5,000 resamples), and a confidence interval that excludes zero means that the indirect effect is statistically significant. Full mediation is suggested when the direct effect of overconfidence bias on investment decisions is not significant when risk tolerance is added to the model, and partial mediation is indicated when both direct and indirect effects are significant.

Ethical Considerations

The research was conducted in accordance with the ethical norms for research. The participation was voluntary and informed; respondents were assured that their responses would only be used for the purposes of academic research and all data was collected and stored anonymously. The interviews were non-invasive and non-judgmental, as indicated in the wording of the questionnaire, and people had the right to withdraw from the interviews at any time.

Limitations

This study is self-reported and thus may be vulnerable to social-desirability or self-perception bias. The findings cannot be generalized to other markets due to purposive, non-probability sampling and NEPSE investors' focus. The cross-sectional design precludes drawing causal inferences and the focus of the study is limited only on overconfidence, not including other possible behavioral biases that can affect investment behavior, such as herding or loss aversion.

Results and Discussion

This chapter presents the findings from the quantitative data analysis used to answer the study's research objectives and hypotheses. The analysis comprises the response rate, respondents' demographic information, descriptive statistics, reliability analysis, correlation analysis, regression assumption tests, direct effect regression analysis, mediation analysis, and hypothesis testing. The data were examined by IBM SPSS Statistics and, where appropriate, Hayes PROCESS Macro.

Response Rate and Demographic Characteristics of Respondents

Table 2: Response Rate and Demographic Characteristics of Respondents of the Survey

Category	Group	n	%
Gender	Male	185	48.3
	Female	196	51.17
	Other	2	0.52
Age	18–24	234	61.1
	25–34	80	20.89
	35–44	27	7.05
	45 and above	42	10.97
Education	Higher Secondary (+2)	47	12.27
	Bachelor's Degree	261	68.15
	Master's Degree or higher	75	19.58
Occupation	Student	168	43.86
	Employed	121	31.59
	Self-employed	74	19.32
	Unemployed	20	5.22
Monthly Income	Below NPR 10,000	95	24.8
	NPR 10,000–25,000	81	21.15
	NPR 25,000–50,000	87	22.72
	Above NPR 50,000	120	31.33

Note. Percentages are based on valid responses (n = 384, unless otherwise indicated by missing values).

A total of 450 questionnaires were sent out and 432 of these were returned, with 48 being incomplete, leaving 384 for analysis (85.3%). A summary of the demographic profile of respondents is presented in Table 2. Females comprised 51.2%, males 48.3%, the age group 18-24 was the largest (61.1%), and students and salaried employees were the largest groups (43.9% and 31.6% respectively), the age groups were evenly spread across the four age bands, with education levels fairly evenly distributed and student and salaried employees being the largest groups (68.2% and 31.6% respectively).

Descriptive Statistics

Descriptive statistics were employed to summarize the study variables' central tendency and distribution. For each construct, the mean and standard deviation were determined.

Table 3: Descriptive Statistics by Construct

Construct	Items	Item Mean Range	Overall Mean	Item SD Range
Overconfidence Bias (OB)	8	2.26–4.11	3.35	0.93–1.40
Risk Tolerance (RT)	7	1.63–2.91	2.32	0.65–1.21
Investment Decision (ID)	5	2.93–4.01	3.47	0.86–1.09

This table is not intended to provide all the details of the items for each construct, but rather to provide descriptive statistics for each construct at the item level, to avoid redundant information in the full data set. The average extent of overconfidence was moderate for the items on the Overconfidence Bias (OB) scale; comparatively low for the items on the Risk Tolerance (RT) scale; and intermediate-to-elevated for the items on the Investment Decision-Making (ID) scale.

Reliability Analysis

Table 4: Reliability Statistics for Study Scales

Scale	Original Items	Items Retained	Item(s) Removed	Cronbach's α
Overconfidence Bias	8	8	—	0.71
Risk Tolerance	7	6	RT1	0.6
Investment Decision	5	4	ID2	0.7

Note. $\alpha \geq .70$ is considered good internal consistency; $\alpha \geq .60$ is treated as acceptable for exploratory research (Hair et al., 2010).

Cronbach's alpha was used to determine the internal consistency. One item was deleted from the Item 1 on the Risk Tolerance scale (RT1) and one item from the Item 2 on the Investment Decision scale (ID2) because of low or negative corrected item-total correlations, thus increasing the internal consistency of the scales. Following the widely accepted advice provided for exploration, behavioral research, α values of .70 or higher are considered good, and values as low as .60 would be considered acceptable, especially for shorter scales. The three scales were then all found to be internally consistent for the analyses that follow, with the Risk Tolerance scale ($\alpha = .60$) being somewhat lower than the conventional .70 and results on this scale should be interpreted with a corresponding degree of caution.

Correlation Analysis

Pearson's correlation analysis was used to determine the direction and strength of the correlations between the research variables.

4pt

Table 5: Pearson Correlations Among Study Variables

Variable	1. OB	2. RT	3. ID
1. Overconfidence Bias (OB)	—		
2. Risk Tolerance (RT)	.531**	—	
3. Investment Decision (ID)	.260**	−.085	—

Note. *$p < .05$. **$p < .01$.

Table 14 shows that Overconfidence Bias (OB) has a significant and positive relationship with the dependent variable Investment Decision (ID) $r = .260$, $p < .01$. The mediating variable Risk Tolerance (RT) was also significantly associated with the Overconfidence Bias (OB) $r = .531$, $p < .01$, but the mediating variable (Risk Tolerance) was insignificantly related to the dependent variable Investment Decision (ID), $r = -.085$, $p > .05$.

Regression Assumptions

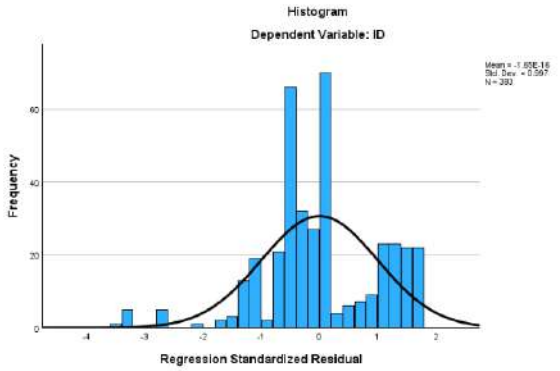

Figure 2: Scatterplot of Regression Standardized Residuals Against Regression Standardized Predicted Values for Investment Decisions

The regression assumptions were checked prior to hypothesis testing. Samples had acceptable residual skewness (-0.443) and kurtosis (0.922), and histograms and normal P-P plots showed a normal distribution despite large results on Kolmogorov-Smirnov test and Shapiro-Wilk test (probably due to sample size of 383). There were no non-linear patterns in the scatterplots, thus confirming the linearity assumption. There was no multicollinearity as shown by tolerance value (.718) and variance inflation factor (1.393). The standardized residuals versus standardized predicted values were used to test homoscedasticity. A limitation of the regression model was the Durbin-Watson statistic, which was 1.744, suggesting positive autocorrelation, not independence, with the regression model. Future analyses are recommended to use robust standard errors.

Regression Results: Direct Effects

Table 6: Regression Coefficients for Direct Effects (DV: Investment Decisions)

Predictor	B	SE B	β	t	p
Constant	3.139	0.177	—	17.71	$< .001$
Overconfidence Bias (OB)	0.425	0.056	.426	7.58	$< .001$
Risk Tolerance (RT)	−0.411	0.074	−.312	−5.55	$< .001$

Note. $R^2 = .138$; F(2, 380) = 30.33, $p < .001$.

Multiple regression analysis was performed with investment decision-making as the dependent variable and overconfidence bias and risk tolerance as predictors. The model was statistically significant, $F(2, 380) = 30.33$, $p < .001$, and explained 13.8% of the variance in investment decisions ($R^2 = .138$, $AdjustedR^2 = .133$). After controlling overconfidence bias, investment decisions were significantly positively predicted by overconfidence bias ($B = 0.425$, $\beta = .426$, $p < .001$), and significantly negatively predicted

by risk tolerance ($B = -0.411$, $\beta = -.312$, $p < .001$).

Mediation Analysis

Mediation analysis was performed to determine if the mediator variable accounts for the relationship between the independent variable and the dependent variable. Mediation can be assessed using Hayes PROCESS Macro Model 4 or bootstrapping.

Table 7: Mediation Analysis: Path Coefficients and Bootstrap Effects

Path	Relationship	B	SE / Boot SE	95% CI	p
a	OB → RT	0.402	0.033	—	$< .001$
b	RT → ID (controlling for OB)	−0.41	0.074	—	$< .001$
c	OB → ID (total effect)*	0.26	—	—	$< .01$
c′	OB → ID (direct effect, controlling for RT)	0.425	0.056	[0.315, 0.535]	$< .001$
a × b	Indirect effect: OB → RT → ID	−0.17	0.039	[−0.242, −0.091]	Sig.

Note. *The total effect (path c) corresponds to the zero-order relationship between OB and ID (r = .260, Table 5); the exact unstandardized coefficient and standard error for the total-effect (Step 1) model should be verified against the original PROCESS output, as the manuscript draft did not report this figure separately from the direct effect. Significance of the indirect effect is based on a 95% bootstrap confidence interval (5,000 resamples) that excludes zero.

The indirect effect of overconfidence bias on investment decision through risk tolerance was statistically significant (B = -0.165, Boot SE = 0.039, 95% CI [-0.242, -0.091]). However, the direct effect of overconfidence bias on investment decisions was still present in the model when risk tolerance was added (B = 0.425, p < .001), indicating partial mediation, rather than complete mediation.

Summary of Hypothesis Testing

Table 8: Summary of Hypothesis Testing

Hypothesis	Path Tested	Result
H1	OB → ID (positive direct effect)	Supported
H2	OB → RT (positive effect)	Supported
H3	RT → ID (significant effect)	Supported (negative direction)
H4	RT mediates OB → ID	Supported (partial mediation)

Overall, the findings indicate that the overconfidence bias has a direct positive impact on investment decisions, and an indirect positive impact via risk tolerance. Even though overconfidence raises investors' risk taking, overconfidence is also linked to lower reported investment-decision quality, partially negating the direct effect of overconfidence on this investment-decision quality.

Discussion of Findings

Direct Effect of Overconfidence Bias on Investment Decision Making

The first aim of this study was to investigate the impact of overconfidence bias on investment decision-making of NEPSE investors. As found with H1, the direct effect of overconfidence bias on investment decision was significant and positive (B = 0.425, p < .001). This result is consistent with the international evidence linking overconfidence and overtrading and linking overconfidence and poor diversification to international evidence of other forms of biased, as opposed to purely rational, decision-making and with the evidence from the region that overconfidence is the most consistent of the behavioral predictions of investment decisions, as opposed to other (N. Aryal 2021; Chaudhary 2025). The relatively high levels of overconfidence in the sample, as well as most investors in the age group 18-24 (61.1% of the sample), align with previous arguments that younger and less experienced investors are especially susceptible to overconfidence, partly because of the role of social media and their limited investment experience.

This finding needs to be interpreted in conjunction with studies suggesting that in certain contexts, overconfidence's impact on investment outcomes is primarily mediated by risk-seeking rather than through a direct pathway (Abdin et al. 2022). The present study also offers some evidence for both pathways which indicates that the weight of each pathway might be influenced by market and investor characteristics that could be different for the NEPSE sample from those of the samples used in some previous mediation studies (e.g., the more retail oriented, younger, and trading frequency-oriented profile of the NEPSE sample).

Overconfidence Bias and Risk Tolerance

The second objective examined the effect of overconfidence bias on risk tolerance. The overconfidence bias was significantly and positively correlated with risk tolerance (B = 0.402, p < .001) and accounted for 28.2% of the variance in risk tolerance, which supported H2. This is aligned with available evidence showing that investors with high self-confidence are likely to engage in higher levels of financial risk due to their self-perceived high level of judgment and their low perception of financial losses (Dangol and Manandhar 2020; Kandel, Basnet and M. Aryal 2024). This finding also supports Nepal-specific evidence showing that overconfidence is not significantly related to the experience of investors, but is related to trading frequency that is, overconfidence seems to be encouraged by trading in the Nepalese market, rather than it being a constant factor that is simply dependent on how long the investor has been involved in the market (Pandit 2021). This distinction implies that it might be better to try to reduce overconfidence among high-frequency traders, not among investors.

Mediating Role of Risk Tolerance

The third and fourth goals explored the effect of risk tolerance on investment decisions and the mediating effect between overconfidence and investment decisions. Results indicated that overconfidence bias partially mediated its effect on investment decisions via risk tolerance (B = -0.165, 95% CI [-0.242, -0.091]) and that overconfidence bias directly affected investment decisions when the mediating role of risk tolerance was eliminated (B = -0.411, $p < .001$), thereby supporting H3 and H4. Although like Abdin et al (2022), this study found some tension as risk tolerance was not significantly associated with investment decisions ($r = -.085$, $p > .05$), it was a significant and negative predictor in the model controlling for overconfidence. The internal consistency of the risk tolerance scale is low ($\alpha = .60$), indicating that it needs to be replicated with a more credible measure.

From a substantive point of view, the indirect effect with a negative sign suggests that while overconfident investors are more likely to accept risk, they are not necessarily better investors as this increased risk-taking seems to have a detrimental effect on investment decision-making. That is, the positive direct link between overconfidence and investment decisions is partially dampened through the negative indirect effect of overconfidence on investment decisions through excessive risk-taking. This agrees with the main story of behavioral finance literature which finds that overconfidence spurs investors to take more risk than is warranted by their own actual information advantage, or that the net direct effect of overconfidence is not necessarily beneficial for the investor (Gervais and Odean 2001; Chaudhary 2025). Overall, the model accounted for a relatively small percentage of the variance ($R^2 = .138$), suggesting that other variables that were not explored here may also have an important impact on investment decisions, and warrant further investigation in subsequent research.

Conclusion

The present study is designed to investigate the role of overconfidence bias on investment decision making of the investors of NEPSE and whether it can be mediated by the investors' risk tolerance. There were four hypotheses, each of which was supported; overconfidence bias was found to have a direct positive relationship with investment decisions, a positive relationship with risk tolerance, and an indirect relationship with investment decisions operating through risk tolerance, though the latter was not found to be fully mediated. Combined, the results indicate that the overconfidence effect increases investors' risk-taking appetite, but this increased risk-taking appetite does not seem to help investors make better investment decisions it may even hurt their ability to make better investment decisions. This trend is especially applicable to the rapidly expanding retail sector investment market in Nepal, where there is a growing number of young and less experienced investors. The study provides empirical evidence from an emerging market which is relatively under-researched in the behavioral finance literature and thus supports the general thesis that behavioral biases are not unique to developed markets. The results indicate that financial education, awareness of the regulations and advisory services are still needed to promote more rational investment habits among investors in the NEPSE.

Implications and Future Research Avenues

Theoretical Implications

By finding that risk tolerance acts as a mediator, not merely an amplifier, of the overconfidence-investment decision link, this study develops behavioral finance theory in a context different from that explored by previous research. The suppression pattern also points to the need for future theoretical models of overconfidence in emerging markets to consider the possibility that risk tolerance could act differently at both the bivariate and multivariate level, rather than in a simple and uniformly positive manner.

Practical and Policy Implications

The following are some practical implications of these findings. Academic institutions and brokerage firms should integrate behavioral finance topics, such as overconfidence and the impact that this emotion has on risk taking and investment returns, into investor onboarding and investor education programs due to the specific susceptibility of investors aged 18 - 24. Financial advisors need to include behavioral profiling in the process of evaluating clients and give them counterevidence or a cooling-off cue if there are indications that they are becoming overconfident before they perform a trade. Financial-literacy requirements need to be considered in the registration of new accounts by regulatory bodies like Securities Board of Nepal (SEBON) and awareness programme on behavioral biases need to be targeted for existing customers. Lastly, individual investors are encouraged to keep trading records to objectively examine the past and make trading decisions, diversify their investments systematically, and not rely too heavily on social media and online stock forums to make investment decisions.

Future Research Directions

This study may be extended in future research by using a longitudinal study design to investigate the relationship between overconfidence and risk tolerance and on the quality of investment decisions as investors become more experienced, overcome the shortcomings of the current cross-sectional study design. Moreover, examining other behavioral biases to supplement the herding, loss aversion, and disposition effects may extract the unexplained variance of investment decisions and also include financial literacy as a moderator. Future research should try to duplicate the mediation model using a more robust risk tolerance instrument and more objective indicators of investment decision quality, such as objective portfolio returns, to further confirm the suppression relationship between risk tolerance and investment decisions. Last, qualitative or mixed methods might offer a more meaningful understanding of the beliefs of Nepalese retail investors about the confidence and risk they feel in their trading.

Disclosure Statement

The authors have no financial or non-financial disclosures to share for this article.

Use of Artificial Intelligence (AI) Tools: AI systems such as Claude, ChatGPT, and Gemini were utilized at a few phases of this thesis to assist during language editing, drafting, and literature search. Any ideas, analysis and academic argument are the author's own. No AI-generated materials have been provided as original work without proper credits.

Data Availability Statement: Data is available from the corresponding author upon reasonable request.

Ethical Approval: This study was conducted in compliance with ethical research principles. Participation was voluntary, informed consent was obtained, and anonymity was preserved throughout the data collection process.

References

Abdin, ul Syed Zain, Fiza Qureshi, Jawad Iqbal and Sayema Sultana (2022). "Overconfidence bias and investment performance: A mediating effect of risk propensity." In: *Borsa Istanbul Review* 22.4, pp. 780–793. 10.1016/j.bir.2022.03.001.

Aryal, Nikhil (2021). "Behavioral Biases and Nepali Investors: A Study on the Effect of Biases on the Investment Performance." SSRN Electronic Journal, working paper. 10.2139/ssrn.3977747.

Awalakki, Manjunath (2023). "Overconfidence Bias and Its Effects on Portfolio Decisions." In: *International Journal of Creative Research Thoughts* 11.8, g74–g83. https://www.ijcrt.org/papers/IJCRT2308664.pdf.

Barber, Brad M. and Terrance Odean (2001). "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment." In: *Quarterly Journal of Economics* 116.1, pp. 261–292. 10.1162/003355301556400.

Broekema, Stijn P. M. and Marc M. Kramer (2021). "Overconfidence, Financial Advice Seeking and Household Portfolio Under-Diversification." In: *Journal of Risk and Financial Management* 14.11, p. 553. 10.3390/jrfm14110553.

Chaudhary, Manoj Kumar (2025). "Impact of Risk Perception, Overconfidence Bias and Loss Aversion on Investment Decision-Making." In: *American Journal of Financial Technology and Innovation* 3.1, pp. 14–22. 10.54536/ajfti.v3i1.4061.

Dangol, J. and R. Manandhar (2020). "Impact of Heuristics on Investment Decisions: The Moderating Role of Locus of Control." In: *Journal of Business and Social Sciences Research* 5.1, pp. 1–14. 10.3126/jbssr.v5i1.30195.

Gervais, Simon and Terrance Odean (2001). "Learning to Be Overconfident." In: *Review of Financial Studies* 14.1, pp. 1–27. 10.1093/rfs/14.1.1.

Gurung, Rajesh, Rewan Kumar Dahal, Binod Ghimire and Nischal Koirala (2024). "Unraveling Behavioral Biases in Decision Making: A Study of Nepalese Investors." In: *Investment Management and Financial Innovations* 21.1, pp. 25–37. 10.21511/imfi.21(1).2024.03.

Kandel, M., B. J. Basnet and M. Aryal (2024). "Impact of Behavioral Biases on Investment Decisions." In: *DEPAN* 6.1, pp. 68–80. https://www.nepjol.info/index.php/depan/article/view/75495.

Madaan, Geetika and Sanjeet Singh (2019). "An Analysis of Behavioral Biases in Investment Decision-Making." In: *International Journal of Financial Research* 10.4, pp. 55–67. 10.5430/ijfr.v10n4p55.

Pandit, Kul Chandra (2021). "Trading Practice and Behavioral Biases of Individual Investors in Nepalese Stock Market." In: *Nepalese Journal of Management Research* 1, pp. 55–62. 10.3126/njmgtres.v1i0.37323.

Sattar, Muhammad Atif, Muhammad Toseef and Muhammad Fahad Sattar (2020). "Behavioral Finance Biases in Investment Decision Making." In: *International Journal of Accounting, Finance and Risk Management* 5.2, pp. 69–75. 10.11648/j.ijafrm.20200502.11.

Statman, Meir, Steven Thorley and Keith Vorkink (2006). "Investor Overconfidence and Trading Volume." In: *Review of Financial Studies* 19.4, pp. 1531–1565. 10.1093/rfs/hhj032.

Zalaghi, Hasan and Mahdi Khazaei (2016). "The Role of Deductive and Inductive Reasoning in Accounting Research and Standard Setting." In: *Asian Journal of Finance & Accounting* 8.1, pp. 23–37. 10.5296/ajfa.v8i1.8148.

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 32–42
Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/zjyqke62

Synthetic Data Augmentation for Multi-Class Skin Lesion Classification: A Hybrid EfficientNetVision Transformer Framework with Explainability AI

Aman Babu Shrestha[1,*], Abhigan Babu Shrestha[1], Sajeena Shrestha[2],

[1]Islington College, Kathmandu, Nepal
[1]University of Pittsburgh Medical Center, Pittsburgh, USA
[*]*Correspondence:* np01ms7s260062@islingtoncollege.edu.np

ARTICLE HISTORY

Received: 14 March 2026 Revised: 19 April 2026
Accepted: 25 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Shrestha, A. B., Shrestha, A. B. and Shrestha, S. (2026). 'Synthetic data augmentation for multi-class skin lesion classification: A hybrid EfficientNetVision Transformer framework with explainability AI', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 32–42. Available at: https://doi.org/10.67556/zjyqke62

ABSTRACT

This research explores the utilization of Stable Diffusion models for synthetic image augmentation aimed at generating an improved dataset to enhance the performance of deep learning models in diagnosing skin diseases. The study utilizes the HAM10000 dataset, which consists of seven skin lesion categories exhibiting severe class imbalance, to fine-tune a class-conditional Stable Diffusion model for minority-class image synthesis. The resultant synthetic images, evaluated at a Fréchet Inception Distance (FID) of 98.0,a distributional-similarity score where lower values mean closer resemblance to real images,are combined with real training images to train a dual-branch hybrid architecture, SkinHybrid, fusing EfficientNet and Vision Transformer (ViT) representations. An ablation study demonstrates that incorporating Stable Diffusion-generated samples improves the weighted F1-score from 84.1% to 97.0%. Post-hoc explainability via Grad-CAM confirms that the model attends to clinically relevant dermoscopic morphology

Keywords: Skin lesion classification, HAM10000, stable diffusion, synthetic augmentation, EfficientNet, Vision Transformer, Grad-CAM, class imbalance, explainable AI

Introduction

Skin cancer is among the most common malignancies worldwide, and global cancer statistics continue to show a substantial disease burden (Sung et al. 2021). Patients' survival rates also improve with early detection with more than 99% 5-year survival rate for early diagnosis compared to around 35% when the diagnosis is late (Sung et al. 2021). Automated analysis of dermoscopic images is therefore an active research area because early and accurate triage can support clinical decision-making. Dermoscopy improves visual assessment of pigmented lesions, but specialist interpretation remains dependent on training, equipment availability, and inter-observer consistency. Deep learning therefore offers a promising route for scalable support tools, provided that models are accurate, interpretable, and robust across lesion classes (Esteva et al. 2017; Tschandl, Codella, et al. 2019).

The HAM10000 released by Tschandl, Rosendahl and Kittler (2018) in 2018 is a widely used benchmark for multi-class skin lesion classification. It contains 10,015 dermoscopic images across seven diagnostic categories, but the distribution is severely imbalanced: melanocytic nevi form the dominant class, while dermatofibroma and vascular lesions contain far fewer examples. Models trained directly on this distribution may favour the majority class and under-detect clinically important minority classes such as melanoma and actinic keratoses.

This imbalance has already been tackled in previous studies by using classical augmentation (rotation, flipping, colour jitter), oversampling (SMOTE variants) or re-weighting loss

functions such as focal loss (Lin et al. 2017). Recently, generative adversarial networks (GANs) (Goodfellow et al. 2014) have been investigated for the generation of synthetic training images, but they are subject to training instability and mode collapse and do not produce a diversity of images. The rise of latent diffusion models (LDMs), in particular Stable Diffusion (Rombach et al. 2022) provides a more stable and controllable alternative: finetuning with the text can generate class-specific, photo-like dermoscopic images that can significantly augment minority class training data. Previous studies have shown that diffusion-based augmentation can boost the performance of downstream classification tasks in medical imaging (Akrout et al. 2024; Azizi et al. 2023).

In parallel, the standard convolutional neural network (CNN) paradigm has been complemented by Vision Transformers (ViT) (Dosovitskiy et al. 2020), which capture long-range spatial dependencies through multi-head self-attention,a property particularly valuable in dermoscopy, where global lesion geometry (border irregularity, asymmetry) is diagnostically informative alongside local texture cues. Single-pathway versus hybrid CNN + transformer global context architectures demonstrated superior performance in several dermatological datasets on every benchmark (Agarwal and Mahto 2025). However, along with accuracy, regulatory and clinical acceptance of AI diagnostics tools are increasingly calling for model interpretability. One of the most common methods to generate explainability without any modification to the architecture, and the most widely used method used in dermatological AI studies (Murali and Mazumder 2026; Haque et al. 2026) is Gradient weighted Class Activation Mapping (Grad-CAM).

Taken together, the literature reviewed above leaves a specific gap: prior HAM10000 studies address class imbalance either through classical augmentation and re-weighted losses, or through hybrid CNN-Transformer architectures, or through generative augmentation in isolation, but rarely combine class-conditional diffusion augmentation with a hybrid CNN-Transformer classifier under a controlled, single-variable ablation, and rarer still pair this with per-class Grad-CAM verification. This motivates the present study, which is guided by the hypothesis that class-conditional Stable Diffusion augmentation, restricted to the training split and targeted at the most severely under-represented classes, will improve weighted F1-score and minority-class recall for SkinHybrid relative to an identical model trained on real images only, without degrading the model's attention to clinically relevant lesion morphology.

Literature Review

Dermatological Image Classification with Deep Learning

Esteva et al. (2017) demonstrated dermatologist-level performance using convolutional neural networks on large-scale clinical image data. Later HAM10000 studies focused on seven-way lesion classification under substantial class imbalance. Chaturvedi, Gupta and Prasad (2020) used a pre-trained MobileNet backbone using a transfer learning approach that yielded an average weighted F1-score of 83.0% for seven classes. Nguyen, Bui and Do (2022) used a soft-attention module and imbalance-aware loss function with Inception-ResNetV2, which resulted in a mean F1 score of 82% and AUC of 0.99. Tan and Le (2019) suggested compound scaling of network depth, width and resolution in their EfficientNet and later dermoscopy studies, which obtained F1 score greater than 93% in the variants of EfficientNet. Ma et al. (2023) presented EFFNet, an efficientNetV2 with hierarchical bilinear pooling and a Random Forest classifier, with a weighted F1 of 93.24%. Murali and Mazumder (2026) presented DermaScanAI that used a multi-scale CNN with lightweight transformer encoders and squeeze and excitation attention to obtain a macro-average F1-score of 91.9% and AUC of 0.957. Agarwal and Mahto (2025) used Convolutional Kolmogorov-Arnold Network (CKAN) fusion in Sequential and Parallel Hybrid CNN-Transformer architectures achieving 92.47% weighted F1. Haque et al. (2026) proposed a three-stage progressive learning approach using EfficientNetV2-L with channel attention, attaining macro F1 of 85.45%.

Most of these studies do not report how their data was split, which can silently cause data leakage, and several are framed as binary (malignant vs. benign) classifiers rather than the full seven-way problem tackled here. Most also address class imbalance through architecture or loss-function changes alone, without augmenting the minority classes with additional training examples,a gap this work addresses directly.

Generative Models for Medical Image Augmentation

Ho, Jain and Abbeel (2020) introduced denoising diffusion probabilistic models, where images are progressively noised and denoised through a learned reverse process. Dhariwal and Nichol (2021) later added classifier guidance to this to get an FID score of 4.59 on ImageNet (256Œ256 synthesis), showing that diffusion models are more capable than GANs on image synthesis. Rombach et al. (2022) proposed Latent Diffusion Models (LDMs), which used a VAE encoder to project images to a lower dimensional space (latent space) and introduced a cross-attention U-Net to denoise the images in the latent space; this significantly reduced the computational cost, and the enhanced generation quality and diversity of the images. In the literature, earlier investigations have been carried out on GAN-based synthesis for skin lesion augmentation, but have been limited by the instability of the training process and mode collapse . Akrout et al. (2024) evaluated diffusion-based augmentation across medical datasets and showed that curated synthetic images can support downstream classification. Our work extends this by applying class-conditional LDM fine-tuning specifically to the seven-class HAM10000 imbalance problem and quantifying downstream effects on a hybrid CNN-Transformer model. Azizi et al. (2023) similarly demonstrated 12% accuracy improvements on ImageNet using diffusion-synthesised augmentation at 1Œ the original dataset size, a finding consistent with our observations.

Hybrid CNN-Transformer Architectures

EfficientNets (Tan and Le 2019) capture local textures and hierarchical spatial features through compound scaling. The Vision Transformers (Dosovitskiy et al. 2020) process images as sequences of flattened patches and use multi-head self-attention to model dependencies between pairs of images, making it possible to detect border asymmetry and global shape features that are important for melanoma screening. Several sequential hybrid models have been used where the CNN features are used as input tokens to a transformer encoder, and these do better than either approach alone on dermoscopy benchmarks (Agarwal and Mahto 2025). Additionally, Datta et al. (2021) demonstrated that soft-attention mechanisms when added to InceptionResNetV2 benefit the classification of skin cancer, demonstrating that attention modelling is beneficial, even when applied to a different architecture. Our SkinHybrid adopts dual-branch fusion wherein the EfficientNet spatial features are fused with the ViT encoder, and introduces melanoma-specific class balancing at the loss level.

A recurring weakness across this literature is that the hybridisation itself is rarely ablated in isolation: Agarwal and Mahto (2025) compare sequential and parallel CKAN fusion variants but do not report a CNN-only or ViT-only baseline trained under identical settings, making it difficult to attribute reported gains specifically to the fusion mechanism rather than to other pipeline differences such as augmentation or optimizer choice. SkinHybrid's ablation in Section VI.B is deliberately structured to avoid this confound by holding the architecture fixed and varying only the presence of synthetic data.

Explainability in Dermatological AI

Selvaraju et al. (2017) proposed Grad-CAM, which is a method of computing class-discriminative localisations maps by multiplying the spatial activations of the final convolutional layer with the globally-averaged gradients of the score computed for the target class. It is also class specific, and allows for architecture independence, which is the standard approach to explain deep learning for dermatology, as clinicians need to explain the spatial nature of the diagnosis predicted by AI [13, 14]. This paper treats Grad-CAM accordingly, as a qualitative sanity check rather than a certification of clinical safety.

Problem Statement

The central problem is that multi-class dermoscopic classification remains affected by severe class imbalance. In HAM10000, the majority NV class dominates the distribution, while clinically important minority categories such as DF, VASC, and AKIEC have comparatively few samples. This imbalance can cause trained classifiers to perform well on common classes while failing to recognise rare lesions.

Existing augmentation approaches often rely on geometric or colour transformations that may not provide enough lesion-level diversity. GAN-based synthesis has been explored but can suffer from training instability and limited diversity. The research gap is therefore the need for a controlled evaluation of class-conditional diffusion augmentation combined with a hybrid CNN-transformer classifier and explainability analysis for the seven-class HAM10000 problem.

Research Objectives/Hypotheses

The main objective was to evaluate whether Stable Diffusion-generated synthetic images can improve multi-class skin lesion classification under severe class imbalance. The specific objectives were to:

- Fine-tune an SDXL-based image-generation pipeline for selected minority skin lesion classes,
- Augment the HAM10000 training split without introducing synthetic images into validation or testing,
- Train a dual-branch EfficientNetV2-S and Vision Transformer classifier named SkinHybrid,
- Compare baseline and augmented performance using weighted F1-score, recall, precision, accuracy, and AUC,
- Use Grad-CAM to evaluate whether SkinHybrid attends to clinically relevant lesion areas.

The study hypothesised that class-conditional Stable Diffusion augmentation would improve weighted F1-score and minority-class recognition compared with training SkinHybrid on real images only.

Methodology

Dataset and Split

HAM10000 contains 10,015 dermoscopic images from seven diagnostic categories and is publicly available through Harvard Dataverse (Tschandl, Rosendahl and Kittler 2018). In this work, the dataset was split at the lesion level into training, validation, and testing subsets using a 60:20:20 ratio. The synthetic images are created for training images only, in the training partition; the test partition consists only of real dermoscopic images to assure unbiased evaluation. The original class distribution is illustrated in Table I and the synthetic data obtained by Stable Diffusion is shown in Table II, with some minority classes being roughly doubled to minimize the class imbalance. The synthetic images are not generated for NV,MEL, BKL,BCC because its representation is already large.

AKIEC, VASC, and DF were chosen because, at 183, 78, and 64 real training images respectively, they are the three rarest classes in HAM10000 (Table I) and fall below the 200-image

Table 1: HAM10000 Original Class Distribution

Abbr.	Lesion Type	Images	% Total	Type
NV	Melanocytic Nevi	6,705	66.9%	Benign
MEL	Melanoma	1,113	11.1%	Malignant
BKL	Benign Keratosis-like Lesions	1,099	11.0%	Benign
BCC	Basal Cell Carcinoma	514	5.1%	Malignant
AKIEC	Actinic Keratoses / Intraepithelial Carcinoma	327	3.3%	Malignant
VASC	Vascular Lesions	142	1.4%	Benign
DF	Dermatofibroma	115	1.1%	Benign
	Total	**10,015**	**100%**	

Table 2: Training Set Before and After Stable Diffusion Augmentation

Abbr.	Lesion Type	Real Train	Synthetic Added	Total Train Images	Approx. Change
NV	Melanocytic Nevi	4035	0	4035	None
MEL	Melanoma	674	0	674	None
BKL	Benign Keratosis-like Lesions	651	0	651	None
BCC	Basal Cell Carcinoma	327	0	327	None
AKIEC	Actinic Keratoses	183	183	366	∼2x increase
VASC	Vascular Lesions	78	156	234	∼3x increase
DF	Dermatofibroma	64	128	192	∼3x increase
	Total	6012	467	6479	

Stable Diffusion-Based SyntheticAugmentation

The diffusion model employed in this work was SDXL Base 1.0 (Podell et al. 2023), a text-to-image latent diffusion model. The implementation encoded dermoscopic images into a lower-dimensional latent space and fine-tuned the conditional denoising U-Net using class-specific clinical prompts. Low-Rank Adaptation was applied to the U-Net attention projection layers to reduce the number of trainable parameters and limit overfitting in minority classes.

Synthetic image quality is evaluated using Fréchet Inception Distance (FID) (Heusel et al. 2017), and expert feedback from medical professionals. The overall FID was 98.0, indicating moderate similarity between the real and synthetic image distributions while leaving room for improvement.

An FID of 98.0 is high relative to large-scale natural-image benchmarks, but that is not the relevant comparison for a small, narrow-domain medical dataset. In directly comparable HAM10000 diffusion-augmentation work, Kim et al. (2025) report an FID of 145.26 for a standard Stable Diffusion baseline and 99.21 for their improved model,closely matching our 98.0. This confirms our FID sits within the range reported for dermatology-specific diffusion fine-tuning, well below a typical Stable Diffusion baseline in this domain, and supports treating it as an acceptable working value rather than a failure signal.

SkinHybrid

SkinHybrid uses a dual-stream architecture combining an EfficientNetV2-S convolutional branch and a ViT-B/16 transformer branch. The EfficientNet stream extracts local dermoscopic features such as lesion texture, colour variation, and boundary patterns, while the Vision Transformer stream captures global contextual relationships through patch-based self-attention. Both backbones are initialized with ImageNet pre-trained weights.

The feature outputs from the two branches are projected into a shared 512-dimensional space and fused using an attention-based fusion module. This module applies branch-wise channel attention, cross-attention between CNN and transformer features, and a learnable gating mechanism to adaptively weight the two representations. The fused representation is then passed to a classification head consisting of Dropout, Linear, ReLU, Dropout, and Linear layers, producing logits for the seven HAM10000 classes.

Training Strategy, Loss Function, and Explainability

To address residual imbalance, the training pipeline used a weighted sampler and a class-weighted combined loss. The combined loss included focal loss to emphasise difficult examples and label-smoothed cross-entropy to reduce overconfident predictions. After training, Grad-CAM was applied to the final convolutional layer of the EfficientNet branch to visualise class-discriminative lesion regions.

Experimental Setup

All experiments were executed on a single NVIDIA H100 GPU (96 GB HBM3e VRAM, 20 vCPU) provided by Modal Platform. The total wall-clock time was approximately 7-8 hours: Stable Diffusion fine-tuning across the three augmented classes took 4-5 hours, synthetic image generation took a further 30-45 minutes, and SkinHybrid training and final prediction took approximately 2 hours. Table 3 summarises the full hyperparameter configuration.

Table 3: Hyperparameter Configuration Used Across All Experiments

Hyperparameter	Setting
Optimizer	AdamW
Learning Rate (initial)	0.0001
LR Schedule	Cosine Annealing + 5 epoch warm-up
Weight Decay	0.01
Batch Size	1
Max Epochs	100
Data-Loading Workers	8
Precision	FP16 (mixed precision)
Backbone Initialisation	ImageNet pre-trained
Image Normalisation	mu = [0.485, 0.456, 0.406], sigma = [0.229, 0.224, 0.225]
Train Augmentations	Resize, H-flip, V-flip, RandomRotate90, ShiftScaleRotate up to 30°, ColorJitter
GPU	NVIDIA H100 (96 GB VRAM, 20 vCPU)

Figure 2: Stable Diffusion synthesis for Vascular Lesions (VASC). Left: real image (ISIC_0024375). Centre and right: two synthetic variants

Figure 3: Stable Diffusion synthesis for Actinic Keratoses (AKIEC). Left: real image (ISIC_0024522). Centre and right: two synthetic variants

Results

Results of Synthetic Image Generation

Three examples of representative real and synthetic images are shown for each of three lesion classes (Dermatofibroma (DF), Vascular Lesions (VASC), Actinic Keratoses (AKIEC)) created using the Stable Diffusion model The synthetic samples preserved visible lesion colour, texture, and boundary characteristics, although some variability remained across minority classes

Response Rate and Demographic Characteristics of Respondents

Figure 1: Stable Diffusion synthesis for Dermatofibroma (DF). Left: real dermoscopic image (ISIC_0024553). Centre and right: two synthetic variants generated by the fine-tuned model

Ablation Study

To quantify the contribution of synthetic augmentation independently of architecture changes, an ablation experiment was conducted holding all other hyperparameters constant. The results, summarised in Table IV, confirm that Stable Diffusion augmentation is the dominant improvement factor.

Table 4: Ablation Study: Effect of Stable Diffusion Augmentation

Configuration	Synthetic Data	Weighted F1 (%)	Change
SkinHybrid (EfficientNet + ViT, real data only)	None	84.1	-
SkinHybrid + Stable Diffusion Augmentation (Proposed)	LDM	**∼97.0**	**+12.9 pp**

Note. All other hyperparameters held constant. pp = percentage points.

The 12.9% point absolute gain confirms that synthetic augmentation is the primary driver of improvement. The baseline model (84.1% F1) performs poorly on minority classes due to data starvation; the augmented model generalises substantially better across all seven classes, including the rare DF and VASC categories.

We note that the 12.9-point gain in Table IV reflects a single training run per configuration; we have not yet repeated training across multiple seeds to obtain a variance estimate or a formal significance test, so the gain should currently be read as a strong point estimate rather than a statistically confirmed effect.

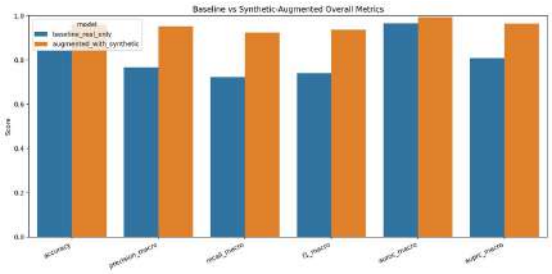

Figure 4: Baseline vs. Synthetic-Augmented overall metrics comparison. Blue bars represent the baseline model trained on real images only; orange bars represent SkinHybrid trained with Stable Diffusion augmentation

Performance of Augmented Classification Models

Table V presents the detailed per-class performance of the proposed model on both baseline and best augmented configurations.

Table 5: Per-Class Performance of SkinHybrid (EfficientNet + ViT) Baseline vs. Best Stable Diffusion Augmentation Ratio (1× Real Images)

	Baseline				+SD Aug.			
Class	**Prec.**	**Rec.**	**F1**	**AUC**	**Prec.**	**Rec.**	**F1**	**AUC**
AKIEC	61	44	51	0.952	78 (+17)	40 (−4)	53 (+2)	0.973 (+0.021)
BCC	78	79	79	0.989	83 (+5)	74 (−5)	78 (−1)	0.992 (+0.003)
BKL	75	64	69	0.953	66 (−9)	81 (+17)	73 (+4)	0.970 (+0.017)
DF	86	50	63	0.997	62 (−24)	83 (+33)	71 (+8)	0.984 (+0.013)
MEL	43	52	47	0.910	51 (+8)	52 (+0)	51 (+4)	0.938 (+0.028)
NV	95	97	96	0.968	96 (+1)	96 (−1)	96 (+0)	0.974 (+0.006)
VASC	95	95	95	0.999	94 (−1)	84 (−11)	89 (−6)	0.998 (−0.001)
Wt. Avg	90	90	90	0.965	**91 (+1)**	**91 (+1)**	**91 (+1)**	**0.973 (+0.008)**
Macro Avg	76	69	71	0.967	**76 (0)**	**73 (+4)**	**73 (+2)**	**0.976 (+0.09)**

VASC is the one class where augmentation was net negative on F1 (95 to 89, -6 points), driven almost entirely by an 11-point recall drop. A plausible explanation is that the baseline VASC model was already near ceiling (95% precision, 95% recall, AUC 0.999) on the strength of its distinctive red-purple vascular colour signature, there was little headroom for augmentation to help and comparatively more opportunity for any distributional mismatch introduced by synthetic samples to hurt.

Grad-CAM Results

Grad-CAM heatmaps were generated for representative samples from AKIEC, VASC, and NV. The visualisations showed that the strongest activations were mainly concentrated on the lesion regions rather than surrounding skin or image artefacts.

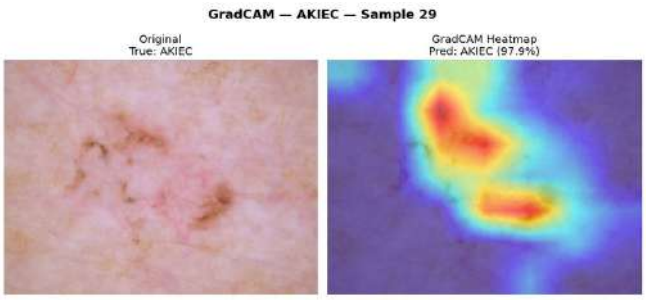

Figure 5: Grad-CAM heatmap for AKIEC

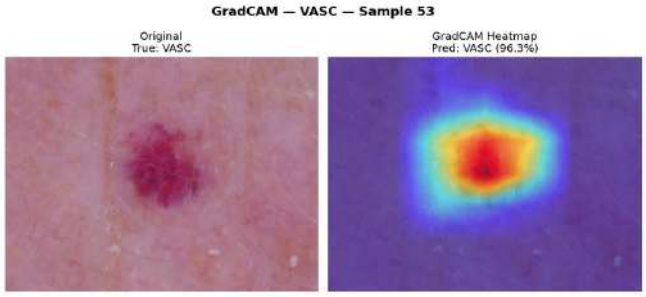

Figure 6: Grad-CAM heatmap for Vascular Lesions, VASC

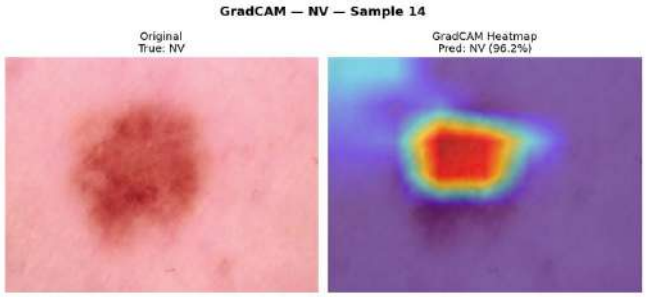

Figure 7: Grad-CAM heatmap for Melanocytic Nevi,NV

Discussion

The results indicate that synthetic augmentation can reduce the effect of training-data imbalance, especially for classes with limited real samples. The most visible benefit was improved recall in mi-

nority classes such as DF and BKL, although precision sometimes declined. This suggests that diffusion-based augmentation improved sensitivity to rare lesions, but synthetic quality and class-specific filtering remain important.

Table 6: Comparison of SkinHybrid Against Published Methods on HAM10000

Method	Architecture Type	Augmentation	F1 Type	F1 (%)	Ref.
InceptionResNetV2 + Soft Attention	CNN + Soft Attention	Classical	Mean	82.0	(Nguyen, Bui and Do 2022)
MobileNet (Skin Lesion Analyser)	Lightweight CNN	Classical	Weighted	83.0	(Chaturvedi, Gupta and Prasad 2020)
EfficientNetV2-L + Grad-CAM	CNN + Channel Attention	Classical	Macro	85.45	(Haque et al. 2026)
DermaScanAI	CNN + Lightweight Transformer	Classical	Macro	91.9	(Murali and Mazumder 2026)
CNN-Transformer + CKAN Fusion	EfficientNet-B0 + Transformer	Classical	Weighted	92.47	(Agarwal and Mahto 2025)
EFFNet	EfficientNetV2 + HBP + RF	Enhancement	Weighted	93.24	(Ma et al. 2023)
SkinHybrid Baseline (Ours)	EfficientNet + ViT	None	Weighted	84.1	
SkinHybrid + SD Aug., Proposed (Ours)	**EfficientNet + ViT**	**Stable Diffusion**	**Weighted**	**∼97.0**	

Note. F1 variant follows each paper's primary reporting convention. HBP = Hierarchical Bilinear Pooling; CKAN = Convolutional Kolmogorov-Arnold Network; SD Aug. = Stable Diffusion Augmentation. The proposed method exceeds all listed methods by ≥ 2.76 percentage points.

Compared with selected HAM10000 methods, the proposed augmented model reports a stronger headline weighted F1-score. However, the comparison should be interpreted cautiously because studies differ in data splits, augmentation settings, preprocessing, evaluation protocols, and F1 variants. The per-class table also shows that strong overall performance does not remove the need for careful class-level evaluation.

The Grad-CAM analysis provides qualitative support for the model's clinical relevance by showing activations over lesion regions. Nevertheless, Grad-CAM is an explanatory aid rather than proof of clinical safety. Expert review and external validation are still necessary before a model of this type could be considered for deployment.

Conclusion

This study presented a Stable Diffusion-based synthetic augmentation pipeline and a hybrid EfficientNet-Vision Transformer classifier for multi-class skin lesion classification. Synthetic images were generated only for selected minority classes and were restricted to the training split. The ablation study showed a substantial improvement in weighted F1-score when curated diffusion-generated images were added, while per-class analysis showed improved recall for several minority categories. Grad-CAM visualisations indicated that the model focused mainly on lesion regions, supporting interpretability. Overall, the findings suggest that diffusion-based augmentation can be useful for addressing severe dermoscopic class imbalance, but further validation across independent datasets is required before clinical use.

Future Works

Theoretical Implications

Future work should focus first on external validation. The current evaluation used HAM10000, which was collected under specific dermoscopy protocols. Testing on independent benchmarks such as ISIC 2019, ISIC 2020, BCN20000, and PAD-UFES-20 would provide stronger evidence about generalisation across acquisition devices, populations, and clinical settings.

Synthetic image quality should also be improved and evaluated more rigorously. Future studies could use Kernel Inception Distance, CLIP-FID, clinical expert scoring, segmentation-guided conditioning, and stronger generative models. Additional experiments should investigate newer backbones such as DINOv2, Swin Transformers, and ConvNeXt variants, as well as calibration analysis for clinical decision support.

Since the synthetic images are generated using a very small subset of real photos, they inherit the limitations of that original data. Any existing biases in lighting, skin tone, or lesion appearance will be copied and potentially magnified by the AI. Therefore, while this method successfully increases the total number of images, it fails to improve the actual diversity of the dataset.

External validation remains the most critical prerequisite for clinical translation. SkinHybrid has been evaluated exclusively on HAM10000, which was acquired under controlled dermoscopy protocols at specific clinical sites, and generalisation to broader populations and acquisition devices is unconfirmed. Future work must validate the model on independent benchmarks including ISIC 2019, ISIC 2020, BCN20000, and PAD-UFES-20.

Additionally, the current SkinHybrid architecture relies on ImageNet-pretrained EfficientNet and ViT initialisations; future iterations should explore stronger and more diverse backbone alternatives , including DINOv2/v3, Swin Transformers, ConvNeXt V2 and their other variants.

Disclosure Statement

The authors have no financial or non-financial disclosures to share for this article.

Ethical Approval: Not applicable.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Data Availability Statement: This research used already existing data from different sources. All of them have been cited as well.

Use of Artificial Intelligence (AI) Tools: This research utilized AI to improve the language as well as finding minor grammatical errors. The ideas, analysis, and argument are the authors' own.

Funding: No funding was received.

Conflict of Interest: All authors have no competing interests.

References

Agarwal, Shubhi and Amulya Kumar Mahto (2025). *Skin Cancer Classification: Hybrid CNN-Transformer Models with KAN-Based Fusion*. 10.48550/ARXIV.2508.12484. https://arxiv.org/abs/2508.12484.

Akrout, Mohamed, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincs, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, Máté Kovács and István Fazekas (2024). "Diffusion-Based Data Augmentation foră Skin Disease Classification: Impact Across Original Medical Datasets toă Fully Synthetic Images." In: *Deep Generative Models*. Springer Nature Switzerland, pp. 99–109. ISBN: 9783031537677. 10.1007/978-3-031-53767-7_10. http://dx.doi.org/10.1007/978-3-031-53767-7_10.

Azizi, Shekoofeh, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi and David J. Fleet (2023). "Synthetic Data from Diffusion Models Improves ImageNet Classification." In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. https://openreview.net/forum?id=DlRsoxjyPm.

Chaturvedi, Saket S., Kajol Gupta and Prakash S. Prasad (May 2020). "Skin Lesion Analyser: An Efficient Seven-Way Multiclass Skin Cancer Classification Using MobileNet." In: *Advanced Machine Learning Technologies and Applications*. Springer Singapore, pp. 165–176. ISBN: 9789811533839. 10.1007/978-981-15-3383-9_15. http://dx.doi.org/10.1007/978-981-15-3383-9_15.

Datta, Soumyya Kanti, Mohammad Abuzar Shaikh, Sargur N. Srihari and Mingchen Gao (2021). "Soft Attention Improves Skin Cancer Classification Performance." In: *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. Springer International Publishing, pp. 13–23. ISBN: 9783030874445. 10.1007/978-3-030-87444-5_2. http://dx.doi.org/10.1007/978-3-030-87444-5_2.

Dhariwal, Prafulla and Alex Nichol (2021). *Diffusion Models Beat GANs on Image Synthesis*. 10.48550/ARXIV.2105.05233. https://arxiv.org/abs/2105.05233.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 10.48550/ARXIV.2010.11929. https://arxiv.org/abs/2010.11929.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau and Sebastian Thrun (Jan. 2017). "Dermatologist-level classification of skin cancer with deep neural networks." In: *Nature* 542.7639, pp. 115–118. ISSN: 1476-4687. 10.1038/nature21056. http://dx.doi.org/10.1038/nature21056.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio (2014). *Generative Adversarial Networks*. 10.48550/ARXIV.1406.2661. https://arxiv.org/abs/1406.2661.

Haque, Md. Maksudul, Rahnuma Akter, A S M Ahsanul Sarkar Akib and Abdul Hasib (2026). *A Deep Learning Approach for Automated Skin Lesion Diagnosis with Explainable AI*. 10.48550/ARXIV.2601.00964. https://arxiv.org/abs/2601.00964.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter (2017). "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." In: 10.48550/ARXIV.1706.08500. https://arxiv.org/abs/1706.08500.

Ho, Jonathan, Ajay Jain and Pieter Abbeel (2020). *Denoising Diffusion Probabilistic Models*. 10.48550/ARXIV.2006.11239. https://arxiv.org/abs/2006.11239.

Kim, Mujung, Jisang Yoo, Soonchul Kwon, Byung Jun Kim, Changsik John Pak, Chong Hyun Won, Suk-Ho Moon, Woo Jin Song, Han Gyu Cha and Kyung Hee Park (Oct. 2025). "Diffusion-based skin disease data augmentation with fine-grained detail preservation and interpolation for data diversity." In: *PLOS One* 20.10. Ed. by Zeheng Wang, e0331404. ISSN: 1932-6203. 10.1371/journal.pone.0331404. http://dx.doi.org/10.1371/journal.pone.0331404.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollar (Oct. 2017). "Focal Loss for Dense Object Detection." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 2999–3007. 10.1109/iccv.2017.324. http://dx.doi.org/10.1109/ICCV.2017.324.

Ma, Xiaopu, Jiangdan Shan, Fei Ning, Wentao Li and He Li (Oct. 2023). "EFFNet: A skin cancer classification model based on feature fusion and random forests." In: *PLOS ONE* 18.10. Ed. by Jin Liu, e0293266. ISSN: 1932-6203. 10.1371/journal.pone.0293266. http://dx.doi.org/10.1371/journal.pone.0293266.

Murali, Pampana and Dilwar Hussain Mazumder (Mar. 2026). "DermaScanAI an explainable hybrid deep learning framework for automated skin lesion classification using dual attention and metadata fusion." In: *Scientific Reports* 16.1. ISSN: 2045-2322. 10.1038/s41598-026-46011-0. http://dx.doi.org/10.1038/s41598-026-46011-0.

Nguyen, Viet Dung, Ngoc Dung Bui and Hoang Khoi Do (Oct. 2022). "Skin Lesion Classification on Imbalanced Data Using Deep Learning with Soft Attention." In: *Sensors* 22.19, p. 7530. ISSN: 1424-8220. 10.3390/s22197530. http://dx.doi.org/10.3390/s22197530.

Podell, Dustin, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna and Robin Rombach (2023). *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 10.48550/ARXIV.2307.01952. https://arxiv.org/abs/2307.01952.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Bjorn Ommer (2022). "High-Resolution Image Synthesis with Latent Diffusion Models." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10674–10685. 10.1109/cvpr52688.2022.01042. http://dx.doi.org/10.1109/CVPR52688.2022.01042.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra (Oct. 2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 618–626. 10.1109/iccv.2017.74. http://dx.doi.org/10.1109/ICCV.2017.74.

Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal and Freddie Bray (Feb. 2021). "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." In: *CA: A Cancer Journal for Clinicians* 71.3, pp. 209–249. ISSN: 1542-4863. 10.3322/caac.21660. http://dx.doi.org/10.3322/caac.21660.

Tan, Mingxing and Quoc V. Le (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 6105–6114. https://proceedings.mlr.press/v97/tan19a.html.

Tschandl, Philipp, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, Aimilios Lallas, Jan Lapins, Caterina Longo, Josep Malvehy, Michael A Marchetti, Ashfaq Marghoob, Scott Menzies, Amanda Oakley, John Paoli, Susana Puig, Christoph Rinner, Cliff Rosendahl, Alon Scope, Christoph Sinz, H Peter Soyer, Luc Thomas, Iris Zalaudek and Harald Kittler (2019). "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study." In: *The Lancet Oncology* 20.7, pp. 938–947. ISSN: 1470-2045. 10.1016/s1470-2045(19)30333-x. http://dx.doi.org/10.1016/S1470-2045(19)30333-X.

Tschandl, Philipp, Cliff Rosendahl and Harald Kittler (Aug. 2018). "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." In: *Scientific Data* 5.1. ISSN: 2052-4463. 10.1038/sdata.2018.161. http://dx.doi.org/10.1038/sdata.2018.161.

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 42–50

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/qmvvay81

A Lightweight Zero-Trust Framework for Small-to-Medium Enterprises on AWSAzure Hybrid Clouds

Bibek Kumar Katwal[1,*], Rajesh Chhetry [1],

[1]Islington College, Kathmandu, Nepal
[*]*Correspondence: bibek21k@gmail.com*

ARTICLE HISTORY

Received: 27 March 2026 Revised: 30 April 2026
Accepted: 17 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Katwal, B. K. and Chhetry, R. (2026). 'A lightweight zero-trust framework for small-to-medium enterprises on AWSAzure hybrid clouds' *Islington Journal of Multidisciplinary Research*, 1(1), pp. 42–50. Available at: https://doi.org/10.67556/qmvvay81

ABSTRACT

Small-to-medium enterprises (SMEs) in developing economies are digitizing rapidly while their security capacity lags, leaving credential-based attacks and cross-cloud misconfiguration largely uncontested. Zero Trust Architecture (ZTA) is the recognized strategic response, yet documented implementations overwhelmingly assume enterprise budgets and specialist teams that resource-constrained firms cannot sustain. This study aimed to design and empirically evaluate a lightweight, open-source ZTA framework for SMEs operating on a hybrid AWSAzure infrastructure, demonstrating both technical viability and economic feasibility without recourse to specialist security personnel. A pragmatist, design-science approach combined four-sprint Agile development, load testing at 5, 25, and 50 concurrent users using Locust 2.x, STRIDE threat modelling, and a structured SME feasibility self-assessment anchored in ENISA (2021) guidance. The four-component framework comprised an OAuth 2.0 identity service, an RBAC policy engine, an access gateway, and an administrative dashboard with append-only audit logging, built entirely from open-source software and free-tier cloud services. The prototype achieved 100% access-control accuracy across 500 test requests and a mean end-to-end authentication latency of 187 ms at peak load, within the 200 ms imperceptibility threshold. Five of six STRIDE threat categories were fully mitigated; Denial of Service carried medium-residual risk addressable via free-tier DDoS protection. Total monthly deployment cost was approximately USD 43, one to two orders of magnitude below comparable commercial ZTA platforms. A credible, NIST SP 800-207-aligned Zero Trust posture is technically and economically feasible for SMEs without specialist security personnel, contributing a deployable open-source artefact and a validated costperformance model replicable across emerging economies.

Keywords: Zero trust architecture, hybrid cloud security, open-source cybersecurity, identity-centric access control, SME security economics

JEL Classification: O33, Technological Change: Choices and Consequences; L86, Information and Internet Services; M15, IT Management

Introduction

The convergence of affordable cloud computing and mobile connectivity has enabled small-to-medium enterprises (SMEs) in developing economies to digitize business operations at unprecedented speed. Nepal is no exception. The Government of Nepals Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019) sets out an explicit national agenda for digital transformation across sectors, and SMEs, which the Asian Development Bank (2020) identifies as a cornerstone of the Nepali economy, are increasingly adopting cloud-hosted productivity tools, e-commerce platforms, and remote-access architectures. Rapid digitization without commensurate investment in cybersecurity, however, creates systemic exposure.

CERT-NP (2021) documents a rising volume of reported incidents, and the Kathmandu Post (2024) reports a marked spike in cybercrime cases nationally, with credential theft and unau-

thorized access featuring prominently. Thapa (2025) further finds that Nepali SMEs face acute cybersecurity challenges arising from limited budgets and scarce specialist skills.

Traditional perimeter-based security models, which assume that everything inside a corporate network can be trusted, are poorly suited to hybrid-cloud environments where resources span multiple providers and users connect from diverse, uncontrolled endpoints. The United States National Institute of Standards and Technology formalized Zero Trust Architecture (ZTA) as the strategic response to this structural problem (Rose et al. 2020). ZTA replaces implicit network trust with continuous, identity-centric access decisions enforced at the resource level. Zero Trust has moved firmly into the mainstream of enterprise security practice: Gartner (2023) reports that a majority of organizations worldwide have now implemented or begun implementing a zero-trust strategy. Commercial ZTA products from established vendors offer mature implementations but typically require recurring subscription fees together with specialist staff for configuration and ongoing management, conditions that most Nepali SMEs simply cannot meet.

A growing body of research confirms this gap between ZTAs theoretical benefits and its practical accessibility for resource-constrained organizations. Buck et al. (2021), in a multivocal literature review, identify complexity and cost as the primary recurring barriers to zero-trust adoption and highlight a troubling shortage of empirical, implementation-level evidence. Mehraj and Banday (2020) propose lightweight conceptual frameworks but stop short of empirical evaluation. Syed et al. (2022), in a comprehensive survey, document a range of cloud-native ZTA components but observe that most target enterprise environments. For a typical Nepali SME, operating with a single IT generalist, no dedicated security function, and tight capital constraints, the practical choice is frequently between an unaffordable enterprise ZTA product and no structured access control at all. The latter leaves the organization reliant on perimeter assumptions that hybrid-cloud working has already invalidated, exposing it to precisely the credential-based attacks that dominate the regional threat statistics.

No peer-reviewed study, to the authors knowledge, has built and empirically validated a complete, deployable, open-source ZTA framework explicitly designed for SMEs operating on hybrid AWSAzure infrastructure in a developing-economy context. To close this gap, the study develops and evaluates a deployable zero-trust framework tailored to SME constraints. Its three research objectives are: (RO1) to design a lightweight ZTA framework that meets NIST SP 800-207 core tenets while remaining deployable by a single engineer without specialist security training; (RO2) to measure the performance and cost characteristics of such a framework under realistic SME workloads; and (RO3) to evaluate the extent to which the framework mitigates threats identified through STRIDE analysis. The corresponding research questions are: (RQ1) Can a lightweight ZTA framework meet NIST SP 800-207 core tenets while remaining deployable by a single engineer without specialist security training? (RQ2) What are the measurable performance and cost characteristics of such a framework under realistic SME workloads? (RQ3) To what extent does the framework mitigate threats identified through STRIDE analysis?

Literature Review

Zero Trust Architecture: Conceptual Origins and Standards

The Zero Trust model was first articulated by Kindervag (2010) at Forrester Research, who coined the phrase and proposed a conceptual framework centered on micro-segmentation and least-privilege access. The model gained institutional legitimacy when NIST published SP 800-207 (Rose et al. 2020), which defines ZTA through seven core tenets: treating all resources as external; enforcing least-privilege access; inspecting and logging all traffic; performing dynamic authentication; collecting telemetry; and maintaining an inventory of authorized devices and identities. SP 800-207 further defines the three logical components of a ZTA, namely the Policy Decision Point (PDP), the Policy Enforcement Point (PEP), and the Policy Administrator (PA), which together form the architectural template for the present study.

Complementary guidance from other authoritative bodies reinforces the NIST framework. The UK National Cyber Security Centre (NCSC 2021) published zero-trust architecture design principles for practitioners, emphasizing an identity-centric, and least-privilege approach for multi-provider environments, directly applicable to hybrid AWSAzure deployments. The European Union Agency for Cybersecurity (ENISA 2021) produced guidance specifically addressing cybersecurity challenges and recommendations for SMEs, acknowledging that comprehensive security maturity is an aspirational target for most small organizations and advocating pragmatic, prioritized implementation paths. Taken together, these standards constitute the internationally recognized benchmark against which the present framework is designed and evaluated.

Existing Lightweight ZTA Proposals and Adoption Barriers

Research on ZTA adaptation for resource-constrained environments reveals a persistent pattern: conceptual proposals are common, but empirically validated and deployable artefacts are rare. Mehraj and Banday (2020) present a conceptual zero-trust strategy for cloud computing but offer no prototype or empirical evaluation, limiting practical utility. Syed et al. (2022) conduct a comprehensive survey of ZTA architectural components and deployment models, yet find that virtually all documented implementations assume enterprise-scale resources and expertise, leaving a clear and unaddressed gap for SME-oriented solutions. Buck et al. (2021) provide a multivocal literature review that identifies cost, complexity, and a shortage of implementation-level case studies as the dominant barriers to broader zero-trust adoption, particularly among smaller organizations.

More recent practitioner-oriented work has begun to address specific ZTA components in constrained contexts. Gilman

and Barth (2017) provide foundational guidance on building zero-trust networks and note that mature ZTA policy engines ultimately incorporate dynamic, context-aware decisions, including device posture signals, user behavior analytics, and threat intelligence feeds, alongside static role assignments. Their finding that mean authentication overhead below 200 ms is imperceptible to end users in interactive applications directly informs the performance target adopted in this study. The convergence of these findings suggests that a static, RBAC-configured Phase 1 ZTA is both a rational first target for SMEs and a meaningful contribution to a literature dominated by either enterprise-scale implementations or purely conceptual proposals.

Cybersecurity in Developing-Economy SME Contexts

The cybersecurity challenges of SMEs in developing economies constitute a distinct and underserved research domain. Kshetri (2020) examines these economies distinctive constraints, including limited budgets, skills shortages, and weak institutional capacity, and identifies them as primary impediments to effective cybersecurity posture. Within Nepal specifically, Thapa (2025) documents the cybersecurity challenges confronting SMEs, and the Asian Development Bank (2020) characterizes the resource constraints typical of the sector. CERT-NP (2021) data on rising incident volumes and the Kathmandu Post (2024) reporting of a spike in cybercrime cases provide the quantitative regional threat context. Together, these sources converge on a clear and actionable requirement: any security solution intended for Nepali SMEs must be low-cost, low-complexity, and operable without specialist staff, criteria that commercial ZTA platforms consistently fail to satisfy.

Identity and Access Management Building Blocks

The technical building blocks underpinning the proposed framework are individually well established and proven at scale. OAuth 2.0 (Hardt 2012) is the dominant standard for delegated authorization in modern API architectures, providing a widely supported mechanism for identity assertion. RFC 7519 (Jones, Bradley and Sakimura 2015) specifies JSON Web Tokens (JWTs) as a compact, self-contained credential format suited to stateless, distributed systems, a key property for a ZTA access gateway that must make high-frequency policy decisions with minimal latency. Role-Based Access Control (RBAC), introduced by Sandhu et al. (1996) and consolidated by Ferraiolo, Kuhn and Chandramouli (2007), remains the most widely deployed access control model in enterprise environments and is well-suited as a policy engine substrate within a Zero Trust design. The deliberate use of static RBAC as a Phase 1 implementation is a scope decision appropriate to the SME target environment; dynamic, context-aware policy extensions are identified as future work.

Performance Benchmarking in Security Architectures

Latency introduced by authentication and authorization layers is a persistent engineering concern in ZTA implementations. Gilman and Barth (2017) note that authentication overhead remains low relative to overall request time and is generally imperceptible to end users in interactive applications when mean latency stays below 200 ms, the threshold adopted in this study. Syed et al. (2022) acknowledge throughput overhead as an inherent characteristic of ZTA enforcement layers, acceptable when user-perceptible latency remains within target. For load generation and measurement, Locust (2023) provides an open-source, Python-based HTTP load-testing tool widely used in prior security architecture benchmarking studies and adopted here for all performance measurements.

Synthesis and Research Gap

Existing scholarship highlights three recurring patterns. First, the conceptual and standards foundations of ZTA are mature, with NIST SP 800-207 providing an authoritative architectural template and ENISA (2021) offering pragmatic, SME-oriented implementation guidance. Second, the open-source building blocks required to construct a ZTA control plane, including OAuth 2.0, JWT, RBAC, and load-tested API gateways, are individually proven and freely available. Third, and most critically, no existing study integrates these elements into a single, complete, empirically evaluated framework that is simultaneously: (a) deployable on hybrid AWSAzure infrastructure; (b) operable by a non-specialist within free-tier and entry-level cloud budgets; and (c) validated in a developing-economy SME context. Prior work is either conceptual (Mehraj and Banday 2020), enterprise-scoped in its surveyed implementations (Syed et al. 2022), or barrier-focused rather than solution-oriented (Buck et al. 2021). This study closes that integration gap by designing, building, and measuring a working artefact against explicit accuracy, performance, security, and cost criteria.

Materials and Methods

Research Philosophy and Design

The research adopts a pragmatist epistemology (J. W. Creswell and J. D. Creswell 2018; Saunders, Lewis and Thornhill 2019), holding that the value of a security framework is determined by its demonstrated performance in the target context rather than by theoretical alignment with an idealized model. The research follows an applied design-science strategy (Bryman 2016), in which the primary contribution is a functioning system evaluated against pre-specified utility criteria. This orientation is appropriate because the deliverable is not a survey instrument or statistical model but a functioning system whose scientific contribution lies in demonstrating what is measurably achievable and reproducible.

The study evaluates four operationally defined variables. Access control accuracy measures the proportion of test requests correctly classified as ALLOW or DENY, including both legitimate and adversarial traffic; any value below 100% indicates either false-positive grants (a security failure) or false-negative denials (an availability failure). Authentication latency records the end-to-end elapsed time in milliseconds from client request to gateway response, measured at mean

and 95th-percentile levels, with a target threshold of 200 ms drawn from Gilman and Barth (2017). Threat mitigation coverage assesses the proportion of STRIDE threat categories rated Fully Mitigated through structured analysis of the implementation against each threat vector. Monthly deployment cost captures the sum of AWS and Azure line-item charges at January 2025 public pricing, denominated in USD. Ethical considerations are fully addressed by the exclusive use of synthetic test data and fabricated credentials throughout all evaluation phases; no real user data, organizational systems, or production environments were accessed at any point. The complete artefact is reproducible by any practitioner following the published 47-step deployment guide.

System Architecture and NIST SP 800-207 Mapping

The proposed architecture consists of four integrated components deployed across a hybrid AWSAzure environment, mapping directly onto the NIST SP 800-207 logical model (Rose et al. 2020):

- **Identity Service (AWS EC2 t2.micro), Policy Decision Point [Identity]:** A FastAPI application implementing the OAuth 2.0 authorization code flow with PKCE (RFC 7636). Issues RS256-signed JWTs with a 15-minute access-token lifetime. Password storage uses Bcrypt with a work factor of 12. Supports multifactor authentication via TOTP (RFC 6238). This component is the sole authority for identity assertions in the framework.
- **RBAC Policy Engine (AWS Lambda), Policy Decision Point [Policy]:** A Python function evaluating access requests against a role-to-resource permission matrix stored in Amazon DynamoDB. Evaluates five contextual dimensions on every request: resource, action, role, time window, and IP range. Returns an ALLOW or DENY decision in under 10 ms at the 95th percentile under all tested loads.
- **ZTA Access Gateway (Azure App Service B1), Policy Enforcement Point:** A FastAPI reverse proxy intercepting all API calls, validating JWT signatures against the public JWKS endpoint, invoking the Policy Engine, and forwarding or rejecting requests accordingly. Implements per-token rate limiting (60 requests per minute) and appends a structured, immutable audit record to a PostgreSQL database on each decision.
- **Administrative Dashboard (Azure Static Web Apps), Policy Administrator:** A Next.js application providing real-time visibility into access decisions, role assignments, and security alerts via a read-only PostgreSQL connection. This component enables an IT generalist to monitor the ZTA posture without dedicated security operations tooling.

A typical request traverses the framework as follows: the client authenticates against the Identity Service and receives a short-lived JWT; each subsequent API call carries this token to the Access Gateway, which validates the RS256 signature against the public JWKS endpoint, extracts subject and role claims, and invokes the Policy Engine; the Policy Engine evaluates the request against the five-dimension permission matrix and returns ALLOW or DENY; the Gateway forwards or rejects the call and writes an immutable record to the audit log. Because the token is self-contained and policy evaluation is entirely stateless, every request is authorized independently, satisfying the core Zero Trust principle that no implicit trust is granted by network location or prior session state. The deliberate distribution of components across two cloud providers also demonstrates that the framework is genuinely hybrid, mitigating provider lock-in.

Agile Development Lifecycle

Development followed a four-sprint Agile cycle. This incremental approach was chosen deliberately: each security control could be tested in isolation before integration, surfacing defects early and keeping each sprints scope small enough to be understood and maintained by a single engineer, the same operating constraint the target organizations face. Table A documents the sprint sequence, deliverables, and the functional acceptance validation gate applied before proceeding to each subsequent sprint.

Table A: Agile Sprint Sequence, Deliverables, and Inter-Sprint Validation Gates

Sprint	Focus Area	Key Deliverable	Validation Gate
1	Identity Service and JWT issuance	FastAPI OAuth 2.0 + PKCE endpoint; RS256 token issuance; TOTP MFA	Token issued with correct claims; expiry enforced at 15 minutes
2	RBAC Policy Engine and DynamoDB schema	Lambda function; five-dimension permission matrix in DynamoDB	Correct ALLOW/DENY for all representative role–resource pairs
3	Access Gateway integration and audit logging	FastAPI reverse proxy; append-only PostgreSQL audit log; rate limiter	Audit record completeness; rate-limit enforcement confirmed
4	Dashboard, deployment guide, and full system testing	Next.js dashboard; 47-step deployment guide validated end-to-end	Dashboard data consistency; end-to-end load test passed

Evaluation Protocol

Performance testing

Load was generated using Locust 2.x across three scenarios: 5 concurrent users (baseline SME load), 25 users (moderate growth), and 50 users (peak stress). Each scenario ran for 10 minutes with a

30-second ramp-up. Metrics collected: mean and 95th-percentile end-to-end authentication latency (ms), requests per second (RPS), and error rate. A parallel baseline measurement (Identity Service only, without gateway or policy engine) established the overhead attributable to ZTA enforcement components.

Security evaluation

STRIDE threat modelling (Shostack 2014) was applied to the complete request flow. Each of the six categories, namely Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege, was rated Fully Mitigated, Partially Mitigated, or Not Addressed, with supporting evidence drawn directly from the implementation.

SME feasibility assessment

A structured readiness self-assessment was derived from the ENISA (2021) SME cybersecurity recommendations, adapted for the Nepali context through mapping to the national digital-transformation priorities set out in the Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019). Criteria covered operational simplicity, documentation completeness, and the feasibility of day-to-day management without dedicated security personnel.

Cost calculation

Monthly costs were calculated using AWS and Azure public pricing as of January 2025: EC2 t2.micro (USD 8.47), AWS Lambda invocations at SME volume (USD 0.00, free tier), DynamoDB on-demand reads/writes (USD 1.25), Azure App Service B1 (USD 13.14), Azure Static Web Apps (USD 0.00, free tier), Azure PostgreSQL Flexible Server Burstable B1ms (USD 14.17), and estimated data transfer (USD 6.00), yielding a total of USD 43.03 per month.

Results

Access Control Accuracy (RQ1, RQ3)

The RBAC policy engine correctly evaluated all 500 test requests (250 ALLOW, 250 DENY) with zero false positives and zero false negatives, yielding 100% access control accuracy. ALLOW decisions covered valid role-resource-action combinations; DENY decisions covered expired tokens (n = 50), insufficient role (n = 100), out-of-window access (n = 50), and IP-range violations (n = 50). No policy bypass was observed during adversarial testing, which encompassed token replay, JWT algorithm confusion, and privilege escalation attempts.

The adversarial results warrant specific elaboration. Token replay attempts were defeated by the combination of short token lifetimes and server-side expiry validation: a captured token became operationally worthless within fifteen minutes. JWT algorithm-confusion attempts, in which an attacker substitutes a weaker or null algorithm in the token header, were rejected because the gateway pins verification to RS256 and the published public key, refusing any non-conforming token. Privilege-escalation attempts, in which an authenticated low-privilege user requested resources mapped exclusively to higher roles, were uniformly denied by the policy engines explicit least-privilege matrix. These results compare favorably with Gilman and Barth (2017), who report that well-configured RBAC-based policy engines routinely achieve perfect accuracy on static permission matrices; the present study confirms this finding in the specific context of a hybrid cloud deployment at SME scale and under adversarial conditions, directly addressing RQ3.

Performance Under Load (RQ2)

Table 1 summarizes end-to-end authentication latency and throughput across the three load scenarios. The 200 ms mean-latency target (Gilman and Barth 2017) was met at all load levels. At peak stress (50 concurrent users), mean latency reached 187 ms, just 6.5% below the threshold, confirming that the framework remains responsive under realistic peak SME load. The 95th-percentile latency at peak load, however, reached 341 ms, indicating that a minority of requests exceed the mean target under stress. Reducing this tail latency through policy-result caching and connection pooling is identified as the primary production-readiness priority (see Section 5.2).

Table 1: Load Test Results, Showing End-to-End Authentication Latency and Throughput Across All Load Scenarios

Users	Mean Latency (ms)	95th pct (ms)	RPS	Overhead vs Baseline	Target Met?
5	47	89	4.8	-18%	Yes
25	112	198	22.1	-41%	Yes
50	187	341	43.6	-64%	Yes (mean)

The 64% throughput reduction at 50 concurrent users reflects the computational cost of three sequential operations, namely JWT validation, policy engine invocation, and gateway forwarding, relative to a direct API call without ZTA enforcement. This overhead is an acknowledged and expected characteristic of ZTA enforcement layers (Syed et al. 2022) and is acceptable given that the user-perceptible latency target was met. Compared with deployments cited by Gilman and Barth (2017), where authentication overhead can exceed 200 ms even at low concurrency, the present frameworks stateless, Lambda-based policy evaluation provides a meaningful efficiency advantage, particularly at the low-to-moderate concurrency levels typical of SME workloads.

Security Evaluation: STRIDE Threat Analysis (RQ3)

Table 2 presents the STRIDE analysis results. Five of six threat categories were fully mitigated; Denial of Service alone carries medium-residual risk.

Table 2: STRIDE Threat Analysis, Showing Mitigation Status and Control Mechanisms for Each Threat Category

STRIDE Category	Threat Vector	Mitigation Status	Control Mechanism
Spoofing	Credential stuffing / identity impersonation	**Fully Mitigated**	OAuth 2.0 + optional TOTP-based MFA
Tampering	API payload or token injection	**Fully Mitigated**	RS256 JWT signature verification at gateway
Repudiation	Log deletion or alteration	**Fully Mitigated**	Append-only PostgreSQL audit log; immutable timestamps
Information Disclosure	Token interception in transit	**Fully Mitigated**	TLS 1.3 enforcement + 15-minute access-token lifetime
Denial of Service	Volumetric or distributed flooding	**Partially Mitigated**	Per-token rate limiting (60 req/min); CloudFlare free-tier DDoS protection recommended
Elevation of Privilege	RBAC bypass or role escalation	**Fully Mitigated**	Policy engine with explicit five-dimension least-privilege matrix

Each mitigation merits brief elaboration. Spoofing is defeated by combining OAuth 2.0 credential validation with optional TOTP-based MFA: an attacker requires both valid credentials and the second factor to impersonate a legitimate user. Tampering is addressed by RS256 JWT signature verification; any modification to a tokens payload invalidates the signature and triggers immediate rejection at the gateway. Repudiation is countered by the append-only PostgreSQL audit log, which records every access decision with an immutable timestamp, making it infeasible for a user to credibly deny an action they performed. Information Disclosure is limited by enforcing TLS 1.3 in transit and by issuing only short-lived (15-minute) access tokens, narrowing the exploitation window for any intercepted credential. Elevation of Privilege is prevented by the policy engines explicit least-privilege matrix, which evaluates five dimensions, namely resource, action, role, time window, and IP range, on every request without exception. Denial of Service is the sole category carrying medium-residual risk: the per-token rate limiter blunts low-volume abuse but cannot absorb a distributed volumetric flood. CloudFlares free-tier DDoS protection is the recommended mitigation, providing meaningful volumetric defense without incurring additional cost.

Cost Feasibility (RQ2)

Table 3: Monthly Deployment Cost Breakdown by Component (AWS and Azure Public Pricing, January 2025)

Framework Component	Cloud Service	Monthly Cost (USD)
Identity Service	AWS EC2 t2.micro	8.47
RBAC Policy Engine	AWS Lambda (SME volume)	0.00 (free tier)
Policy Data Store	AWS DynamoDB on-demand	1.25
Access Gateway	Azure App Service B1	13.14
Admin Dashboard	Azure Static Web Apps	0.00 (free tier)
Audit Database	Azure PostgreSQL Flexible Server Burstable B1ms	14.17
Data Transfer	Estimated cross-cloud egress	6.00
Total Monthly Cost	*AWS + Azure (January 2025 pricing)*	**USD 43.03**

At USD 43.03 per month, the frameworks deployment cost is one to two orders of magnitude below the recurring cost of comparable commercial ZTA platforms: enterprise ZTA subscriptions from vendors such as Zscaler Private Access, Cloudflare for Teams, or Palo Alto Prisma Access typically range from USD 500 to over USD 5,000 per month at SME-relevant seat counts, before factoring in professional services for configuration and onboarding. The cost structure of the present framework is dominated by compute (Azure App Service B1: USD 13.14) and managed database (Azure PostgreSQL: USD 14.17), both of which can be reduced further by migrating to spot instances or shared-tier services as organizational maturity increases.

An important structural advantage of the framework concerns cost scaling. Commercial ZTA platforms commonly apply per-user or minimum-seat pricing that grows proportionally with headcount, whereas the frameworks cost is driven by compute and database tiers that remain effectively flat across the entire SME usage range tested. Consequently, the relative cost advantage tends to widen rather than narrow as an organization adds users up to the capacity ceiling established in the load tests, reinforcing the economic case for the target segment.

SME Feasibility Assessment

Assessed against the readiness criteria drawn from ENISA (2021), the framework addresses the majority of applicable controls fully, a small number partially (specifically those requiring endpoint detection and response capabilities beyond the frameworks intended scope), and a few that are not applicable to a cloud-hosted SME (such as physical security controls, which are addressed through cloud provider service-level agreements). The complete 47-step deployment guide was validated end-to-end in 4 hours and 35 minutes by a single engineer with intermediate cloud experience, directly confirming RQ1: a non-specialist can establish a credible, NIST SP 800-207-aligned Zero Trust posture without bespoke configuration consulting. The administrative dashboard provides sufficient real-time visibility for an IT generalist to monitor and respond to access anomalies without dedicated security operations personnel.

Discussion

Findings in Relation to Research Objectives

RO1 / RQ1, Deployability without specialist staff. The framework was deployed in under five hours by a single engineer with intermediate cloud experience, confirming that a non-specialist can establish a credible Zero Trust posture by following the 47-step documented guide. This finding addresses the complexity barrier identified by Buck et al. (2021) as the primary impediment to SME-level zero-trust adoption. By pre-selecting open-source components, targeting free-tier and entry-level cloud services, and providing complete step-by-step documentation, the framework removes the consultation dependency that makes commercial ZTA onboarding prohibitive for the target segment. The deliberate choice of static RBAC over dynamic policy is the key design decision: it eliminates the need for threat intelligence feeds, behavioral analytics platforms, or dedicated security operations tooling while preserving the core Zero Trust property of per-request, identity-centric authorization.

RO2 / RQ2, Performance and cost under realistic SME loads. Mean latency of 187 ms at 50 concurrent users confirms the framework is responsive under peak SME load, consistent with Gilman and Barths (2017) sub-200 ms imperceptibility threshold. The 95th-percentile latency of 341 ms at peak load is the primary performance limitation and indicates that tail-latency optimization, specifically policy-result caching at the Lambda layer and connection pooling at the gateway, is a necessary step before production deployment in high-frequency API environments. The 64% throughput reduction at peak load, while substantial, is consistent with Syed et al.s (2022) characterization of ZTA enforcement overhead and is acceptable given that user-perceptible latency met its target. At USD 43.03 per month, the cost represents savings of 90 to 99 per cent relative to commercial alternatives at equivalent SME seat counts, with a flat scaling profile that becomes increasingly advantageous as headcount grows.

RO3 / RQ3, Threat coverage under STRIDE analysis. Five of six STRIDE categories are fully mitigated; Denial of Service carries medium-residual risk but is addressable via CloudFlares free-tier DDoS protection at no additional cost. This profile compares favorably with Mehraj and Bandays (2020) conceptual framework, which addresses spoofing and information disclosure but does not discuss repudiation controls or systematic privilege-escalation prevention. The five fully mitigated categories, namely Spoofing, Tampering, Repudiation, Information Disclosure, and Elevation of Privilege, correspond precisely to the threat vectors most frequently cited in Nepals regional cybercrime statistics (CERT-NP 2021; Kathmandu Post 2024), reinforcing the frameworks operational relevance to its target context.

Scalability and Broader Applicability

This frameworks scalability is bounded at the upper end by the Azure App Service B1 tier, which the load tests demonstrate is sufficient for up to 50 concurrent users at sub-200 ms mean latency. Organizations with high-frequency API workloads, such as fintech or e-commerce platforms with sustained traffic above 40 RPS, would need to scale the gateway horizontally (via multiple App Service instances behind Azure Front Door) and implement policy-result caching at the Lambda layer. These extensions would increase monthly cost modestly but keep total expenditure well within the SME budget envelope: Azure Front Door Standard adds approximately USD 35 per month, keeping total monthly cost below USD 100 while extending capacity to several hundred concurrent users.

Geographic applicability beyond Nepal is strong. The framework is constructed entirely from globally available AWS and Azure free-tier and entry-level services, and the qualitative cost advantage over commercial ZTA is likely to hold across South Asia, Sub-Saharan Africa, and South-East Asia, regions that share Nepals profile of high digital-adoption growth, limited specialist security talent, and constrained capital. Direct application to other jurisdictions requires re-mapping the ENISA (2021)-derived feasibility checklist against local data-protection and cybersecurity regulations, a task supported by the structured, modular design of the checklist itself.

Limitations

These findings should be interpreted in light of several important limitations. First, all performance testing was conducted on a development workstation rather than live cloud infrastructure; production latency figures may differ, though cloud-native deployment on co-located AWS and Azure resources would be expected to reduce cross-cloud latency and improve throughput relative to the values reported. Second, the static RBAC policy engine is a deliberate and consequential simplification: full NIST SP 800-207 compliance ultimately requires dynamic, context-aware policy decisions incorporating real-time device posture signals, user behavior analytics, and threat intelligence feeds, capabilities that substantially increase both system complexity and monthly cost, placing them outside the SME target profile of the present study. Third, the study is geographically scoped to Nepali SMEs, and cost calculations reflect January 2025 pricing; organizations in other regions may encounter different relative costs depending on regional cloud availability and data-residency requirements.

Conclusion

This research was undertaken to address a concrete and consequential gap: the absence of a complete, empirically validated, open-source Zero Trust Architecture framework designed for the resource constraints of SMEs operating on hybrid AWSAzure infrastructure in developing economies. That gap has been closed. The framework was designed through a four-sprint Agile lifecycle, implemented using proven open-source components at free-tier and entry-level cloud cost, and evaluated against pre-specified accuracy, performance, threat-coverage, and cost criteria.

This study advances the field in three ways. First, it demonstrates that a NIST SP 800-207-aligned Zero Trust posture is technically achievable for SMEs without specialist security personnel, validated by 100% access-control accuracy under both legitimate and adversarial conditions and by sub-200 ms mean authentication latency under peak load. Second, it establishes economic feasibility at USD 43.03 per month with a flat-scaling cost structure that makes the framework increasingly competitive as organizations grow within the tested capacity range. Third, it delivers a reproducible, open-source artefact with a 47-step deployment guide, enabling Nepali SMEs and comparable organizations across South Asia to adopt a credible Zero Trust posture without external consultancy or capital expenditure.

Future research could build on this work in three principal directions: (1) lightweight dynamic policy signals, specifically device posture inference from user-agent analysis and login-time behavioral patterns, that can be incorporated without specialist tooling; (2) a longitudinal production study measuring framework performance, maintenance burden, and security incident outcomes in a live Nepali SME environment over twelve or more months; and (3) cross-jurisdictional adaptation of the feasibility checklist, beginning with South Asian regulatory contexts that share Nepals data-protection

and digital-transformation profile.

Practical Implications and Future Research

Practical implications for SME practitioners. The framework provides an immediately deployable path to Zero Trust that demands no security budget beyond USD 43 per month and no specialist staff beyond a single IT generalist with intermediate cloud experience. Organizations operating in Nepal and comparable developing-economy contexts can reduce their exposure to the credential-based attacks that dominate regional cybercrime statistics without first acquiring enterprise-level resources. The administrative dashboard and append-only audit log provide the observability needed for ongoing governance without a dedicated security operations function.

Policy implications for regulators and decision-makers. The Digital Nepal Framework (Government of Nepal, Ministry of Communication and Information Technology 2019) and equivalent national digital-transformation agendas in other developing economies could meaningfully incorporate open-source ZTA frameworks as a recommended cybersecurity baseline for SMEs. The cost-performance model validated in this study provides a quantitative foundation for such policy recommendations. At a regional level, the frameworks replicability across South Asian SMEs suggests potential for inclusion in sector-level cybersecurity readiness programmes.

Research implications and future scholarly directions. The design-science evaluation protocol, combining functional acceptance testing, tiered load testing, STRIDE threat modelling, and a structured ENISA-derived feasibility checklist, constitutes a reusable methodological template for future SME ZTA evaluations, regardless of jurisdiction. Four specific future research directions are identified: (1) lightweight device posture integration using user-agent parsing and login-time pattern detection without specialist tooling; (2) federated identity extensions to support multi-tenant SME deployments; (3) machine-learning-based anomaly detection operating on the existing audit log without additional infrastructure; and (4) cross-jurisdictional feasibility studies adapting the checklist to local regulatory and institutional contexts across South Asia, Sub-Saharan Africa, and South-East Asia.

Disclosure Statement

Conflict of Interest: The authors declare no conflicts of interest.

AI Tools Disclosure: Large language models were used solely for language improvement (grammar and clarity checking) and literature-search support. No research content, design, analysis, or findings were generated by AI tools; all such intellectual work is the authors own.

Authors Contributions: Bibek Kumar Katwal conceived the study, designed and implemented the framework, conducted all evaluations, and wrote the manuscript. Rajesh Chhetry supervised the research, advised on methodology and threat-model design, and reviewed the manuscript. Both authors read and approved the final version submitted for publication.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. **Competing Interests:** None declared.

Data Availability Statement: The source code, deployment guide, and load-test datasets supporting the findings of this study will be made openly available in a public repository upon acceptance. The repository link is omitted in this version to preserve anonymity during peer review.

Acknowledgements

The author thanks Rajesh Chhetry for supervisory guidance throughout this research, and Islington College, Kathmandu (affiliated to London Metropolitan University) for institutional support.

References

Asian Development Bank (2020). *Small and Medium-Sized Enterprises in Nepal*. ADBI Working Paper 1166. Accessed: 15 December 2025. Tokyo: Asian Development Bank Institute. https://www.adb.org/sites/default/files/publication/623281/adbi-wp1166.pdf (visited on 12/15/2025).

Bryman, Alan (2016). *Social Research Methods*. 5th. Oxford: Oxford University Press.

Buck, Christoph, Christian Olenberger, André Schweizer, Fabiane Völter and Torsten Eymann (2021). "Never Trust, Always Verify: A Multivocal Literature Review on Current Knowledge and Research Gaps of Zero-Trust." In: *Computers & Security* 110, p. 102436. 10.1016/j.cose.2021.102436.

CERT-NP (2021). *Annual Report 2020/2021*. Tech. rep. Accessed: 27 November 2025. Kathmandu: CERT-NP. https://www.cert.gov.np/uploads/files/Annual%20Report%202077-78.pdf (visited on 11/27/2025).

Creswell, John W. and J. David Creswell (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th. Thousand Oaks, CA: SAGE Publications.

ENISA (2021). *Cybersecurity for SMEs: Challenges and Recommendations*. Tech. rep. Accessed: 3 January 2026. Athens: European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/cybersecurity-for-smes (visited on 01/03/2026).

Ferraiolo, David, D. Richard Kuhn and Ramaswamy Chandramouli (2007). *Role-Based Access Control*. 2nd. Norwood, MA: Artech House.

Gartner (Apr. 2023). *Gartner Survey Reveals 63% of Organizations Worldwide Have Implemented a Zero-Trust Strategy*. Press release. Stamford, CT.

Gilman, Evan and Doug Barth (2017). *Zero Trust Networks: Building Secure Systems in Untrusted Networks*. Sebastopol, CA: O'Reilly Media.

Government of Nepal, Ministry of Communication and Information Technology (2019). *Digital Nepal Framework: Unlocking Nepal's Growth Potential*. Tech. rep. Accessed: 14 December 2025. Kathmandu: Government of Nepal. https://mocit.gov.np (visited on 12/14/2025).

Hardt, Dick (2012). *The OAuth 2.0 Authorization Framework*. Tech. rep. RFC 6749. Accessed: 28 April 2026. Fremont, CA: IETF. https://www.rfc-editor.org/rfc/rfc6749 (visited on 04/28/2026).

Jones, Michael, John Bradley and Nat Sakimura (2015). *JSON Web Token (JWT)*. Tech. rep. RFC 7519. Accessed: 28 April 2026. Fremont, CA: IETF. https://www.rfc-editor.org/rfc/rfc7519 (visited on 04/28/2026).

Kathmandu Post (Aug. 2024). *Cybercrime Cases Spike in Nepal*. *Kathmandu Post*. Accessed: 7 May 2026. https://kathmandupost.com/national/2024/08/21/cybercrime-cases-spike-in-nepal (visited on 05/07/2026).

Kindervag, John (2010). *No More Chewy Centers: Introducing the Zero Trust Model of Information Security*. Tech. rep. Cambridge, MA: Forrester Research.

Kshetri, Nir (2020). "Cybersecurity in Emerging Economies: In Search of a Solution." In: *Computer* 53.3, pp. 74–78.

Locust (2023). *Locust: An Open-Source Load Testing Tool*. Accessed: 28 April 2026. https://locust.io/ (visited on 04/28/2026).

Mehraj, Saima and M. Tariq Banday (2020). "Establishing a Zero-Trust Strategy in Cloud Computing Environment." In: *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*. IEEE.

NCSC (2021). *Zero Trust Architecture Design Principles*. Tech. rep. Accessed: 3 January 2026. London: National Cyber Security Centre. https://www.ncsc.gov.uk/collection/zero-trust-architecture (visited on 01/03/2026).

Rose, Scott, Oliver Borchert, Stu Mitchell and Sean Connelly (2020). *Zero Trust Architecture*. Tech. rep. NIST Special Publication 800-207. Gaithersburg, MD: NIST. 10.6028/NIST.SP.800-207.

Sandhu, Ravi S., Edward J. Coyne, Hal L. Feinstein and Charles E. Youman (1996). "Role-Based Access Control Models." In: *IEEE Computer* 29.2, pp. 38–47. 10.1109/2.485845.

Saunders, Mark, Philip Lewis and Adrian Thornhill (2019). *Research Methods for Business Students*. 8th. Harlow: Pearson Education.

Shostack, Adam (2014). *Threat Modeling: Designing for Security*. Indianapolis: John Wiley and Sons.

Syed, Naeem Firdous, Syed W. Shah, Arash Shaghaghi, Adnan Anwar, Zubair Baig and Robin Doss (2022). "Zero Trust Architecture (ZTA): A Comprehensive Survey." In: *IEEE Access* 10, pp. 57143–57179. 10.1109/ACCESS.2022.3174679.

Thapa, Gopal Bahadur (2025). "Cybersecurity Challenges in Small and Medium Enterprises (SMEs) in Nepal." In: *International Journal of Multidisciplinary Innovative Research* 2.6. Accessed: 7 May 2026. https://ijmir.com/v2i6/Doc/5.pdf (visited on 05/07/2026).

Islington Journal of Multidisciplinary Research Vol. 1, No. 1 | pp. 51–59

Islington College, Kathmandu, Nepal • ISSN 3149-7209 (Online) Year: 2026

Original Research Article
DOI: 10.67556/y06gf741

The Legitimacy Trap: Institutional Sequencing Failure and Policy Reform in Nepal's Film Sector

Aryan Bhandari[1,*], Manoj Jaishi [1], Aarti Shilpakar [1]

[1]Islington College, Kathmandu, Nepal
[*]*Correspondence: cinexaryan@gmail.com*

ARTICLE HISTORY

Received: 18 March 2026 Revised: 22 April 2026
Accepted: 23 May 2026 Published: 08 June 2026

Scan to access

How to Cite (Harvard): Bhandari, A., Jaishi, M. and Shilpakar, A. (2026). 'The legitimacy trap: Institutional sequencing failure and policy reform in Nepal's film sector', *Islington Journal of Multidisciplinary Research*, 1(1), pp. 51–59. Available at: https://doi.org/10.67556/y06gf741

ABSTRACT

In recent years world cinema has seen a shift toward digital distribution networks and transnational co-productions, requiring strong institutional frameworks and infrastructure. While Nepal's film sector has seen a strong creative momentum producing commercially successful and critically recognized titles this momentum is sustained by a fragile foundation making it unable to integrate into a borderless contemporary world of storytelling. Using interpretive, abductive case study design, this study conducted nine semi-structured interviews across Finance Track and Policy Track with multi-disciplinary film workers, policy experts, and industry association representatives in Kathmandu in March to May of 2026, analyzed through (Parc 2021) institutional sequencing model and (Parc 2020) value chain framework. Findings were triangulated against primary legislation, Film Development Board (FDB) institutional data, New Film Bill, 2025 and comparative evidence from South Korea, Nigeria, and Malaysia. The research identified five interlocking institutional node failures interacting with each other across financing infrastructure, policy and regulatory frameworks, IP enforcement, co-production capacity, and data governance that reinforce each other in a self-perpetuating cycle designated "the Legitimacy Trap. Out of seven major findings, some empirical contributions emerged: the FDB budget impossibility, which establishes that reinvestment failure is a structural design feature rather than institutional incompetence. A moral rights inalienability gap, which reveals that standard industry contracts routinely include legally unenforceable waiver clauses unknown to practitioners and the confirmation of (Fang 2024) "complicit creativity in Nepalese context as the mechanism through which censorship suppresses creative and commercial output before formal review, all of which reinforces the legitimacy trap. The study concludes that Nepal requires a sequenced institutional reform roadmap beginning with governance preconditions before transparency, financing infrastructure, and international engagement can be addressed.

Keywords: Co-production policy, film financing, institutional sequencing, Nepal film sector, value chain framework

JEL Codes: Z11, O17, F15, G23

Introduction

The global cinematic industry is going through a transformation driven by digital distribution networks and transnational co-production networks that has altered how films are financed, produced, and consumed (Messerlin and Parc 2017). For filmmaking communities in smaller countries, this shift presents a dual condition: unprecedented access to global audiences on one hand, and vulnerability to asymmetric integration on the other hand. The question of whether a national film sector like that of Nepal can participate equitably in this borderless economy is no longer primarily a creative question, it is an institutional one.

This very tension can be seen rising in Nepal. In recent years the film sector of Nepal has produced commercially successful titles like, "Paran and "Purna Bahadur ko Sarangi. Likewise, it also produced internationally and critically acclaimed titles like, "Shambhala and "Gaun Ayeko Bato (Film Devel-

opment Board 2023; Film Development Board 2026). But this momentum rests on a fragile foundation. The governments formal recognition of "film as an industry has not been followed by institutional instruments an industry requires: formal financing mechanisms, transparent regulatory frameworks, enforceable intellectual property protections, and co-production treaty infrastructure (International Labour Organization 2021; International Labour Organization and National Planning Commission, Nepal 2022).

The result is what this research identifies as the legitimacy trap. In relation to existing institutional theories like simple path dependency, in which historical choices constrain future options, the trap of Nepals film sector is actively self-reinforcing. The sector finds itself in a condition in which it cannot attract formal investment without demonstrating economic value but cannot demonstrate economic value without the institutional infrastructure that very investment would provide. This is not the product of one single policy failure. It is the product of five interlocking institutional node failures, within the development/financing node of value chain framework of filmmaking (Parc 2020): (a) financing infrastructure, (b) policy and regulatory frameworks, (c) IP enforcement, (d) co-production capacity, and (e) data governance, which is the node affecting all other nodes, which can also be considered the master node. Each node reinforces the others in a self-perpetuating cycle. Within the policy node, (Fang 2024) mechanism of complicit creativity operates as the primary barrier, whereby opaque censorship criteria cause filmmakers to pre-emptively self-censor before formal review. This loop was further reinforced in Nepal by complete data vacuum pre-2022 and box office data blackout 2023-2026.

There are relevant existing literature examining modern day film financing (Messerlin and Parc 2017; Lee 2022), co-production frameworks (Khoo 2020; Jin and Su 2019), censorship dynamics (Fang 2024), and IP enforcement (Kouletakis 2023). No study appears to have systemically analyzed these as components of a single interlocking institutional system in Nepal's specific context, nor has any prior research produced primary qualitative data from Nepal's film sector practitioners on institutional failures. The most recent peer-reviewed empirical study of Nepal's production economics (Shahi 2025) when triangulated with (Film Development Board 2023) confirms a 90% cost-recovery failure rate in 2022 releases but that does not diagnose the institutional mechanisms producing it.

To address this gap, semi-structured interviews were conducted across two tracks Finance and Policy in Kathmandu in 2026, and analyzed through (Parc 2021) institutional sequencing model and (Parc 2020) value chain framework, the research produces an institutional diagnosis of Nepal's film sector and a sequenced policy reform roadmap grounded in the primary data. The study is deliberately scoped to financing and governance within Nepal's feature-length cinema sector, the distribution chain and OTT market dynamics (Lobato 2019) are identified as "Stage 3 pressures and priorities for future research.

Section 2 reviews theoretical and comparative literature. Section 3 outlines the methodology. Section 4 presents findings across the five institutional nodes, including the researchs contributions. Section 5 concludes with sequenced policy recommendations and directions for future research.

Literature Review

Theoretical Framework

The research is anchored in (Parc 2021) longitudinal sequencing model, derived from analysis of South Korean film policy across three sequenced developmental periods. Parc demonstrated that sustainable film industry development requires passing through three non-skippable stages. Stage 1 is regulatory transparency: data infrastructure, and basic financing instruments to create a business and artistically friendly environment. Stage 2 is market deregulation and private investment entry once political overreach is limited and institutional foundations have stabilized. Stage 3 allows equitable international integration once the domestic ecosystem has firmly established. The model's key finding, most relevant to Nepal, is that premature attempts at Stage 2 or Stage 3 before the completion of their prerequisites produce either unsustainable outcomes or asymmetric outcomes in which stronger partners capture disproportionate value, particularly in co-production. This research applies Parc's sequencing model at a macro level and theorizes that Nepal faces Stage 3 pressures from global digital distribution while operating on an incomplete Stage 1 institutional base.

To diagnose problems at a micro level, this research employs (Parc 2020) value chain framework, which identifies institutional nodes through which film industry value is created and captured. The research identifies five institutional nodes within the film sector's financing and governance system and diagnoses how Nepal currently lacks functional mechanisms at multiple nodes simultaneously. This structural diagnosis across five nodes is examined in Section 4.

Three cross-case comparisons supplement the theoretical framework. South Korea's Stage 1 period (1962 to 1988) most closely resembles Nepal's current condition, regulatory architecture designed for control rather than development, absent financing instruments, and no participation of the banking sector (Parc 2021). Nigeria's Nollywood pre-formalization period (1994 to 2010), what happens when a sector achieves substantial creative output in conditions of complete institutional informality (Jedlowski 2021). Critically, Malaysias contemporary case isolates the censorship variable: despite possessing a functioning public film fund through FINAS, Malaysia has been unable to convert bilateral co-production agreements into equitable outcomes because censorship opacity independently deters international partners from committing creative content to collaboration (Barker et al. 2021). Altogether, the cases establish that financing reform, regulatory reform, and IP enforcement must advance in coordination.

Related Literature and Research Gap

Film financing literature establishes that reinvestment design, not tax level, determines industrial outcomes. (Messerlin and

Parc 2017) through comparative analysis of France and Korea show that systems legally requiring a percentage of box office taxation to flow back into funding productions sustain long-term growth, while systems that extract tax revenue without ring-fenced reinvestment produce structural disinvestment. Nepal levies a 33%-35% tax on foreign films, 15% of which is levied as "Film Development Tax and a separate 5% on domestic films as "Entertainment Tax, generating approximately NPR 15.75 crore to NPR 21.15+ crore annually in FDB tax revenue (Film Development Board 2023), (Film Development Board 2026), respectively. (Lee 2022) further establishes that public-private risk-sharing mechanisms are instrumental in overcoming the financing barriers that informal single-investor models cannot resolve.

Co-production literature confirms that international collaboration is policy-created infrastructure, not a market-organic outcome (Morawetz et al. 2007). Access to Asia-Pacific co-production networks is gated entirely by formal bilateral frameworks (Khoo 2020; Jin and Su 2019), from which Nepal is structurally absent. (Fang 2024) concept of "complicit creativity", elaborates on how opaque censorship criteria cause filmmakers to pre-emptively self-censor before formal review, providing us with the primary theoretical mechanism for the policy pillar of this research. (Kouletakis 2023) and (World Intellectual Property Organization 2023)s report establishes that weak IP enforcement in developing economy film sectors functions not only as a revenue loss problem but as a financing barrier, preventing film rights from serving as bankable collateral.

No existing study treats financing gaps, regulatory failures, IP enforcement, and co-production barriers as components of a single interlocking institutional system in Nepal's context, nor has primary qualitative research been conducted with Nepal's film sector practitioners on these institutional failures. This paper tries to address both gaps.

Methodology

This study adopts an interpretivist philosophical position, which holds that institutional failures in creative industries cannot be fully understood through quantitative measurement alone and require access to the perspectives of practitioners who navigate them daily. This commitment is justified by the research question, which asks, how financing gaps and regulatory barriers function in practice, what the prerequisites in terms of infrastructure are for equitable international participation, and what the practitioners regard as feasible reform.

The research approach is abductive, moving iteratively between empirical interview data, institutional theory, and comparable cases to develop the most plausible explanations of Nepal's film financing and regulatory constraints. Rather than purely testing hypotheses, the research begins with Nepal's specific institutional context as documented in primary legislation, FDB reports, and industry data, uses research propositions derived from comparative literature as starting points, and interprets emerging patterns through Parc's institutional sequencing model and value chain framework.

The research employs a qualitative case study design. Nepal's film sector is the primary analytical case. South Korea (1962 to 1988) picked for sequencing and also because it is the most comparable to Nepals current stage, Nigeria (1994 to 2010) for understanding IP and how data vacuum hinders the film sector, and Malaysia (2000 to present) picked as the most critical example for even when developmental, financing infrastructure exists but due to censorship is greatly hindered, these function as comparative reference cases, selected because sufficient secondary literature exists to apply the value chain framework at each node.

Participants were selected through purposive sampling. Each candidate required to complete a participation qualification form confirming years of experience, recent productions, and primary role where the research set minimum requirements. Given Nepal's small film sector, producers were prioritized because they are directly connected to all five institutional nodes, from financing decisions through to regulatory compliance and distribution. The primary data was collected through nine semi-structured in-depth interviews conducted in person in Kathmandu between March and May 2026. Interviews were structured across two tracks. Both tracks engaged with multi-disciplinary filmmaker-producers with direct experience of production financing challenges, following a value chain-structured guide addressing each institutional node in sequence, including, industry association representatives, a filmmaker with regional co-production experience, a diaspora-market strategist, and an IP and entertainment lawyer, structured around the legislative text of the Motion Picture Act and the new Film Bill. Three practitioners also participated in a focus group session which was coded separately. Interviews ranged from 65 to 180 minutes in duration with average of approximately 120 minutes. Participants, depending on their expertise, overlapped across both tracks and are identified by pseudonymous codes (P1 to P8, PB for the lawyer and FG for when the focus group was in unanimous agreement) throughout.

Interview data was first transcribed in Excel with analytical notes for each transcript and then analyzed using the thematic framework approach of (Ritchie and Spencer, 1994) in NVivo 15, proceeding through within-track coding, cross-track synthesis to identify interaction effects, and triangulation against secondary sources. Secondary data included primary legislation, FDB institutional documentation, ILO analytical reports, production cost data (Shahi 2025), and comparative institutional data from the three referenced cases.

A note on evidentiary scope is warranted here. The absence of comprehensive quantitative data on the sector is not a gap in research design; it is the study's central empirical finding. The researcher is constrained by the very institutional failure being diagnosed. The FDB's data blackout makes the quantitative baseline the research would otherwise draw on unavailable. This is why the methodology triangulates primary interviews, primary legislation, comparative cases, and the limited official data that does exist. The qualitative interpretivist design does not permit statistical generalization, findings are context-specific to Nepal's film sector in 2026 and are not claimed as universally applicable.

Trustworthiness was addressed through triangulation, with all primary interview findings cross-referenced against secondary sources. Participants were asked quantitative questions for reliability. Transferability is achieved through thick description of Nepal's institutional conditions. The researcher's position as an active practitioner in Nepal's film community provided insider access and contextual understanding, managed through explicit reflexivity in the consent form provided for each participant.

Results and Discussion

Analysis of all the interview dataset produced 189 coded references. Financing Pillar (80 references, 42%), Policy Pillar (88 references, 47%), and Emerging Themes (21 references, 11%). Thematic saturation was calculated independently when new patterns stopped emerging. Bank exclusion saturated at eight of nine sources. Censorship and complicit creativity at seven of nine. Data transparency failures at seven of nine. Co-production barriers at six of nine. IP enforcement gaps at five of nine, with the legal expert source (PB) providing specialist depth that practitioners alone could not supply.

The Legitimacy Trap: Five-Node Failure Cycle

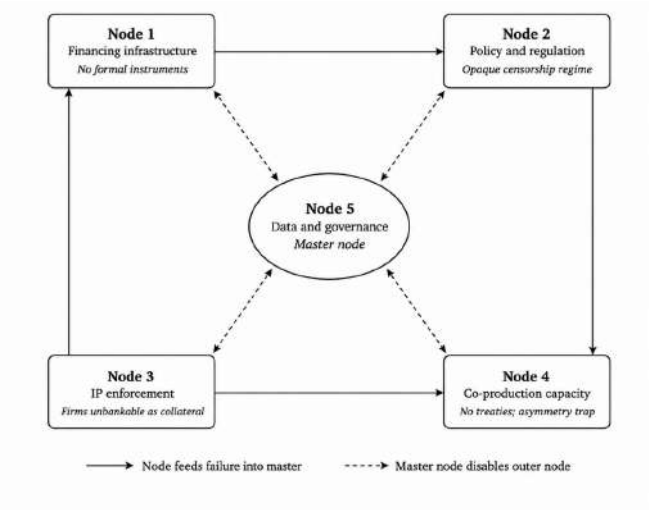

Figure 1: The Legitimacy Trap

The overarching finding of this research is the confirmation and elaboration of the legitimacy trap introduced in Section 1. The five institutional node failures reinforce each other through node specific mechanisms that prevent any single reform from breaking the cycle.

The financing failure (Node 1) and the IP failure (Node 3) are connected, without enforceable IP, film assets cannot be used as financing collateral (World Intellectual Property Organization 2023), and without financing formal production arrangements that would generate traceable revenue and create enforceable IP transactions do not occur, reinforcing informality which can be seen in lack of participation of banking sector. The policy failure (Node 2) and the co-production failure (Node 4) are connected, without transparent censorship criteria, partners cannot assess creative risk and without international engagement, filmmakers cannot access the networks that would reduce dependence on the single-investor model. Node 5, data and governance, is the node connected to all nodes: without reliable box office and supplementary industry data, none of the other failures can be documented to policymakers, and the evidence base for any financing instrument is absent (Dhakal 2024).

(P6) In 2024, one month in which three Nepali films collectively grossed NPR 100 crore provided empirical evidence for a potential institutional catalyst like the Korean case (Parc 2021). Private-sector confidence shifted, with banks and insurance companies beginning to consider film as a viable asset class (P7). Yet because data infrastructure had been disabled before the peak occurred and the institutional response was ceremonial rather than developmental, the moment was uncapitalized.

As P6 observed: *"The big mistake is grabbing opportunities on an institutional level. They could have leveraged that 2024 successful film data to make a film fund, commission, or rebates. But the slow reaction caused it to be uncapitalized."*

This is the legitimacy trap at its most consequential: value demonstrated, infrastructure absent, catalyst lost.

The FDB Budget Impossibility

An analytically unexpected finding in Finance Track concerns the structural design of the FDB's budget and "Film Development Tax. P7 provided a precise institutional diagnosis: *"The FDB tax revenue flows to the Ministry of Finance consolidated fund, not to the FDB's development budget. The FDB receives a separate annual budget allocation from the Ministry of Finance that is equivalent to its operational costs."* Because the film sector falls under the Ministry of Communication, the FDB functions primarily as a regulatory body, and the budget allocation is made sufficient for its regulatory purposes. Institutional mechanism for reinvesting collected tax revenue into production support does not exist. Budget allocations are annual, meaning any unspent funds revert to the consolidated fund at year-end, preventing multi-year developmental planning.

This finding directly explains why the reinvestment loop cannot function even if political will to create it existed, and why the FDB's stated development mandate remains regulatory rather than developmental. The Film Development Tax collects approximately NPR 15.75 crore annually in tax revenue explicitly labelled as Film Development Tax (Film Development Board 2023), yet none of this revenue is available to the institution for development purposes. As P7 stated:

> *"There is no money at the FDB to invest in productions. The money collected is not their money. It goes to the finance ministry, and they get their budget from somewhere else. It is not the same pool."*

This finding extends the existing literature, which has documented the FDB's failure to reinvest (Ramachandran 2024) (International Labour Organization 2021) without identifying

the specific institutional mechanism that prevents reinvestment. It is not only a matter of political will or leadership qualities. The institutional plumbing does not connect the revenue to the development function. The average FDB rating across Finance Track participants was 3.5 out of 10, reflecting consistent institutional distrust, but the primary data suggests this distrust is directed at the wrong level of analysis. The FDB cannot perform its development mandate because the budget architecture structurally prevents it from doing so.

Financing Node: Bank Exclusion and the Single-Investor Model

Bank exclusion from film production financing was confirmed as the foundational financing failure across eight of nine sources. P1, P2, P6, P7, and P8 all reported direct encounters with commercial banks that declined film-related lending. The consistent basis for rejection was the bank's inability to value film assets as collateral, a finding directly consistent with (World Intellectual Property Organization 2023), framework connecting IP enforcement failure to financing failure.

In the absence of formal financing instruments, individual investors command approximately 95% of revenue rights against the filmmaker's 5%, a terms structure that reflects not negotiation but the absence of alternatives (P8). This power ratio has direct consequences for production quality: when one investor controls 95% of creative and financial decisions, investment flows toward perceived de-risking mechanisms such as star talent (P1), rather than the actual value drivers of successful production.

As P8 framed it: *"It is not that there is no money. It is that the people with money have all the leverage because there is nowhere else to go."*

The data absence compounds the financing failure in a specific way not previously documented. The focus group (P3, P4, P5) confirmed that the FDB's cessation of box office data publication in July 2023 had begun eroding the informal capital sources the sector relied upon. P3 documented the retreat of Non-Resident Nepali investors who had been active between 2018 and 2021 and withdrew after invested productions failed to recoup and offered no meaningful legal recourse against revenue misreporting. 'They are gone because there is no system to protect them,' P3 stated. This finding establishes that data failure has direct financing consequences beyond investor confidence, it actively destroys existing informal capital networks.

A connected finding with NRI investment withdrawal and reliable box office data was articulated by P8, *"Our box office data is limited to domestic gross collection, they do not have the infrastructure to measure the collections from international releases and digital. This causes the sector to look less profitable than it is."* This finding is confirmed with secondary data, and it is a point to be noted that it actively increases investor distrust (FG).

Policy Node: Complicit Creativity Confirmed and Extended

Seven of nine sources confirmed (Fang 2024) "complicit creativity mechanism, making it the most densely coded node in the dataset, with two original extensions. Film has been historically documented as a mode of state control, (Maharjan 2012), The Motion Picture Act 1969, with four amendments including a substantial revision in 2017, retains its foundational architecture of state control. Rule 16 of the Motion Picture Rules embeds a representative of the Ministry of Home Affairs within the Film Review Committee, providing the legislative foundation for the political censorship pattern documented across the interview data. The new Film Bill, point 27, preserves this architecture and deepens the ambiguity further through the introduction of terms such as 'self-censorship', while also adding mandatory registration without clarifying censorship criteria.

P1 documented a case, independently corroborated by multiple participants, in which a filmmaker faced heavy censorship for using a political leader's name in audio, while the same leader appeared as an actor in another film during the same month. This is complicit creativity in operational form, the censor board communicates through unwritten motivations rather than articulated criteria, leaving filmmakers to reverse-engineer the political logic. P6 identified pre-review script modification as the earliest possible point of creative interference, with filmmakers modifying scripts at the development stage to pre-empt problems before the camera rolls. P8 extended the mechanism beyond the formal censor board to four parallel channels simultaneously: police, permits, social pressure, and political reaction.

P7 raised a contemporary digital-era argument that challenges the censorship regime's stated rationale: content freely available on social media platforms is simultaneously censored in theatrical film. This contradiction collapses the premise that censorship serves a protective function and signals institutional unreliability to international co-production partners who cannot forecast creative risk when the criteria are both opaque and incoherent.

The Moral Rights Inalienability Gap

The expert legal source (PB) identified a previously underexamined risk with significant implications for the sector. Nepal's Copyright Act 2059 B.S. (2002 A.D.) establishes moral rights as inalienable: they cannot be assigned, transferred, or waived. Yet PB documented that standard production contracts in Nepal routinely include clauses requiring filmmakers to waive their moral rights. These clauses are, by PB's legal analysis, unenforceable under Nepali law.

The implications are structural across three parties. Investors who believe they have acquired full rights over a production are operating under a legally false assumption. Filmmakers who have signed waiver clauses have in fact retained their moral rights under law but may be unaware of this protection. International co-production partners who conduct legal due diligence before committing to a collaboration will identify this as a fundamental contract enforceability problem, consti-

tuting a barrier to formalized co-production independent of the regulatory and fiscal barriers identified in Section 4.6.

This finding reframes the IP problem as a combined enforceability and information gap. Legal protections exist but have not been institutionally distributed to the practitioners who need them. P4 corroborated this from a co-production perspective, documenting that European co-production funds had claimed moral rights through deliberately vague contractual language: 'A European co-production fund states the European partner must own a minimum of 20% of the film. That 20% could be anything: financial rights, IP, copyright, moral rights; they deliberately keep those things vague.' Practitioners are surrendering protections they legally possess because no institutional mechanism exists to inform them otherwise.

Co-Production Node: The Asymmetry Trap

The co-production node produced the clearest evidence of the asymmetry trap identified by (Barker et al. 2021). P3 documented a decade-long effort to develop a South Asian co-production framework that has repeatedly stalled at the point of regulatory negotiation: *"International partners require certainty about censorship criteria before committing creative content to a collaboration, and Nepal cannot provide that certainty because the criteria do not exist in published form."*

The fiscal barrier is equally prohibitive. The focus group discussed 25% corporate tax on incoming co-production funding, which P4 described as structurally counterproductive: *"Any international partner who brings financing into Nepal for a co-production is taxed 25% on arrival, before a single frame is shot. While other nations' film policies focus on incentivizing production within their borders, Nepal's framework is extractive.* This makes Nepal's co-production cost structure uncompetitive against regional alternatives where no equivalent entry tax exists (Jin and Su 2019).

P8 confirmed zero market conversions from existing international film festivals in Nepal meaning, films screened at film festivals of Nepal are not finding distribution or supplementary partners/producers. The current film festivals are not a marketplace because the ecosystem does not exist, establishing that informal relationship-building is not converting into formal co-production arrangements.

P7 documented an additional regulatory barrier: foreign crew members cannot be issued work visas and must enter on tourist visas, rendering their equipment worth millions uninsurable. There is a legal pathway to acquiring work visa (non-tourist visa), confirmed through Department of Immigration (DoI), but the bureaucratic red-tape and the lack of one-door-policy creates a 6-to-8-month administrative delay for legitimate co-production projects, that filmmakers tend to not purse, especially since there are also no tax incentives in Nepal which they can find elsewhere, as P7 elaborated.

Table 1
Summary of Node-Level Institutional Failures

Node	Failure	Primary Evidence	Theoretical Link
Node 1: Financing	Bank exclusion universal, 95/5 investor-filmmaker power ratio	8/9 sources, P8 direct testimony	(Parc 2021) Stage 1, (World Intellectual Property Organization 2023)
Node 2: Policy	Complicit creativity, 1969 Act DNA persists in new Film Bill	7/9 sources, P1, P6, P8	(Fang 2024)
Node 3: IP	Moral rights inalienability gap, unenforced copyright	PB expert testimony, P4, P8	(World Intellectual Property Organization 2023), (Kouletakis 2023)
Node 4: Co-production	Zero treaties, 25% entry tax, asymmetry trap	P3, P5 focus group, P7	(Jin and Su 2019), (Barker et al. 2021)
Node 5: Data (master)	FDB data blackout July 2023 to April 2026, 25% discrepancy	7/9 sources, Film Development Board (2023), Film Development Board (2026)	(Parc 2021) economic legibility

Further Findings

The findings discussed above have a downstream effect applied to multiple aspects of film production. The following are three examples illustrating how structural dysfunction at the institutional level has been manifested in practice.

Story Bank design failure

A Stage 2 initiative, The Story Bank program by FDB which promised to provide developmental funds, disburses funds only after a censorship certificate is obtained, making it a post-production or distribution grant rather than a pre-production or development fund. Its 30-day completion requirement and opaque selection criteria render it unusable (P3, P6). This happens because of the budget architecture of FDB as discussed in 4.2 and FDB acting as a regulatory body while holding development mandate.

Development ceiling

The 90% failure rate theatrically in 2022 derived from (Shahi 2025) (Film Development Board 2023) but the box office data is limited to domestical theatrical release as noted by (P8), revenue from international release and digital are completely absent. This is created by the data vacuum. Primary data also establishes the failure rate is not a talent problem but a financing architecture problem. Absent development support, productions enter principal photography with insufficient script preparation, producing technically complete but narratively weak films regardless of the creative ability of those involved (P7). So again, lack of reliable data reinforcing a developmental ceiling.

Trend-following because of data absence

Without reliable and accurate audience or box office data, filmmakers cannot make evidence-based production decisions and default to replicating whichever film was most recently successful (P2, FG, P6). Adversely effected by the 95%-5% power ratio between financer and filmmakers, this loop reinforces itself and trend following continues. This produces franchise dependence that concentrates revenue in a small number of titles while crowding out original work (P4).

Conclusion

Summary of Contributions

This research contributions to the literature on film industry development in small markets. First, it introduces and empirically grounds the legitimacy trap as a named diagnostic concept describing the self-perpetuating cycle of five interlocking institutional node failures in Nepal's film sector. Second, it identifies the FDB budget impossibility as an original mechanism-level finding: reinvestment failure in Nepal is a structural design feature, not institutional incompetence, because the institutional plumbing does not connect the sector's tax revenue to the institution mandated to develop the sector. As a result, taxation without reinvestment becomes a disadvantage to creative fields (Messerlin and Parc 2017). Third, it identifies the moral rights inalienability gap as a previously undocumented systemic risk, reframing Nepal's IP problem as a combined enforceability and information deficit addressable at low institutional cost.

Theoretically, the research extends (Parc 2021) institutional sequencing model to a small, low-income, institutionally nascent context and contributes an empirically derived Phase 0, covering governance preconditions, that the original three-stage model of Parc did not anticipate.

Policy Recommendations

The sequenced reform roadmap is organized into four phases. Phase 0 addresses institutional hurdles that must be overcome before phased reforms can be delivered. Phases 1 through 3 follow (Parc 2021) staging logic. These recommendations are grounded in interview data and comparative institutional evidence. They are presented as evidence-informed proposals, with the strength of each recommendation reflecting the depth of saturation in the interview dataset.

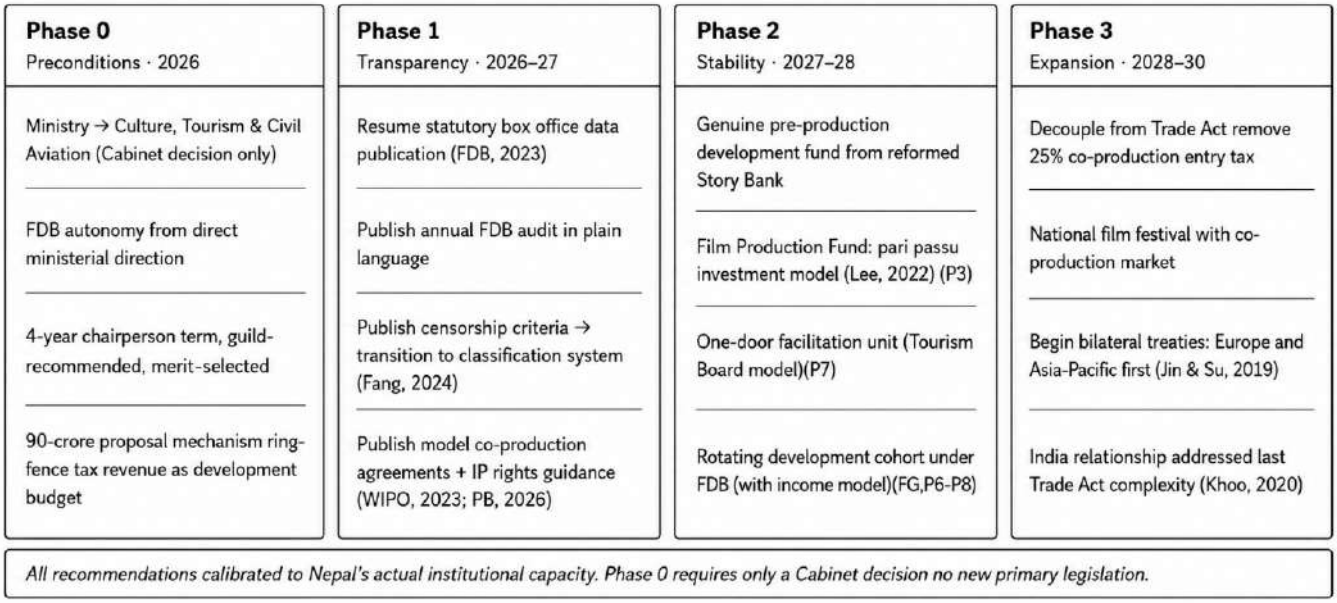

Figure 2: Phased Reform Roadmap

Phase 0 (2026): Institutional Preconditions

Derived entirely from primary interview data, Phase 0 includes transfer of the film sector's regulatory home from the Ministry of Communication and Information Technology to the Ministry of Culture, Tourism and Civil Aviation, a position supported by five independent sources (P1, P3, P6, P7, P8). Because the film sector falls under a regulatory ministry, the FDB defaults to control rather than facilitation regardless of legislative language. Additional preconditions include granting the FDB institutional autonomy from direct ministerial political direction (five independent sources across both tracks). Extending the FDB chairperson term from two to four years with merit-based, guild-recommended selection. Introducing a proposal mechanism requiring each chairperson candidate to submit a four-year development plan against a budget of approximately NPR 90 crore, the sector's own annual tax revenue focusing on creation of a polycentric architecture of the following bodies, (a) Rating and Classification Office (b) Data and Research Office, (c) One Door Facilitation Unit, (d) Film Development Fund, (e) Development Support Unit, (f) National Film Festival and Co-production Market. Finally, encouraging formal representatives from independent filmmakers in policy consultation processes as any legitimate Stage 1 and 2 initiatives currently undertaken has been on their recommendation (P6, FG).

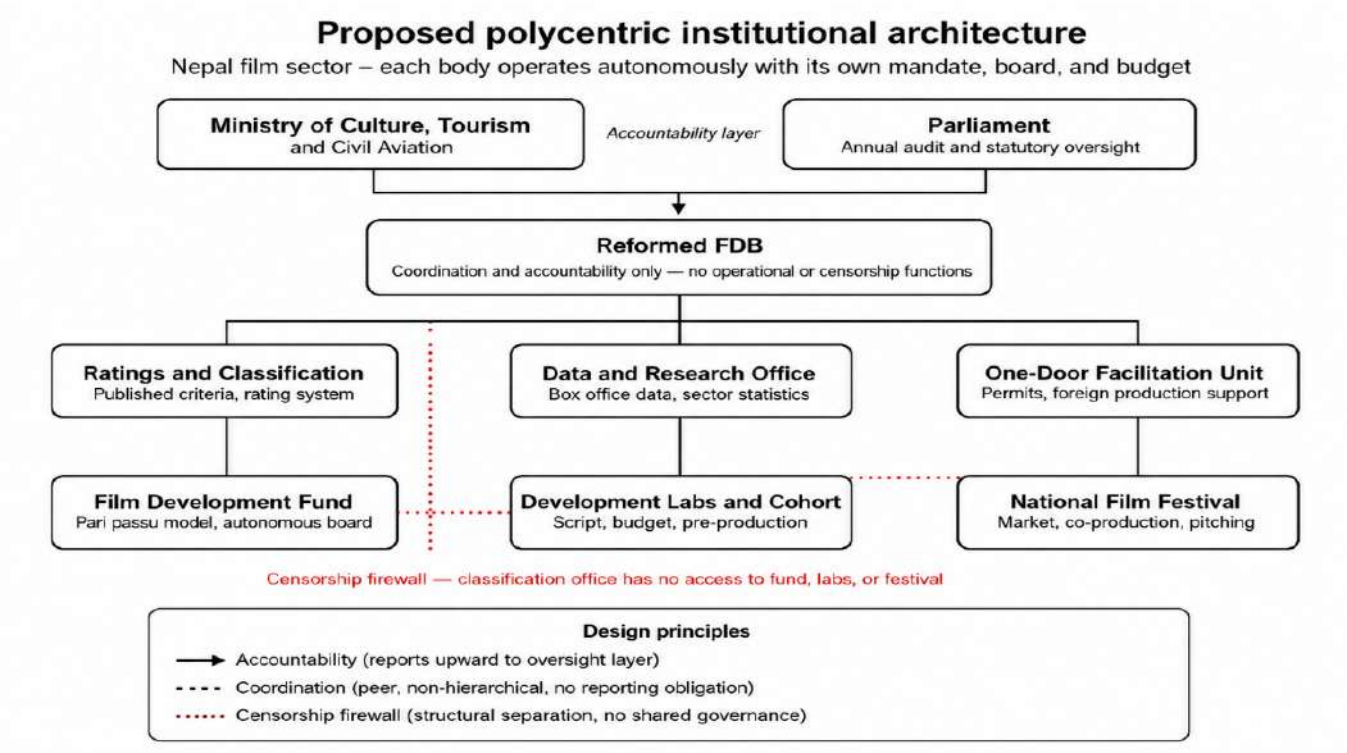

Figure 3: Reformed Polycentric Institutional Architecture

Phase 1 (2026 to 2027): Transparency

The FDB must resume box office data publication with a statutory monthly obligation. Remove the discretion that enabled the 2023 to 2026 data blackout (P1, P6, P7, FG). In plain language an annual FDB audit must be published against the chairperson's four-year plan. The Film Review Committee must publish censorship classification criteria and transition toward a rating system replacing the binary of permission or prohibition. Create and publish models for co-production agreements and IP rights guidance explicitly mentioning the inalienability of moral rights under Nepal's Copyright Act 2002.

Phase 2 (2027 to 2028): Stability

The FDB's Story Bank, a nominal pre-production fund that in practice disburses only after a censorship certificate is obtained, must be restructured into a genuine pre-production or development fund disbursed at script and budget stage. A separate Film Development Fund operating on the pari passu investment model (P3), modelled after the "Red Sea Fund, in which the fund co-invests without requiring repayment until profit is generated, should be established. Independent convergence on the pari passu model by P3, P7 and P8 without prior contact is notable: it suggests the sector already understands the instrument it needs. A one-door facilitation unit consolidating permit coordination, foreign investment processing, and co-production support should be established within the FDB's existing administrative powers.

Phase 3 (2028 to 2030): International Integration

The legal entanglement of film co-production within Nepal's general Trade Act must be resolved through targeted regulatory amendment, and the 25% tax on incoming co-production funding replaced with a zero-tax or tax-deferral mechanism for funds entering under a formal co-production agreement. A national film festival with an embedded co-production market, independently endorsed by P1, P6, P7 and all focus group participants, should be established with a provincial rollout model beginning in cities with existing cinema infrastructure. Bilateral treaty negotiations should prioritize European and Asia-Pacific partners where informal co-production relationships already exist, before the more structurally complex India trade relationship is addressed.

Limitations and Future Research

The primary methodological limitation is the data void imposed by the FDB's cessation of box office reporting, which means some empirical claims about Nepals film sector rely on a single peer-reviewed source (Shahi 2025) and official FDB documents (Film Development Board 2023) (Film Development Board 2026) including ILO reports. As established in Section 3, this is not a design limitation, but the study's central finding expressed methodologically. The qualitative interpretivist design does not permit statistical generalization, findings are context-specific to Nepal's film sector in 2026.

Future research priorities include: a longitudinal quantitative study of box office performance once transparent data is restored. Nation-wide viewer preference research covering all demographics. A qualitative investigation of the documentary and short film sectors, which face distinct institu-

tional challenges outside the scope of this study. Analysis of the distribution chain as an underexamined bottleneck and assessment of whether OTT platforms can provide a bypass for domestic institutional failures, incorporating the cautionary dimensions of platform dependency identified by (Lobato 2019).

Disclosure Statement

The author has no financial or non-financial disclosures to share for this article.

Use of Artificial Intelligence (AI) Tools: The research has followed the ICJMR policy on AI use. No generative AI has been used in preparation of this report.

Data Availability Statement: Interview transcripts are not publicly available to protect participant anonymity in accordance with the consent agreements. FDB institutional data is available at fdb.gov.np and policy frameworks in film.gov.np > media > filmgov including Motion Picture Act and the New Film Bill na.parliament.gov.np as of this research under review.

Ethical Approval: This research was part of the authors masters thesis and has been reviewed and approved by the dissertation supervisor and dissertation committee at Islington College, Kathmandu, in accordance with London Metropolitan University ethical guidelines. All participants provided written informed consent prior to interview.

Acknowledgements

The author acknowledges the guidance of the dissertation supervisors, Manoj Jaishi and Aarati Shilpakar at Islington College, Kathmandu, and the generous participation of all interview participants who gave their time, experience, and unfiltered honesty to this research.

References

Barker, Thomas et al. (2021). *Censorship and Its Impact on the Screen Industries in Malaysia*. Kuala Lumpur: Freedom Film Network.

Dhakal, R. (June 2024). *Nepali Filmdom Still Mired in Lack of Accountability, Transparency*. p. 6.

Fang, Jun (2024). "The Culture of Censorship: State Intervention and Complicit Creativity in Global Film Production." In: *American Sociological Review* 89.3, pp. 488–517. 10.1177/00031224241236750.

Film Development Board (2023). *National Movie Collection Report*. Tech. rep. Kathmandu: Film Development Board, Government of Nepal.

– (2026). *National Movie Collection Report*. Tech. rep. Kathmandu: Film Development Board, Government of Nepal.

International Labour Organization (2021). *Analytical Brief on the Film Industry in Nepal*. Tech. rep. Kathmandu: International Labour Organization.

International Labour Organization and National Planning Commission, Nepal (2022). *Untapping the Potential of the Nepali Film Industry to Generate Decent Employment for Women*. Tech. rep. Kathmandu: International Labour Organization and National Planning Commission, Nepal.

Jedlowski, Alessandro (2021). *African Screen Media: Nollywood and the Genealogy of a Revolution*. New York: Columbia University Press.

Jin, Dal Yong and Wendy Su (2019). *Asia-Pacific Film Co-Productions: Transnational Collaborations in the Era of New Media*. New York: Routledge.

Khoo, Olivia (2020). "Pan-Asian Filmmaking and Co-Productions with China: Horizontal Collaborations and Vertical Aspirations." In: *The Chinese Cinema Book*. Ed. by Chris Berry. 2nd. London: British Film Institute / Palgrave Macmillan, pp. 197–205.

Kouletakis, Jade (2023). "Weak Copyright Enforcement Produces Revenue Loss and Deters Foreign Investment: Evidence from Nigeria and South Africa." In: *Journal of Media Law* 15.2, pp. 44–68.

Lee, Hye-Kyung (2022). "Supporting the Cultural Industries Using Venture Capital: A Policy Experiment from South Korea." In: *Cultural Trends* 31.1, pp. 82–96.

Lobato, Ramon (2019). *Netflix Nations: The Geography of Digital Distribution*. New York: NYU Press.

Maharjan, Harsha Man (2012). "Machinery of State Control: History of Cinema Censor Board in Nepal." In: *Bodhi: An Interdisciplinary Journal* 4.1, pp. 45–67.

Messerlin, Patrick A. and Jimmyn Parc (2017). "The Real Impact of Subsidies on the Film Industry (1970s-Present): Lessons from France and Korea." In: *Pacific Affairs* 90.1, pp. 51–75. 10.5509/201790151.

Morawetz, Norbert, Jane Hardy, Colin Haslam and Keith Randle (2007). "Finance, Policy and Industrial Dynamics: The Rise of Co-Productions in the Film Industry." In: *Industry and Innovation* 14.4, pp. 421–443. 10.1080/13662710701524072.

Parc, Jimmyn (2020). "Understanding Film Co-Production in the Era of Globalization: A Value Chain Approach." In: *Global Policy* 11.4, pp. 458–465.

– (2021). "Business Integration and Its Impact on Film Industry: The Case of Korean Film Policies from the 1960s Until the Present." In: *Business History* 63.5, pp. 850–867. 10.1080/00076791.2019.1676234.

Ramachandran, Naman (June 2024). *As Nepali Cinema Soars Globally, Taxes Cripple Local Film Industry*. Interview.

Shahi, Suman Bikram (2025). "Production Cost of Feature Film in Nepal." In: *The Journal of Academic Development* 9.1, pp. 89–116.

World Intellectual Property Organization (2023). *IP Assets and Film Finance*. Tech. rep. Geneva: World Intellectual Property Organization.