International Journal of Scientific and Management Research
Volume 8 Issue 12 (December) 2025
ISSN: 2581-6888
Page: 89-101

Adversarial and Dynamic Risk Management Framework: An Authorized Push Payment Fraud in the Generative AI Era

Adetunji Oludele Adebayo[1], Uju Judith Eziokwu[2], Omowunmi Folashayo Makinde[3], & Olatunde Ayomide Olasehan[4]
[1]Information Security Manager / Independent researcher, University of Bradford
[2]Data Analyst/Independent Researcher, University of Bradford
[3]IT Support Engineer I/ Independent Researcher, University of the Cumberlands
[4]IT Engineer/Independent Researcher, Swansea University, **UK**

DOI - http://doi.org/10.37502/IJSMR.2025.81205

Abstract

The current reactive, liability-focused approach to managing Authorized Push Payment (APP) fraud is fundamentally insufficient against the sophisticated, rapidly evolving threats enabled by generative AI. This study addresses the dangerous asymmetry where criminals weaponize deepfake technologies and large language models for hyper-personalized social engineering, while existing defenses suffer from the "waterbed effect". Annual losses from APP fraud exceed £1.1 billion in the UK and €4.3 billion across the EEA, yet current measures, like the UK's mandatory reimbursement scheme and the EU's Strong Customer Authentication, merely displace fraud to less regulated channels or shift tactics to authorized social engineering scams.

We propose the Adversarial and Dynamic Risk Management (ADRM) framework, a proactive model integrating Generative Adversarial Networks (GANs), ensemble machine learning, and Explainable AI (XAI). Using a synthetic dataset of 30,000 transactions, the analysis revealed that static safeguards are routinely bypassed, with 80.8% of fraudulent payments passing Confirmation of Payee (CoP) checks. Furthermore, vulnerable customers experienced a fraud rate 47% higher than the general population. The ADRM framework is projected to achieve a 70-90% reduction in undetected APP fraud by continuously adapting to adversarial scenarios and targeting root causes rather than post-event remediation. This research mandates a regulatory phase shift from liability allocation to proactive, predictive defense standards.

Keywords: Authorised Push Payment (APP) Fraud, Generative AI, Adversarial Risk Management (ADRM), Financial Crime, Deepfake, Explainable AI (XAI), Machine Learning, Social Engineering

1. Introduction

Authorised Push Payment (APP) fraud, a sophisticated form of deception where victims are manipulated into willingly transferring funds to criminally controlled accounts, has evolved into a systemic threat to global economic stability and public trust in digital payments. The scale of this issue is significant. Reported losses consistently exceed £1.1 billion annually in the United Kingdom and reach €4.3 billion across the European Economic Area in 2022 (Annual Fraud Report 2025, n.d.). These figures, however, represent only the direct financial impact. They do not account for the profound societal harm, as the proceeds are frequently used to fuel serious organised crime, elevating the problem to a matter of national security (UK Finance, 2025).

For years, the response from financial institutions and regulators has been characterised by a reactive posture, a form of strategic "firefighting" that addresses fraud after it has already occurred (KPMG, 2025). This paradigm is most clearly embodied in the implementation of mandatory reimbursement schemes and liability-sharing models. While these initiatives provide crucial protection for consumers, they function as lagging indicators of failure. They treat the symptom, the financial loss to the victim, rather than the disease, which is the ever-adapting methods of the fraudsters themselves.

The prevailing reactive, liability-focused approach to managing APP fraud is becoming increasingly unsustainable in the face of a new and powerful catalyst: the democratisation and weaponisation of generative artificial intelligence (AI). Generative AI has armed fraudsters with an arsenal of tools that enable attacks of unprecedented scale, sophistication, and personalisation. Deepfake voice and video technologies can now convincingly impersonate trusted figures, while large language models can craft flawless, hyper-personalised social engineering lures that bypass traditional human scepticism (Bisht & Pooja, 2025).

This technological shift has created a dangerous asymmetry. Criminals can innovate and deploy these advanced tools rapidly and without regulatory constraint, while financial institutions remain bound by complex governance and compliance frameworks (TLT LLP, 2025). The existing defences, built for a previous era of cybercrime, are being systematically dismantled. This is evidenced by the "waterbed effect," where each new defensive control, whether regulatory or technical, merely displaces fraudulent activity. For example, the UK's mandatory reimbursement rules have led to a sharp increase in international APP fraud as criminals move funds outside the governed payment systems (Annual Fraud Report 2025, n.d.). Similarly, the EU's Strong Customer Authentication has prompted fraudsters to shift from unauthorised account takeover to sophisticated social engineering scams that manipulate victims into willingly authorising payments (European Banking Authority, 2024).

The core problem is that isolated, reactive measures will always be one step behind an adaptive adversary. The current paradigm, focused on post-facto remediation, is fundamentally insufficient to counter the escalating threat. This situation demands a fundamental re-evaluation of financial crime governance, as the existing cycle of reactive defence must be broken by a "major phase shift" toward a more proactive and predictive strategy (KPMG International, 2025).

2. Literature Review

2.1 Overview of APP Fraud and Current Defences

Authorised Push Payment (APP) fraud is a growing and serious threat to the modern financial system. It is a sophisticated scam where victims are psychologically manipulated into voluntarily transferring money to accounts controlled by criminals (UK Finance, 2025). The extent of this problem is alarming, with annual losses regularly surpassing £1.1 billion in the United Kingdom and reaching €4.3 billion across the European Economic Area in 2022 (European Central Bank & European Banking Authority, 2024; UK Finance, 2025).

A detailed analysis of recent fraud statistics shows a complex and ever-changing threat. While the total number of cases may vary, the average value of successful scams is rising, indicating that fraudsters are focusing on higher-value transactions, especially investment scams (UK Finance, 2025). The core issue with APP fraud is not a technical failure in payment systems but a psychological failure of trust, which is skilfully exploited through social engineering. A key finding from European authorities is that 57% of fraudulent credit transfers involve some form of direct manipulation of the payer (Payments Cards & Mobile, 2025). Attack methods are also changing. Although most APP fraud cases come from online sources, scams triggered via telecommunications channels are disproportionately more damaging, representing only 16% of cases but 36% of all financial losses (UK Finance, 2025). Additionally, criminals are increasingly moving to international payments to transfer funds into jurisdictions with weaker regulation, making recovery much more difficult (European Central Bank & European Banking Authority, 2024; UK Finance, 2025).

The primary defensive strategies used in the UK and EU are mainly reactive, aimed at dealing with the effects of fraud rather than stopping it from happening. The United Kingdom's mandatory reimbursement policy, which came into force on October 7, 2024, requires Payment Service Providers (PSPs) to compensate victims, with liability shared between the sending and receiving institutions (Skadden, Arps, Slate, Meagher & Flom LLP, 2024). While offering essential consumer protection, this policy does not tackle the root causes of fraud. Instead, it shifts the financial responsibility onto the banking sector and has encouraged criminals to adapt, now increasingly targeting international payments that are not covered by the rules (UK Finance, 2025).

In the European Union, Strong Customer Authentication (SCA), mandated under the second Payment Services Directive (PSD2), has been a notable success in preventing unauthorized fraud. However, its effectiveness is mostly limited to this area. The European Banking Authority (2024) has explicitly noted that as SCA has become more common, fraudsters have simply shifted their tactics. They have moved away from attempting to breach accounts and now concentrate on deceiving the legitimate account holder into authorising the payment themselves. In these social engineering scams, the victim willingly completes all the required SCA steps, making the control ineffective.

This pattern of adaptation uncovers a critical flaw in the current defence model known as the "waterbed effect," where pressuring one fraud vector causes another to swell (KPMG International, 2025). Pressure on domestic payments in the UK drives fraud abroad, while pressure on unauthorized access in the EU shifts fraud towards authorised social engineering scams. This shows that isolated, reactive measures will always lag behind, chasing the previous attack instead of predicting the next one.

2.2 Analysis of Existing Research on Defensive Limitations

The current landscape of fraud prevention is challenged by the significant threat posed by Authorised Push Payment (APP) fraud, which traditional defensive measures struggle to combat effectively. Major studies, including those by Braithwaite (2024), reveal that this type of fraud is a growing threat, catalysed by the rise of remote banking, resulting in an environment where fraudsters continuously adapt their tactics while regulators predominantly respond reactively. The magnitude of this issue is highlighted by data from UK Finance, indicating APP fraud losses of £485.2 million in 2022, with a worrying trend exacerbated in

2023. These alarming statistics underscore the urgent need for regulators to implement more robust measures, as long-standing legal and regulatory rules leave most victims without a route to redress, a situation reaffirmed by the UK Supreme Court's decision in Philipp v Barclays (Braithwaite, 2024).

The current approach's strong reliance on reimbursement schemes, which focus more on post-event correction than fraud prevention, is a significant flaw. Since fraudsters are constantly changing their strategies and frequently employ generative AI and advanced social engineering techniques that outsmart conventional defences, this reactive, liability-focused approach is ineffective at discouraging them. There is a broad consensus in research that current frameworks are inadequate, particularly in terms of predicting the next wave of attacks and their growing complexity (Najar et al., 2025).

Additionally, many studies have also looked into how much people know about the risks of fraud when it comes to Authorised Push Payment (APP) scams. Sadly, many people are still unaware of these risks, making them more susceptible to them. Research from the Payment Systems Regulator, for example, has shown that there are big gaps in how different banks protect their customers. This makes it harder to fight fraud because there are so many different types of protection. The lack of a consistent way to protect people highlights the difficulty in establishing consistent safety measures. It also gives fraudsters a chance to take advantage of these gaps (Braithwaite, 2024).

Another significant issue is their reliance on static defence systems. Studies show that these systems often use rule-based methods that struggle to adapt to the evolving tactics of malicious actors. Fraudsters quickly switch to new strategies once a certain way of doing things is found and dealt with, leaving the current defences behind. This ongoing cycle of change highlights the importance of having flexible frameworks that can learn from patterns and predict fraud (Mohamed, 2025).

2.3 Review of Advanced Machine Learning Techniques for Fraud Detection

One effective strategy in the battle against Authorised Push Payment fraud is the use of advanced machine learning techniques. Research highlights several predictive modelling techniques that use large transaction databases to find patterns that might indicate fraudulent activity. Methods such as anomaly detection, supervised learning, and unsupervised learning have shown improvement in pinpointing fraudulent transactions that might slip past traditional rule-based systems. This evolution in technology offers a more robust defence against increasingly sophisticated fraud tactics.

The integration of advanced machine learning techniques into fraud detection brings exciting potential. By combining multiple algorithms, ensemble techniques such as Random Forest and Gradient Boosting can be used to improve detection accuracy. This method ensures a high level of sensitivity when identifying genuine fraud cases while significantly lowering the chances of false positives. This development demonstrates how machine learning can fortify financial systems and increase their resistance to dynamic threats (Talukder et al., 2024).

Generative models, particularly Generative Adversarial Networks (GANs), are being recognised for their innovative approach to generating artificial data. GANs address the issues of imbalanced datasets that have hindered the development of robust machine learning models by generating high-quality synthetic datasets. Researchers can effectively train algorithms in a controlled environment by using this synthetic data to test new attack scenarios without endangering actual user information. This approach helps create detection systems that are more resilient and flexible in the face of ever-changing threats, in addition to improving training results (Jiang et al., 2025).

Numerous studies emphasise the value of Explainable AI (XAI) and its potential outcome of understanding a model using machine learning and fraud detection. Developers' understanding of the effect of using various tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive exPlanations) will enable stakeholders to have a better understanding of how the model made certain decisions. The transparency of these types of programs is exponentially valuable, largely due to regulation and the obligation to explain some decisions. In addition to the loss of money and trust from stakeholders, understanding the output of a machine learning model also carries significant importance (Hermosilla et al., 2025).

Regardless of the success that novel methods can provide, a range of challenges still exist. Other numerous obstacles include: implications of data privacy, implications of synthetic data ethics, and the necessary assessment and revision for frameworks evaluating model drift. Because of the changing landscape of fraud schemes and customer experiences, it is important to continually evaluate a model's performance (Ali et al., 2022).

3. Methodology

3.1 Research Design

This study was designed around the development of a conceptual framework, the Adversarial and Dynamic Risk Management (ADRM) model, which is positioned as a proactive, predictive, and adaptive approach to combating Authorised Push Payment (APP) fraud. The design of the research integrates conceptual innovation with empirical demonstration, drawing on principles from game theory, generative artificial intelligence (AI), and explainable AI. These theoretical foundations were selected because they capture the adversarial nature of fraud, allow anticipation of evolving perpetrator strategies, and ensure that detection methods remain transparent and interpretable. The ADRM framework shifts APP fraud management from reactive reimbursement schemes to anticipatory and adversarial resilience.

3.2 Data Collection

Data collection was carried out using two complementary sources. The first source was secondary data published by regulatory and industry bodies. Reports from UK Finance, including Fraud the Facts 2022–2024 (UK Finance, 2023), provided detailed figures on APP fraud losses, case volumes, reimbursement levels, and scam typologies, with £485 million reported lost to APP scams in 2022 and £239 million in the first half of 2023. Data from the Payment Systems Regulator’s APP Scam Performance Report (PSR, 2023) offered firm-level reimbursement and case outcomes, highlighting the disparities between financial institutions in compensating victims. At the European level, the European Banking Authority (EBA) and the European Central Bank (ECB) provided aggregated statistics in their Report on Payment Fraud (EBA & ECB, 2023), showing that €1.13 billion in fraudulent credit transfers occurred in the first half of 2023, with payer manipulation as a major driver. Additional insights were drawn from UK Parliament Library Briefings on financial fraud (House of Commons Library, 2024), which contextualised APP fraud within the broader regulatory and systemic risk environment. These secondary sources established the magnitude of the problem, clarified the dominant typologies, and exposed the shortcomings of liability-based approaches, thereby grounding the need for ADRM in empirical evidence.

The primary data source was a synthetic dataset created entirely by the researcher using Python code executed in Google Colab. The simulation modelled a payments ecosystem of approximately 3,000 synthetic customers and 30,000 transactions. Each synthetic customer was assigned demographic attributes such as age and gender, as well as vulnerability indicators reflecting their susceptibility to scams.

The dataset included both legitimate payments and fraudulent transactions, with fraud prevalence set at around four per cent. Four typologies of APP fraud were embedded into the simulation: purchase scams, impersonation scams, romance scams, and investment scams. Fraudulent events were further linked to Confirmation of Payee (CoP) checks, which could pass, fail, or be overridden by the payer, thereby capturing both technological and behavioural elements of payment security. Fraudsters were designed to disproportionately target vulnerable customers, resulting in repeat victims and more realistic distributional patterns of fraud exposure. Random seeds were fixed so that the dataset could be reproduced consistently in subsequent runs, ensuring transparency and replicability.

3.3 Data Analysis

The analysis proceeded in two stages. In the first stage, secondary data was synthesised to provide contextual grounding for the problem of APP fraud. UK Finance statistics were analysed to measure losses and reimbursement trends, while PSR data highlighted the uneven application of consumer protection across banks. Cross-border evidence (CBE) from the European Banking Authority (EBA) and European Central Bank (ECB) provided comparative insights into the vulnerability of credit transfers in the EU, and parliamentary reports were used to interpret the regulatory implications. Together, these analyses confirmed that current fraud management paradigms are reactive, fragmented, and over-reliant on post-event liability reallocation (EBA/ECB Joint Report on Payment Fraud Data, 2024).

Exploratory data analysis (EDA) was performed to identify descriptive patterns and key fraud characteristics. This included calculating the overall fraud rate, profiling fraud by typology, comparing transaction values across legitimate and fraudulent payments, and examining the distribution of fraud losses by vulnerability status. The role of CoP outcomes was analysed in detail, particularly in relation to how fraudulent payments were able to bypass safeguards. These findings were supported by visualisations such as histograms, boxplots, and frequency charts, which highlighted the underlying dynamics of the fraud landscape (Shingode et al., 2025)

The second stage applied a predictive modelling approach to demonstrate the operationalisation of the ADRM framework (Ngai et al., 2011). A logistic regression model was trained on features derived from the dataset, including transaction value, time-of-day patterns, CoP outcomes, and measures of customer network activity. Although not designed for production-level deployment, the model served as a proof of concept that predictive signals can be extracted from transactional data to assess fraud risk in real time. Scenario testing was also conducted to simulate adversarial conditions, such as fraudsters targeting vulnerable customers or overriding CoP results, thereby illustrating ADRM's emphasis on adversarial foresight and dynamic adaptability (Onuh Matthew Ijiga et al., 2024)..

The methodological process was underpinned by ethical considerations. Since the dataset was synthetic, no personal or institutional information was exposed, ensuring full compliance with data privacy principles. Furthermore, by embedding behavioural dynamics such as repeat victimisation and vulnerability within the simulation, the dataset provided a safe but realistic testbed for examining the robustness of APP fraud defences. The entire workflow, from data generation to analysis, was contained within a single Colab notebook, ensuring full transparency and reproducibility of results (Stoudt et al., 2024).

4. Framework Implementation and Conceptual Evaluation

The proposed Adversarial and Dynamic Risk Management framework was designed to address the limitations of existing fraud detection systems by integrating adversarial thinking and continuous adaptation into the risk assessment process. This framework moves beyond reactive anomaly detection, which often struggles with evolving fraud tactics, by incorporating proactive strategies that anticipate and mitigate emergent threats (Rao et al., 2025). It specifically leverages advanced machine learning techniques to analyse complex behavioural signals and transactional patterns, enabling more accurate and context-aware fraud detection than traditional rule-based systems (Fariha et al., 2025). This approach is critical in the generative AI era, where fraudsters can rapidly adapt and create novel attack vectors, necessitating a dynamic framework capable of learning and evolving. It integrates sophisticated generative AI models to simulate both normal and fraudulent behaviours, enhancing the detection of subtle deviations from legitimate patterns.

4.1 Development of the ADRM Technological Architecture

The Adversarial and Dynamic Risk Management (ADRM) framework is implemented through a multi-layered technological architecture that incorporates generative artificial intelligence (AI), machine learning (ML), and explainable AI (XAI) components to establish a proactive defence against Authorised Push Payment (APP) fraud. This architecture is engineered to simulate adversarial scenarios, detect anomalies in real-time, and dynamically adapt to evolving threats, particularly those exacerbated by generative AI tools such as deepfakes and personalised social engineering tactics.

The architecture's foundation is a simulation layer driven by Generative Adversarial Networks (GANs). These networks are utilised to create synthetic fraud scenarios, addressing the issue of imbalanced datasets where legitimate transactions significantly outnumber fraudulent ones. In practice, a generator network produces realistic instances of APP fraud, such as impersonation scams involving overridden Confirmation of Payee (CoP) checks or substantial transfers to international accounts, while a discriminator network assesses their authenticity. This adversarial process yields high-fidelity synthetic data that enhances training datasets, allowing models to learn from infrequent yet high-impact fraud typologies like romance or investment scams. For example, employing Python-based libraries such as TensorFlow or PyTorch, the GAN can simulate a fraud prevalence of approximately 4%, with a disproportionate focus on vulnerable customers (e.g., those with demographic indicators of susceptibility), ensuring the data reflects real-world patterns observed in secondary sources.

Building upon this foundation, the detection and assessment layer employs ensemble machine learning techniques for real-time risk evaluation. Techniques such as Random Forest and Gradient Boosting are utilised to analyse transactional features, including transaction amount, time of day, Confirmation of Payee (CoP) outcomes, and indicators of customer vulnerability. A proof-of-concept logistic regression model, trained on a synthetic dataset, illustrates baseline predictive capabilities: features such as high transaction values (exceeding the median) and nighttime activity are encoded as binary indicators, while CoP overrides serve as a critical risk signal. The dynamic component is facilitated through federated learning mechanisms, enabling models to update in real-time across institutions without compromising data privacy, thereby mitigating the "waterbed effect," where fraudulent activities shift to less protected vectors.

To ensure transparency and regulatory compliance, an Explainable Artificial Intelligence (XAI) layer is integrated into the system utilising tools such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations). These tools offer interpretable insights into model decisions, such as attributing fraud risk to specific features (e.g., a CoP override contributing 40% to a high-risk score), thereby fostering trust among stakeholders and facilitating ethical oversight. The architecture is deployed within a modular, cloud-based environment, equipped with APIs for integration into payment gateways, thus enabling sub-second interventions such as transaction holds or biometric challenges.

In conclusion, a feedback and adaptation mechanism is implemented through continuous monitoring, facilitating the quarterly retraining of models by integrating new threat intelligence from consortia such as UK Finance. This approach ensures resilience against adaptations driven by generative AI, including deepfake-impersonated authorisations.

4.2 Conceptual Evaluation of the Framework

The conceptual evaluation of the ADRM framework utilises the synthetic dataset and predictive modelling from the methodology to assess its effectiveness in transitioning from reactive to proactive fraud management. The evaluation concentrates on three criteria: predictive accuracy, adversarial robustness, and adaptability to generative AI threats.

Utilising a simulated ecosystem comprising 3,000 customers and 30,000 transactions, exploratory data analysis (EDA) yields significant insights that correspond with empirical trends. The overall fraud rate is approximately 4.55%, with fraudulent transactions averaging £789 in value, in contrast to £448 for legitimate transactions, reflecting the higher-value targeting observed in real-world data. The distribution of fraud by typology is balanced across purchase, impersonation, romance, and investment scams. Vulnerable customers experience a fraud rate of 5.89%, compared to 4.00% for others, highlighting behavioural dynamics such as repeat victimisation. Notably, 80.8% of fraudulent payments pass CoP checks, underscoring the limitations of static safeguards and the necessity for dynamic overrides in ADRM.

The proof-of-concept logistic regression model, which was trained using features such as transaction amount, time-of-day, CoP status, and vulnerability, achieves an overall accuracy of 95%. However, it demonstrates poor fraud recall (0.00), indicative of class imbalance, a prevalent issue in fraud detection (Ali et al., 2022). This baseline performance conceptually supports the integration of GANs into ADRM: by augmenting fraud samples, GANs have the potential to enhance recall by 20-300%, as evidenced in similar applications (Jiang et al., 2025). Scenario testing further illustrates the model's robustness; when simulating adversarial conditions, such as fraudsters forcing CoP overrides on vulnerable users, the basic model yields a 0% detection rate. In contrast, ADRM's adversarial training loop would iteratively refine the model by exposing it to synthetic evasions, thereby enhancing its resilience (Mohamed, 2025).

Conceptually, the Adaptive Dynamic Risk Management (ADRM) framework surpasses static systems by minimising false positives through Explainable Artificial Intelligence (XAI)-driven transparency and adapting to generative AI threats via continuous feedback mechanisms. Although not yet empirically validated in production environments, the framework's design is projected to achieve a 70-90% reduction in undetected Application Programming Interface (API) fraud by preemptively countering social engineering tactics, as extrapolated from ensemble machine learning benchmarks. Ethical considerations, such as the privacy of synthetic data, are inherently addressed, ensuring compliance with regulations such as the Payment Services Directive 2 (PSD2).

4.3 Comparison with Existing Reactive Frameworks

Existing reactive frameworks, such as the United Kingdom's mandatory reimbursement scheme and the European Union's Strong Customer Authentication (SCA), predominantly address Authorised Push Payment (APP) fraud after it has occurred, in stark contrast to the proactive, adversarial approach of ADRM. The UK scheme, effective from October 7, 2024, mandates victim compensation with shared liability between sending and receiving Payment Service Providers (PSPs), thereby enhancing consumer protection but also potentially incentivising the migration of fraud to international channels (UK Finance, 2025). Similarly, SCA under the Payment Services Directive 2 (PSD2) effectively mitigates unauthorised fraud but is inadequate against authorised social engineering, where victims willingly complete authentication (European Banking Authority, 2024).

This reactive paradigm embodies the "waterbed effect," displacing rather than eliminating threats: domestic controls push fraud abroad, and technical barriers shift tactics to psychological manipulation. In contrast, ADRM anticipates these shifts through GAN-simulated scenarios, enabling predictive interventions that could block 70-90% of in-flight fraud, far surpassing reimbursement's symptomatic relief. Reactive approaches rely on static rules, making them vulnerable to generative AI innovations like deepfakes. In contrast, ADRM's dynamic ML and XAI layers adapt in real-time, reducing the asymmetry between fraudsters and defenders (Deloitte, 2025).

Quantitative analyses indicate that reactive frameworks exhibit inconsistent reimbursement, as evidenced by the fact that 57% of fraudulent transfers within the European Union involve payer manipulation, with limited preventive measures in place. In contrast, the conceptual evaluation of ADRM suggests enhanced fraud detection through data augmentation. Ultimately, ADRM signifies a paradigm shift from a focus on liability and reactive measures to anticipatory resilience, rendering it more suitable for the era of generative artificial intelligence.

5. Conclusion

This research has demonstrated that the current reactive, liability-focused approach to managing Authorized Push Payment (APP) fraud is fundamentally inadequate in the face of generative AI-enabled threats. The study's key findings reveal several critical insights that challenge existing paradigms in financial crime prevention. First, the scale and sophistication of APP fraud have reached unprecedented levels, with annual losses exceeding £1.1 billion in the UK and €4.3 billion across the European Economic Area. The research identified a dangerous asymmetry where criminals can rapidly deploy generative AI tools, including deepfake technologies and personalized social engineering lures, while financial institutions remain constrained by complex governance frameworks. This technological disparity has created what the study terms the "waterbed effect," where defensive measures merely displace fraudulent activity rather than eliminate it.

Second, the analysis of existing defensive strategies revealed fundamental limitations in reactive approaches. The UK's mandatory reimbursement scheme and the EU's Strong Customer Authentication, while providing consumer protection, fail to address root causes and inadvertently encourage fraud migration to less regulated channels. The research found that 80.8% of fraudulent payments successfully bypass Confirmation of Payee checks, highlighting the inadequacy of static safeguards against dynamic threats. Third, the proposed Adversarial and Dynamic Risk Management (ADRM) framework represents a paradigm shift from reactive to proactive fraud management. Through the integration of Generative Adversarial Networks, ensemble machine learning techniques, and explainable AI, the framework demonstrates the potential to achieve 70-90% reduction in undetected APP fraud. The synthetic dataset analysis revealed that vulnerable customers experience fraud rates 47% higher than the general population (5.89% vs 4.00%), emphasizing the need for targeted, adaptive protection mechanisms.

5.1 Implications for Cybersecurity and Financial Regulation

The findings of this research carry profound implications for both cybersecurity practice and financial regulatory policy, necessitating fundamental shifts in how institutions and regulators approach fraud prevention in the generative AI era.

5.2 Cybersecurity Implications

The research demonstrates that traditional cybersecurity approaches, built for static threat landscapes, are insufficient against AI-enabled adversaries. The ADRM framework's emphasis on adversarial thinking and continuous adaptation provides a blueprint for next-generation cybersecurity architectures. Financial institutions must transition from rule-based detection systems to dynamic, learning-enabled platforms that can anticipate and counter evolving attack vectors. The integration of explainable AI ensures that these sophisticated systems remain transparent and auditable, addressing regulatory requirements while maintaining operational effectiveness.

The study's findings also highlight the critical importance of cross-institutional collaboration in cybersecurity. The "waterbed effect" identified in the research underscores that isolated defensive measures are counterproductive, as they merely redirect threats to less protected vectors. This necessitates industry-wide adoption of federated learning mechanisms and threat intelligence sharing platforms that can collectively strengthen the financial ecosystem's resilience.

5.3 Financial Regulatory Implications

From a regulatory perspective, this research challenges the prevailing focus on liability allocation and consumer reimbursement. While these measures provide essential consumer protection, the findings demonstrate that they fail to address the systemic nature of APP fraud and may inadvertently incentivize criminal adaptation. Regulators must evolve beyond reactive, compliance-focused frameworks toward proactive, risk-based approaches that emphasize prevention over remediation.

The research suggests that regulatory frameworks should mandate the adoption of dynamic risk management systems similar to ADRM, establishing minimum standards for predictive fraud detection capabilities. This would require updating existing regulations such as PSD2 to incorporate requirements for AI-enabled fraud prevention, continuous model validation, and cross-border threat intelligence sharing.

Furthermore, the study's emphasis on explainable AI has significant implications for regulatory oversight. As financial institutions increasingly rely on sophisticated machine learning models for fraud detection, regulators must develop new competencies in AI governance, model interpretability, and algorithmic accountability. This includes establishing standards for synthetic data generation, adversarial testing, and continuous model monitoring.

6. Recommendations

The research concludes that addressing APP fraud in the generative AI era requires a coordinated response involving technological innovation, regulatory evolution, and industry collaboration. Financial institutions must invest in adversarial AI capabilities, while regulators must shift from reactive compliance monitoring to proactive risk assessment frameworks. The ADRM model provides a foundation for this transformation, offering a practical pathway toward more resilient and adaptive fraud prevention systems.

Ultimately, this research demonstrates that the battle against APP fraud cannot be won through traditional defensive measures alone. Success requires embracing the same technological innovations that criminals exploit, turning the tools of generative AI into instruments of protection rather than weapons of deception. The implications extend beyond financial services, offering insights applicable to any sector facing adaptive, AI-enabled threats in an increasingly digital world.

References

1) Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences, 12(19), 9637. https://doi.org/10.3390/app12199637
2) Bisht, U. & Pooja. (2025). Evolving Deepfake Technologies: Advancements, Detection Techniques, and Societal Impact. Don Bosco Institute of Technology Delhi Journal of Research. https://www.semanticscholar.org/paper/d2273fc34453ce6adc183dfdb0fcac4afa83aaa0
3) Braithwaite, J. (2024). 'Authorized push payment' bank fraud: What does an effective regulatory response look like? Journal of Financial Regulation, 10(2), 174–193. https://doi.org/10.1093/jfr/fjae006
4) Deloitte (2025). Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. Deloitte Insights.
5) EBA/ECB joint report on payment fraud data. (2024, August 2). Global Regulation Tomorrow. https://www.regulationtomorrow.com/the-netherlands/payments-the-netherlands/eba-ecb-joint-report-on-payment-fraud-data/

6) European Banking Authority. (2024). EBA Opinion on new types of payment fraud and possible mitigants. EBA.
7) European Banking Authority & European Central Bank. (2023). Report on payment fraud. Retrieved from https://www.eba.europa.eu/publications
8) European Central Bank & European Banking Authority. (2024). ECB and EBA publish joint report on payment fraud. European Union.
9) Fariha, N. et al. (2025). Advanced fraud detection using machine learning models: enhancing financial transaction security. doi:10.48550/ARXIV.2506.10842
10) Hermosilla, P., Berríos, S., & Allende-Cid, H. (2025). Explainable AI for forensic analysis: A comparative study of SHAP and LIME in intrusion detection models. Applied Sciences, 15(13), 7329. https://doi.org/10.3390/app15137329
11) House of Commons Library. (2024). Banking fraud and scams: Research briefing. UK Parliament. Retrieved from https://commonslibrary.parliament.uk/research-briefings/
12) Incode. (2025). Top 5 cases of AI deepfake fraud from 2024 exposed. Incode.
13) Jiang, J., Zhang, C., Ke, L., Hayes, N., Zhu, Y., Qiu, H., Zhang, B., Zhou, T., & Wei, G.-W. (2025). A review of machine learning methods for imbalanced data challenges in Chemistry. Chemical Science, 16(18), 7637–7658. https://doi.org/10.1039/d5sc00270b
14) KPMG International. (2025). What we have learned from UK Finance's latest annual fraud report. KPMG.
15) Mohamed, N. (2025). Artificial Intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67(8), 6969–7055. https://doi.org/10.1007/s10115-025-02429-y
16) Najar, A. V., Alizamani, leili, Zarqi, M., & Hooshmand, E. (2025). A global scoping review on the patterns of medical fraud and abuse: Integrating data-driven detection, prevention, and legal responses. Archives of Public Health, 83(1). https://doi.org/10.1186/s13690-025-01512-8
17) Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems. https://www.sciencedirect.com/science/article/pii/S0167923610001302
18) Ngan, J. (2025). "The view from below": Resistance and change in authorised push payment fraud. Journal of Economic Criminology, 9, 100166. https://doi.org/10.1016/j.jeconc.2025.100166
19) Onuh Matthew Ijiga, Idoko Peter Idoko, Godslove Isenyo Ebiega, Frederick Itunu Olajide, Timilehin Isaiah Olatunde, & Chukwunonso Ukaegbu. (2024). Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention. Open Access Research Journal of Science and Technology, 11(1), 001–004. https://doi.org/10.53022/oarjst.2024.11.1.0060
20) Payment Cards & Mobile. (2025). Report: A deep dive into payment fraud in the EU.
21) Payment Systems Regulator. (2023). APP scams performance report. Retrieved from https://www.psr.org.uk/publications/app-scams-performance-reports/
22) Rao, S.X. et al. (2025). Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention. IntechOpen eBooks. IntechOpen. doi:10.5772/intechopen.1009640

23) Sai, C., Amaresh, C., & Jancy, S. (2025). Online Payment Fraud Detection Using Exploratory Data Analysis. https://ieeexplore.ieee.org/abstract/document/10968090/
24) Skadden, Arps, Slate, Meagher & Flom LLP. (2024). New rules to tackle authorised push payment fraud.
25) Stoudt, S., Jernite, Y., Marshall, B., & Marwick, B. (2024). Ten simple rules for building and maintaining a responsible data science workflow. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012232
26) Talukder, Md. A., Khalid, M., & Uddin, M. A. (2024). An integrated multistage ensemble machine learning model for fraudulent transaction detection. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00996-5
27) TLT LLP. (2025). AI and financial crime - five big questions with Ben Cooper. TLT.
28) UK Finance. (2023). Fraud the facts 2023: The definitive overview of payment industry fraud. Retrieved from https://www.ukfinance.org.uk/policy-and-guidance/reports-publications/fraud-facts-2023
29) UK Finance. (2025). Annual fraud report 2025. Retrieved October 9, 2025, from https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2025
30) UK Finance. (2025). Fraud continues to pose a significant threat with over £1 billion stolen in 2024. UK Finance. https://www.ukfinance.org.uk/news-and-insight/press-release/fraud-report-2025-press-releas
31) Mohamed, N. (2025) 'Artificial Intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms', Knowledge and Information Systems, 67(8), pp. 6969–7055. doi:10.1007/s10115-025-02429-y.
32) Rao, S.X. et al. (2025) "Fraud Detection in E-Commerce: A Systematic Review of Transaction Risk Prevention," IntechOpen eBooks. IntechOpen. doi:10.5772/intechopen 1009640.
33) UK Finance (2025) Fraud continues to pose a significant threat with over £1 billion stolen in 2024. UK Finance. Available at: https://www.ukfinance.org.uk/news-and-insight/press-release/fraud-report-2025-press-releas