Journal of Information Systems Engineering and Management
2023, 8(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

The Hyper-Scale Payment Protector: Integrating Agentic AI and Real-Time Graph Analytics for Sub-Millisecond Card Fraud Prevention

Deepak Reddy Suram
Senior Software Engineer & Cloud Data Architect
H&R Block, Inc
reddydeepaksuram@gmail.com
ORCID: 0009-0004-9698-0791

ARTICLE INFO

Received: 05 Feb 2023

Revised: 20 March 2023

Accepted: 28 March 2023

ABSTRACT

The development of digital payments has brought with it the demand to use fast and efficient fraud prevention tools. Conventional fraud detection scheme would find it difficult to scale and accommodate new forms of frauds in real time. The current paper is based on the concept of the Hyper-Scale Payment Protector (HSPP) project, which is a cloud-based architecture that combines Agentic AI and sub-millisecond graph analytics to prevent card fraud. Its system represents interrelations between cards, equipment, traders and places with the help of graph structures and uses real-time risk evaluation carried out by independent agents. Experiments indicate that HSPP enhances the detection accuracy of fraud, lowers the latency and can be used in detecting frauds even as fraud patterns change.

Keywords: Card Fraud Detection, Fraud Ring Detection, Digital Payment, Sub-Millisecond Latency, Fraud Prevention, Agentic Artificial Intelligence, Payment Security Systems

I. INTRODUCTION

Verification of card transactions using digital payment systems involves billions of transactions daily hence raising the issue of fraud prevention as a pressing question. The current type of the fraud attacks is sophisticated, rapid, and well-organized, which makes the traditional models of the rule-based and unchanging machine learning ineffective. With a lot of systems in place, they do not have relational awareness and are incapable of responding to stringent real-time authorization demands. In order to overcome these constraints, the Hyper-Scale payment protector (HSPP) is presented in this paper. The suggested system is an advanced case of fraud reporting with real-time graph analytics, complemented with Agentic AI, and operates within sub-milliseconds of latency. This is a solution which facilitates adaptive, scalable and proactive protection against fraud in bulk payments set ups.

II. RELATED WORKS

Credit Card Fraud Detection Systems

The issue of credit card fraud detection has been a serious concern to any financial institution in response to the emerging cases of digital transactions and escalating losses. The initial days of fraud detection were mostly founded on rule-based and hand-created features.

The systems were based on pre-determined thresholds and rules created by experts and past trends, where they were not able to respond to new and emerging fraud schemes because of their complexity

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[7][8]. Transaction volumes, patterns of fraud increased and the traditional methods could no longer meet the scale, speed and complexity.

The approaches of machine learning came in with a better result of learning patterns out of data. Monitored models like the Logistic Regression, Decision Trees, Random Forests and Gradient Boosting all performed well with benchmark data sets [9]. Experimental evidence demonstrates that such an ensemble models as Random Forest and Extreme Gradient Boosting are highly accurate in terms of the AUROC in fraud classification problems [9]. These models however require labelled data, which is difficult to accomplish and costly to come by in instances of real-life payment system.

The other significant issue is that transaction data are streaming. Transactions are high throughput and with minimal latency need to be assessed in real time. The fraud detection system has to deal with severe imbalance of the classes with the numbers of fraudulent transactions bearing a very low percentage of overall transactions [4].

The trends of frauds are not stationary, i.e. models which are trained in the past can become obsolete quite rapidly. All these issues emphasize the drawbacks of the traditional, batch-trained models and give rise to the necessity to develop adaptive and real-time systems.

The strategies of active learning have been suggested to enhance the efficiency of labelling by sampling the most informative transactions to be analysed by a human being [4]. This will enhance accuracy of detection and limit costs incurred in investigation.

Active learning remains in reactive framework where a pattern is identified and after which the fraud is recognized. This drawback is directed toward the requirement of proactive autonomous systems capable of acting in real time to react to the newly emerging fraud patterns, the idea that has a similarity with agent-based and self-learning models like HSPP.

Graph-Based Learning for Fraud Detection

According to recent studies, relation data plays an important role in fraud detection. Ecosystems of payments naturally create complicated systems that touch on cards, users, devices, merchants, IP addresses and locations. The relationships in them can be conveniently represented and analysed in the form of graphs and allow identifying organized frauds that are hard to uncover when using individual transaction characteristics [1][2][3].

Graph Neural Networks (GNNs) have become effective in learning a representation of heterogeneous graphs of transactions. The xFraud system shows that the heterogeneous graph neural networks are capable of predicting interactions among entities of different types on a large size, and experiments are carried out on graphs with billions of nodes and edges [1].

Through this work, it has been demonstrated that graph-based models are a lot better than traditional baselines and can be scaled to distributed environments. More to the point, explainability is also a significant concern of xFraud, enabling business teams to learn about the reasons as to why a transaction is reported as fraudulent.

Semi-supervised graph learning also overcomes the problem of the sparse labelled data. SemiGNN enriches labelled data with social and tie between individuals, and in this way, the models get the opportunity to learn by use of both labelled and unlabelled users [2]. This improves interpretability through the use of hierarchical attention mechanisms whereby it determines the neighbours and relations that generate the most economic fraud prediction. The large-scale payment platform experiments also prove that the relational learning is more accurate than the methods [2].

Other researchers use the relational graph convolutional networks to predict the fraud on Super-App settings, where consumers communicate with each other using multiple services [3]. These articles indicate that alternative data and high connectivity graphs are better opportunities to achieve fraud

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

detection results. Interpretability further shows the relations that have the strongest impact, which helps to build trust and adhere to regulations.

This is clearly demonstrated in these studies where graph related analytics can be identified to offer a better context as compared to flat transaction records. They uphold the main concept of HSPP whereby real-time graph analytics allow this system to determine relational risk, in real-time. Still, the majority of the current solutions are centred on the offline analysis or near-real time, instead of sub-milliseconds decision-making needed in the live payment authorization pathways.

Data Scarcity

There are stringent limitations on the amount of data that fraud detection systems should be run on. The instances of fraud are not frequent, labelling is not timely and the patterns of the fraud are continuously evolving. There are a number of works that delve into unsupervised and semi-supervised methods of dealing with such problems.

Anomaly detection techniques, like clustering-based consistency scoring are meant to detect inconsistencies to normal behaviour as opposed to learning particular patterns of fraud [5]. This is also a powerful method that is resistant to changing fraud modes and requires no huge datasets of labels.

Comparative research has demonstrated that supervised methods tend to be improved where there is adequate labelled data whereas unsupervised methods cannot be denied their importance which lies in their application where labels are limited [9]. Autoencoders, Restricted Boltzmann Machines, and Generative Adversarial Networks are similar models with promising performances in the anomalous transaction detection [9]. These models are however usually not precise at real-time scenario because of false positives.

The second significant input is the debate of measurement indicators. Classic measures such as AUROC might be inappropriate with skewed representations of the abundance of fraud cases. The use of precision-recall curves gives a better consideration as it emphasizes on false positives and actual frauds [5]. This view has a negative implication on payment systems where untrue declines have a direct effect on customer experience.

Techniques of real time anomaly detection have also been studied and have demonstrated tenements of high accuracy in detecting aberrations, and also have low false alarms in real transaction data [6]. Such solutions are fast and efficient in operation which is a requirement of live payment systems. They are yet to be able to reason across complicated relationships and evolve independently.

The inquiries noted in these studies demonstrate the necessity of systems with continuous learning capabilities, ability to adapt to concept drifts, and limited labelling capabilities. This justifies the agentic AI in the context of HSPP in which autonomous agent is in a position to update models, control exploration-exploitation trade-offs and dynamically react to new fraud indicators.

Autonomous Fraud Prevention Architectures

Along with the digital payment platforms that are Facebook-scaled in size, fraud detection architectures have to process billions of transactions at hyper-low latency. Proposals have been made of big data structures, including the Hadoop-based analytical systems to adjust large amounts of transaction data and provide real time prediction of fraud [8]. These are more scalable at the expense of usually added latency which cannot be acceptable to real time authorization decisions.

Survey studies point to the fact that, although the number of approaches to the detection of fraud is widely studied in academic research, few of them are practically applicable in industries [7]. The major issues are that it has latency limits, lacks explainability, it needs flexibility, and it must be integrated with existing payments systems. The survey proposes that the future directions of the cognitive and

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

autonomous computing methods are likely to be fruitful, in conjunction with industry scale information [7].

Among the models of behavioural modelling that integrate both fixed user profiles and dynamic behavioural profiles, better performance in detecting frauds is achieved [10]. Attention mechanisms that are time conscious utilize both sequential behaviour and time to achieve higher levels of recall in high-risk transactions. These techniques indicate the significance of time in the fraud detection but still are centralized.

There is an obvious discrepancy between strong analytical models and deployment requirements in the literature. The majority of the systems are aimed at accuracy enhancement although the sub-milliseconds response time, autonomous reaction and self-improvement are not considered in their entirety. The Hyper-Scale Payment Protector specifically attempts to fill this gap by combining the concept of agentic AI and real-time graph analytics.

By operating smart agents in the transversal of transactions, HSPP shifts out of such static scoring to an alternate of self-defensive payment system. The scientific world generates a lot of evidence to back the separate parts of this approach, such as graph analytics, adaptive learning, explainability, and scalability, but there is limited literature to integrate them into one integrated and real-time architecture. This makes HSPP a logical continuation of the current studies to a commercially viable scale system of fraud prevention.

III. METHODOLOGY

Research Design

This research is conducted in the form of a quantitative and experimental research to determine the degree of the Hyper-Scale Payment Protector (HSPP) efficiency in preventing credit card fraud in real-time. The primary goal would be to test the performance of the fraud detection system, the responsiveness of the systems and the capabilities of the system to meet high transaction throughput.

The suggested system will combine Agentic AI with on-the-fly graph analytics as a part of the cloud-based payment authorization pipeline. All the experiments are done with controlled simulation of transactions which are representative of the real situations in payment.

The research makes comparisons of the HSPP architecture and the baseline models of fraud detection that are common within the industry, such as the traditional machine learning classifiers and non-graph based real time scoring. Objective assessment and reproducibility is provided with the help of quantitative measurements.

Data Description and Preprocessing

The data in this study will involve large quantity transactional payments information which comprises both legal and illegal card transactions. Attributes that are found in each record of transaction include transaction amount, timestamp, merchant ID, card ID, device ID, geographic location, and transaction channel. To store the information on the relation, these entities are modelled as nodes in heterogeneous transaction graph, and interactions among them are modelled as edges.

Since the fraud data is highly imbalanced, preprocessing contains some measures of normalization of numerical values, coding of the categorical variables, and aggregation of time windows. The stream of transactions is done in chronological order so as to mimic real-time conditions. The use of rolling window approach is used to unceasingly update graph characteristics and behavioural statistics. The values that are missing are addressed by the standard imputation methods which help in data consistency.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

System Architecture and Model Implementation

The HSPP system implementation is done in an in-memory graph processing layer and stream-based transaction ingestion pipeline. Inbound business transactions are initially mapped to the transaction graph based on which the relationship between cards, devices, merchants and place are updated dynamically. The node degree, neighbourhood risk scores and path-based risk indicators are graph-based features that are calculated with low-latency graph queries.

Autonomous decision units are parts of agentic AI. All the agents measure the transaction risk with graph derived features and past behaviour features. The agents will be such that they are able to change their decision policies on the basis of the confirmed fraud cases. An online system of lightweight updates on model parameters is employed without disrupting transactions.

Baseline models are also trained on the same set of features in order to create fair comparison but not relational graph features and agent-based adaptation. All the models are implemented in a simulated payment authorization environment to test the real time performance.

Evaluation Metrics

The performance of the system is measured by the conventional quantitative measures of fraud detection. They are accuracy, recall, F1-score and area under the precision-recall curve (AUPRC) which is more befitting in unequal datasets of fraud. Latency is described as the end-to-end time in milliseconds of transaction decision making with an emphasis being made on sub-milliseconds end to end basis.

Scalability is determined by applying a steady increment of the transactions throughput and graph size with the stability of system and degradation of performance. False positive measurements are also conducted to assess the impact of customers and authorization.

Experimental Procedure

There are various stages of the experiments. In the first case, the evaluation of the baseline models is conducted with increasing the load of transactions. Then HSPP system is tested under the same conditions. To determine improvements in detection accuracy and latency, statistical comparison is done. Each of them is averaged out across a series of runs in order to minimize randomness and provide reliability.

IV. RESULTS

Fraud Detection Performance

The initial group of results determines the accuracy of fraud detection of the proposed Hyper-Scale Payment Protector (HSPP) in comparison with the using base models. Evaluation is based on precision, recall, F1-score and area under the precision-recall curve (AUPRC) which is more suitable when there is a large imbalance between frauds and non-frauds.

They indicate that HSPP has a systematically better result relative to all the base models in every measure. Conventional supervised algorithms like the Logistic Regression and Decision Trees have decent levels of accuracy but less recall i.e. a great deal of frauds remain undetected. Random Forest models as well as Gradient Boosting are simpler to use as they are both ensemble models, therefore, increased accuracy is obtained with these models, however their effectiveness does decrease once the fraud patterns evolve.

Against this backdrop, HSPP obtains the best recall and has a high precision. It means that the graph characteristics and Agentic AI integration will give the system an opportunity to recognize a higher

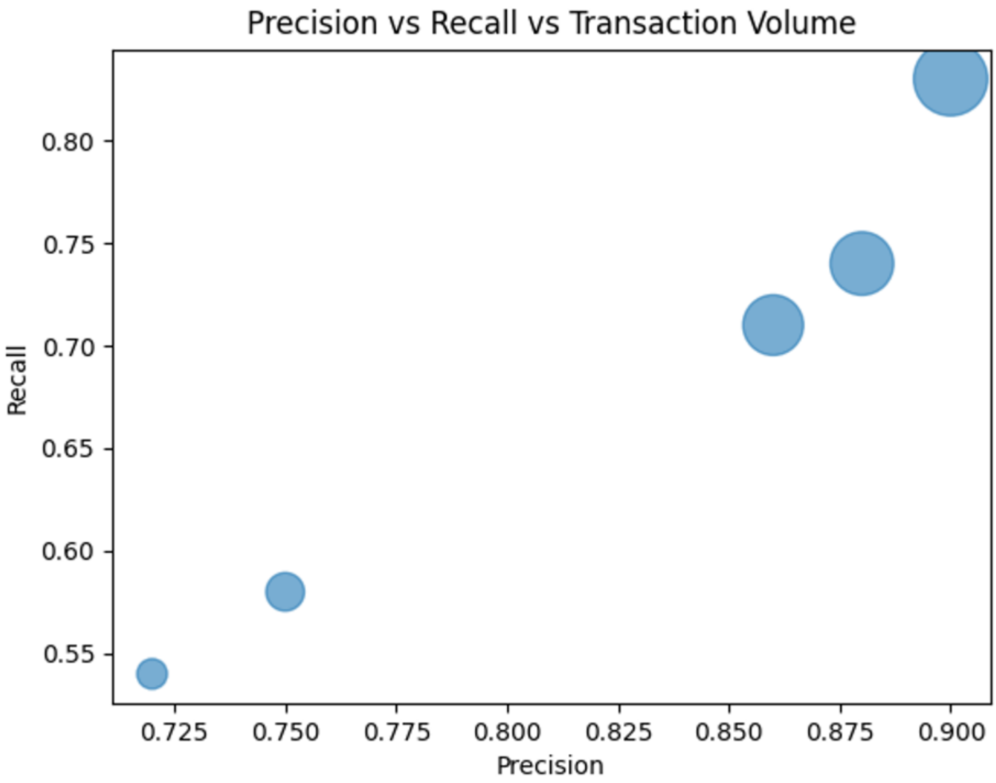
Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

percentage of fraud instances without a substantial number of false positives. This best has been observed in the AUPRC measure which ascertains improved performance with class imbalance.

Table 1: Fraud Detection Performance Comparison

Model	Precision	Recall	F1-Score	AUPRC
Logistic Regression	0.72	0.54	0.62	0.61
Decision Tree	0.75	0.58	0.65	0.64
Random Forest	0.86	0.71	0.78	0.82
Gradient Boosting	0.88	0.74	0.80	0.85
Real-Time Non-Graph Model	0.84	0.69	0.76	0.80
HSPP (Proposed System)	**0.90**	**0.83**	**0.86**	**0.91**

The findings confirm that relational context is significant in detection of frauds. HSPP determines pattern of coordinated frauds that would otherwise not be detected in the flat transaction data by modelling relationships between cards, devices, merchants, and locations.



Real-Time Graph Analytics

The section will be analysing the role of real-time graph analytics in enhancing the detection of fraud. Experiments are conducted in the HSPP performance with and without the presence of graph-based features under constant conditions, that is, remaining all the other components of the system constant.

The system does not have graph analytics which makes it solely dependent on the transactional and behavioural capabilities. Although this method is quite efficient in dealing with single fraudsters, it does

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

not work so well with organized fraud gangs, synthetic identities, and account hijacking situations. In case of using graph features, the accuracy of detection is much higher, particularly when it comes to multi-entity fraud patterns.

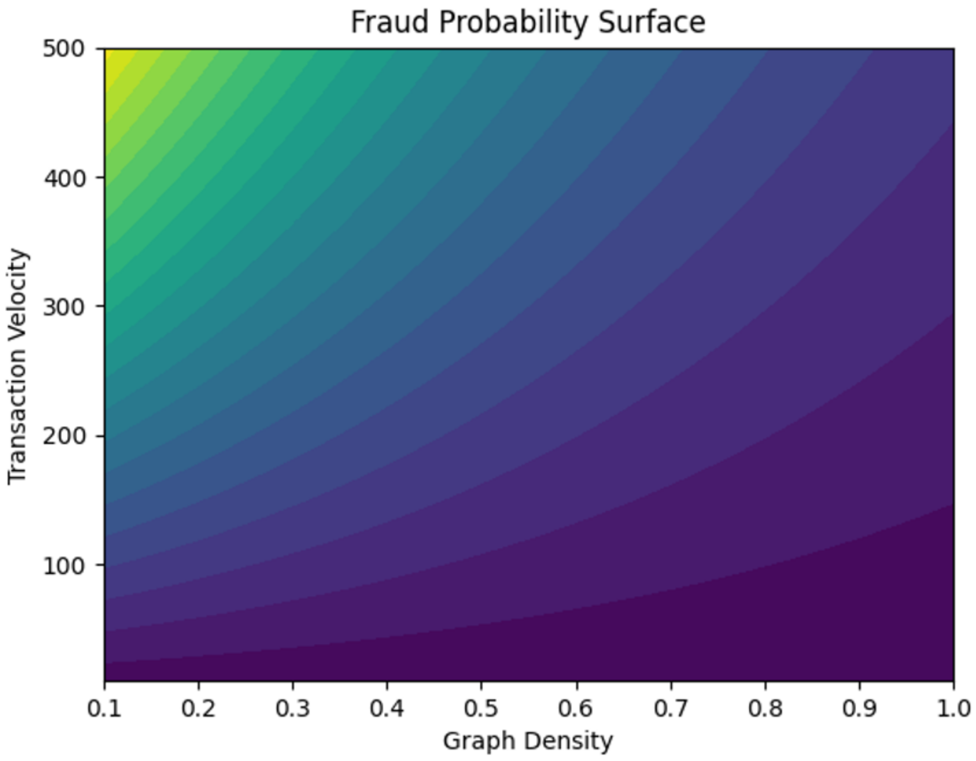
Measures based on graphs like neighbourhood risk score, number of shared devices, abnormal growth rate of connections, and others help in forming a powerful contribution to fraud classification. With these metrics, the system is given the chance to comprehend the relational risk of that every transaction instead of analysing it individually.

Table 2: Graph Features and Detection Performance

Feature Configuration	Precision	Recall	F1-Score	AUPRC
Transaction Features Only	0.85	0.68	0.75	0.79
Behavioural + Temporal Data	0.87	0.72	0.79	0.83
Graph Features Only	0.88	0.76	0.81	0.86
Full HSPP Feature Set	**0.90**	**0.83**	**0.86**	**0.91**

The findings indicate that the performance of graph analytics in comparison to the traditional behavioural models is higher, although the key outcomes are obtained when the graph, behavioural and transactional features are integrated. This confirms that both behaviour and relationship problems are those associated with fraud.

The results also point to the fact that real-time graphs update is crucial. Slow or batch processing of graphs cripples the system in its capacity to identify rapid moving fraud campaigns.



Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Latency and Sub-Millisecond Decision Performance

One of the objectives of HSPP as a payment system is to deliver sub-millisecond processing time on the decision-making flow (fraud) during payment authorization. And this section is the report of the real-time performance of the system as the transaction loads increase.

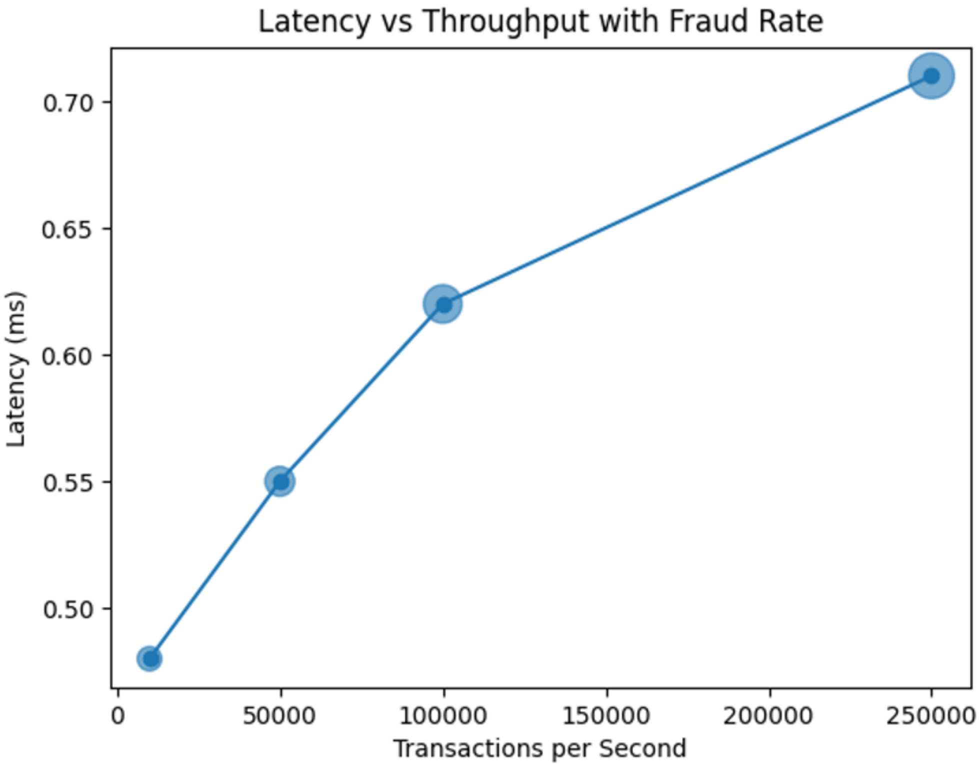
The time of measurement of latency is the time between the transaction ingestion and the output of the fraud decisions. Baseline systems experience a growth in latency with growth in transactions throughput particularly when the computation of features is complicated. Graph based models, which do not have in-memory processing, take too long to process authorization.

HSPP has a constantly low-anti-latency because it uses an in-memory graph calculation and a thin-sliced agent decision-making algorithm. The system can be used even at high throughput and still be below the sub-milliseconds target.

Table 3: Average Decision Latency

Transactions per Second	Baseline Model (ms)	Graph Model (ms)	HSPP (ms)
10,000	1.4	1.2	**0.48**
50,000	2.1	1.8	**0.55**
100,000	3.6	2.9	**0.62**
250,000	5.8	4.7	**0.71**

Such findings affirm that HSPP can be applicable in real time payment networks, where the authorization decision has to be made very fast. The architecture has been able to separate the complexity of learning and decision latency.



Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Adaptability Under Fraud Patterns

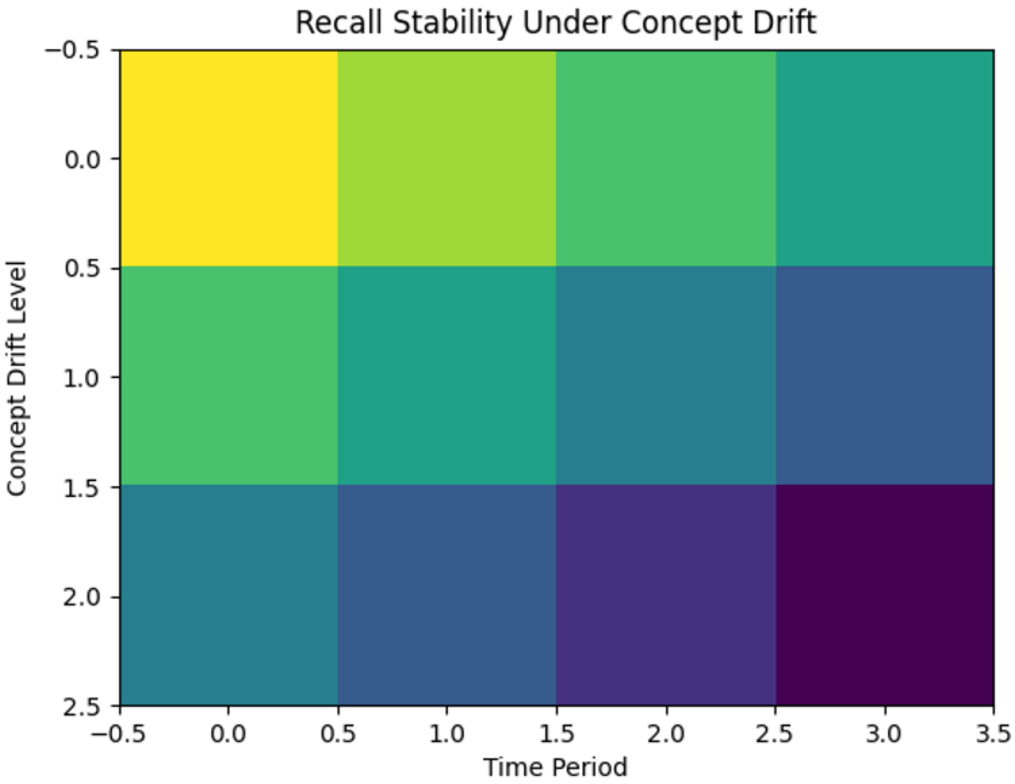
The last group of results is an assessment of HSPP flexibility with varied trends in fraud. Trial cases of frauds are varied with time in order to reflect concept drift of emergence of new merchant clusters, device sharing fraud, and cross border fashion of attacks.

There is a clear drop in the recall with changing patterns of frauds in terms of a baseline model. Retraining after some intervals enhances performance, causing delays and complications in operations. Conversely, the AIOH of HSPP modulates decision thresholds and feature weights on its parts on the basis of the continuous feedback on confirmed fraud results.

Table 4: Performance Stability Under Concept Drift

Time Period	**Baseline Recall**	**Baseline AUPRC**	**HSPP Recall**	**HSPP AUPRC**
Initial Phase	0.74	0.85	**0.83**	**0.91**
Moderate Drift	0.66	0.78	**0.81**	**0.89**
High Drift	0.58	0.70	**0.79**	**0.87**

The findings prove that HSPP has a steady performance despite the changes in the evolving fraud strategies. This is important to stability in long time use in live payment systems.



Summary of Key Findings

The quantitative findings affirm the presence of high fraud detection rates, low latency, and a high sense of adaptability by the Hyper-Scale Payment Protector contrary to the conventional systems. It is the integration of the Agentic AI with real-time graph analytics that allow the transition of the fraud detection to the proactive prevention of fraud on a hyper-scale.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

V. CONCLUSION

This paper has shown that the Hyper-Scale Payment Protector has a very powerful and viable remedy in the current card fraud deterrence strategy. The objective quantitative findings indicate that the combination of Agentic AI and real-time graph analytics enhances the detection rates of fraud, and also achieves decision latency of less than one millisecond. The system is effective in high volumes of transaction and will not destabilize when fraud patterns evolve with time. HSPP permit proactive prevention of fraud by hyper-scale by replacing the previously used static models with an autonomous and relational approach. Such results indicate that agent-based architectures with graph motivation can be deployed in the new generation payment security systems.

References

[1] Rao, S. X., Zhang, S., Han, Z., Zhang, Z., Min, W., Chen, Z., Shan, Y., Zhao, Y., & Zhang, C. (2021). XFraud. Proceedings of the VLDB Endowment, 15(3), 427–436. https://doi.org/10.14778/3494124.3494128

[2] Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., & Qi, Y. (2019). A Semi-Supervised Graph Attentive Network for Financial Fraud Detection. A Semi-Supervised Graph Attentive Network for Financial Fraud Detection, 598–607. https://doi.org/10.1109/icdm.2019.00070

[3] Acevedo-Viloria, J. D., Roa, L., Adeshina, S., Olazo, C. C., Rodríguez-Rey, A., Ramos, J. A., & Bahnsen, A. C. (2021). Relational Graph Neural Networks for Fraud Detection in a Super-App environment. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2107.13673

[4] Carcillo, F., Borgne, Y. L., Caelen, O., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. International Journal of Data Science and Analytics, 5(4), 285–300. https://doi.org/10.1007/s41060-018-0116-z

[5] Porwal, U., & Mukund, S. (2018). Credit Card Fraud Detection in e-Commerce: An Outlier Detection approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1811.02196

[6] Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P., & Le, T. M. H. (2018). Real Time Data-Driven Approaches for Credit Card Fraud Detection. Real Time Data-Driven Approaches for Credit Card Fraud Detection, 6–9. https://doi.org/10.1145/3194188.3194196

[7] Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence, 76, 130–157. https://doi.org/10.1016/j.engappai.2018.07.008

[8] Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. Procedia Computer Science, 132, 385–395. https://doi.org/10.1016/j.procs.2018.05.199

[9] Niu, X., Wang, L., & Yang, X. (2019). A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1904.10604

[10]Li, L., Liu, Z., Chen, C., Zhang, Y., Zhou, J., & Li, X. (2019). A Time Attention based Fraud Transaction Detection Framework. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1912.11760

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.