

International Conference on Internet of Things, Computing, and AI Applications  
ICIoTCAA26-Conference Proceeding | IJMIET  
Doi: [10.67228/30715628/ICIoTCAA-2026P102](https://doi.org/10.67228/30715628/ICIoTCAA-2026P102)  
PP. 17-32

*Original Article*

## Real-Time Overpayment Detection at Point of Payment: Methods and Architecture to Identify Likely Overpayments before Claims are Paid

**Appala Nooka Kumar Doodala**

Manager Quality Assurance at Cognizant, USA.

### ABSTRACT

Overpayments resulting from errors in healthcare claims processing are one of the major causes of financial loss for payer organizations, government programs, and private insurers. Traditional retrospective audits usually locate the overpayments several weeks or months after the funds have already been disbursed and this leads to the recovery processes that are costly, provider abrasion, and the inefficiencies of administration. The real-time or near-real-time identification of overpayment at the POP intends to eliminate the occurrence of such errors by the direct integration of advanced analytics, rule engines, machine learning models, and scenario-based risk scoring into the claim adjudication workflow. The present document discusses a complete system for detecting overpayments in real-time as well as the foundational technical infrastructure, data models, streaming pipelines, business rules, supervised and unsupervised anomaly detection methods, alerting manners, and the junctions with the payer claims systems. The case study illustrates the payer environment-based deployment of a real-time detection engine encompassing system performance, model accuracy, financial impact, and operational considerations. Findings show that real-time detection may bring down overpayments to nearly 40–60% of the levels that retrospective methods can achieve, thereby significantly enhancing payment integrity. The study ends with the suggestions for the next steps including real-time provider behavioral analytics, federated learning for cross-payer intelligence, and integration with national fraud detection networks.

### KEYWORDS

Overpayment Detection, Payment Integrity, Real-Time Analytics, Claim Adjudication, Machine Learning, Predictive Scoring, Healthcare Claims, Fraud Waste and Abuse (FWA), Anomaly Detection, Streaming Architecture, Point of Payment, Risk Scoring.

## I. INTRODUCTION

The healthcare claims ecosystem is a very complex one as it handles a massive number of transactions daily. Large U.S. health plans usually go through millions of claims every day. Each claim is regulated by a complicated network of clinical codes, reimbursement policies, contractual agreements, and utilization management rules. If there are any slight inefficiencies or inaccuracies in claims within this vast operational landscape, they can be translated into huge financial losses. Even though considerable funds have been pumped into the technology of adjudication over the last few decades, the industry is still facing the problem of payments that are not appropriate – overpayments, waste, abuse, and in some instances, deliberate fraud. Collectively, these leakages cost the American healthcare system more than \$100 billion yearly, which is the reason why payer financial performance is hindered, premiums are elevated, and in the end, patients and the healthcare economy as a whole are burdened.

One of the main reasons for these losses is the fact that they still depend on post-payment integrity processes, which are carried out after the detection of anomalies in the files of the disbursed funds. Even though retrospective audits, provider outreach, and recovery programs play a very important role in the protection system, they are still very slow, labor-intensive, and confrontational. What is more, once an improper payment is made, the chances of recovery are very low – providers may contest findings, lack the required documentation, or simply be unwilling or unable to repay. Hence, the commercial and government payers have made prevention at the point of payment their strategic priority.

But to be able to achieve real-time prevention is not an easy task. For instance, claims may differ in structure and clinical context, whereby providers' billing patterns may also differ; fraud schemes, on the other hand, are always changing. On top of that, technical limitations in legacy adjudication platforms, fragmented data sources, and inflexible rules engines are all obstacles to innovation. Any payer who is willing to set up a strong, scalable, and proactive payment integrity program needs to have a modernized detection architecture in place – a system that integrates expert rules, machine learning (ML), anomaly detection, and explainable decisioning while being able to perform seamlessly within millisecond-level adjudication windows.

### 1.1. Challenges in Preventing Overpayments

Efforts to prevent overpayments at the very moment that a payment is made are hindered by a combination of operational, technical, and analytical barriers.

- The sheer number of claims and how fast they come make it very difficult to carry out the analysis in real time. For instance, large payers may handle up to tens of millions of claims every month, hence the adjudication systems that make the payment decisions are left with only milliseconds to do their job.

- The extremely complicated reimbursement rules that cover ICD-10, CPT/HCPCS, DRGs, APCs, modifiers, bundling logic, medical necessity policies, and contract-specific edits, require not only precision but also adaptability.
- The variability of providers causes that the billing patterns may change, thus the rules alone cannot be used to categorize these patterns as legitimate or suspicious with certainty.
- The fragmentation of data in clinical systems, provider directories, historical claims, utilization data, and external databases hinders the unification of analytics during adjudication.
- The fact that post-payment audits are still heavily relied upon, causes that the detection is delayed, there is an administrative burden, providers get abraded, and, in the end, overpayments remain unrecovered.
- False positives may delay the legitimate payments, thus the provider relationships and the member experience can be negatively affected.
- The limitations in integration with the legacy adjudication engines create difficulties in the deployment of advanced ML models or real-time anomaly detection algorithms.

All these hurdles, on the whole, result in a structural gap existing between what is already possible in terms of adjudication and the industry's requirement for payment integrity in real-time.

### 1.2. Problem Statement

Improper payments are still widely spread basically because of inadequate claims systems in spite of major technological advancements made along the road. Current claims systems are typically blamed for the failure in:

- Not enough real-time detection of abnormal billing activities, medical inconsistencies, or new fraud trends.
- Very limited contextual understanding like provider history, peer group comparisons, or past audit results during adjudication.
- Very inflexible rule-based decisioning that can neither rapidly adjust to newly developed schemes nor clinical scenarios deeply.
- The absence of a comprehensive, real-time decisioning structure that integrates rules with sophisticated analytics, machine learning, and anomaly detection is another factor.

### 1.3. Motivation

Modernizing real-time payment integrity is highly necessary, as highlighted by several influential factors or drivers:

- Financial impact: The overpayment and improper claims amount to more than 100 billion dollars yearly in the U.S., thus, there is a great potential for healthcare payers to achieve rapid financial returns from these low-hanging fruits.
- Regulatory expectations: CMS and other agencies are progressively requiring the implementation of program-integrity measures that focus on upstream, preventive activities.

- **Business efficiency:** The implementation of faster and more accurate claims processing will lead to a decrease in the occurrence of rework, appeals, and administrative overhead, thus, releasing resources that can be used in other productive activities.
- **Provider abrasion reduction:** Early issue prevention, thus, avoiding the inconvenience of post-payment audits and documentation requests that disrupt providers' daily routines.
- **Intelligence/ML enhancements:** The new analytic techniques—deep learning, anomaly detection, pattern mining, and LLMs—that AI/ML research has developed enable the detection of capabilities that were previously impossible in real-time environments.
- **Competition differentiation:** Health plans with best-in-class program-integrity systems can reduce loss ratios, improve pricing competitiveness, and increase member/provider satisfaction.

## 2. LITERATURE REVIEW

For a considerable time, one of the major issues that have been at the core of the worries of healthcare payers, government agencies, and regulatory bodies is payment integrity. The writings include the release of traditional retrospective audit methodologies, the first attempts at real-time analytics, and the increasing interest in machine learning-based fraud and overpayment detection. Nevertheless, a great number of these works, i.e., the ones mentioned, have left significant holes in the conversion of analytical advances into scalable, real-time production systems that can function at adjudication speed. This review workers have researched and industry practice has been synthesized to form the review of the modern real-time payment integrity architectures.

### 2.1. Traditional Payment Integrity Approaches

Most times, payment integrity involved going back over the records and thoroughly checking the details of the payments made. It was common that the payers or the vendors hired by them would check the claims and accounts only after the money had been paid. There are quite a few studies showing that post-payment audits lead to the large amounts of money being recovered, though, at the same time, they refer to the drawbacks of these audits such as slow detection, low recovery rates, and high administrative costs for both payers and providers. The retrospective programs usually consist of medical record examinations, statistical sampling, clinical validation, and appeals processes, all of which lengthen the improper payments lifecycle.

**Table 1: Comparison of Payment Integrity Approaches**

| Dimension          | Retrospective Audits                            | Prepayment Review                               | ie Detection at Point of                    |
|--------------------|-------------------------------------------------|-------------------------------------------------|---------------------------------------------|
| Detection Timing   | Weeks or months after payment                   | Before payment, but outside adjudication flow   | During adjudication, before payment release |
| Primary Techniques | Manual audits, sampling, post-payment analytics | Static rules, limited heuristics, manual review | Rules + ML + anomaly detection in real time |
| Recovery           | Low to moderate (often                          | Moderate                                        | High (prevents                              |

|                                    |                                             |                                            |                                                 |
|------------------------------------|---------------------------------------------|--------------------------------------------|-------------------------------------------------|
| Effectiveness                      | <30%)                                       |                                            | payment leakage entirely)                       |
| Operational Cost                   | High (audits, appeals, recovery processes)  | Moderate (manual reviews)                  | Lower long-term cost after automation           |
| Scalability                        | Limited by human effort                     | Limited by review capacity                 | Highly scalable via streaming architectures     |
| False Positives                    | Identified late, costly to correct          | High due to static rules                   | Reduced through contextual ML scoring           |
| Provider Impact                    | High abrasion, disputes, delayed recoveries | Moderate delays and documentation requests | Minimal disruption, fewer post-payment audits   |
| Adaptability to New Fraud Patterns | Low                                         | Low to moderate                            | High (model retraining and anomaly detection)   |
| Regulatory Alignment               | Reactive compliance                         | Partial prevention                         | Proactive payment integrity and compliance      |
| Overall Effectiveness              | Reactive and inefficient                    | Preventive but limited                     | Proactive, precise, and operationally efficient |

Many payers have also either turned towards or settled on prepayment reviews to compensate for the inefficiencies of their retrospective methods. Under these programs, some claims are flagged even before they are judged. Still, in most cases, the prepayment review board depends on manual checks, static rules, or a limited number of ML heuristics, which results in scalability issues and a large number of false positives. The literature argues that if the prepayment interventions are a step forward in prevention, they are still very limited in scope and are, as a rule, used only for high-risk providers or specific billing categories.

Moreover, the narrative around payment integrity has been significantly influenced by the role of third-party vendors in the audit-based recovery process over the years. Usually, these vendors integrate several tools such as a proprietary rule library, contract compliance algorithms, and a human-driven clinical validation module. Even though these vendors are good at identifying the ret

Another primary method is that of the payer's rule-based systems, which are either off-the-shelf or internally developed (for instance, Medicare's National Correct Coding Initiative (NCCI) edits). These systems are designed to incorporate the intricacies of the policy, for instance, the constraints, bundling rules, as well as that of mutually exclusive procedures. While the use of rule-based mechanisms is vital for the uniformity of coding compliance, it is a deterministic method by nature

and, therefore, incapable of discovering new cheating methods or subtle anomalies in provider behavior. As far as coding standards are concerned, they will continue to change, and fraud tactics will become more and more sophisticated. However, at the same time, the inflexibility of rule-based models limits their ability to proactively detect fraudulent activities.

## 2.2. Real-Time Processing and Analytics in Healthcare

With adjudication becoming increasingly automated, the involved parties (researchers and industry stakeholders) have considered the possibility of real-time processing frameworks which would be directly integrated into the claims systems. Initially, the work done in this area was mainly concerned with the embedding of rule engines into adjudication platforms so that it would be faster and more systematic to enforce the established editing rules. Nevertheless, the engines were not intended to include dynamic, data-driven risk scoring or predictive analytics.

A major change was the setting up of the CMS Fraud Prevention System (FPS). It was the first step in applying advanced analytics to Medicare claims at or near the point of payment. The FPS employed predictive models, anomaly detection methods, and provider risk profiling to locate the most fraudulent claims in almost real-time. The system proved that analytics could be used within adjudication windows even though FPS was not fully synchronous and usually required near-real-time batch processing with some latency. Anyway, it gave both a technological and a regulatory precedent for the integration of the predictive models into the payment workflows.

Moreover, vendors from the private sector such as Cotiviti, Optum, SCIO (now EXL), and Change Healthcare have launched analytics-driven payment integrity programs that could be appealing to different groups of customers. In spite of the fact that these platforms incorporate machine learning and have advanced pattern-recognition capabilities, the majority of them execute post-payment or near-real-time batch modes, rather than simultaneously with claim adjudication.

## 2.3. Machine Learning in Claim Fraud and Overpayment Detection

The academic and business literature on machine learning for claims analysis is extensive and covers both supervised and unsupervised methods.

Supervised learning methods are the main focus of most publications and have been illustrated to work effectively by:

- Random Forest (RF) for doctor-level fraud classification and claim-level anomaly scoring,
- XGBoost for high-dimensional fraud pattern extraction and feature interactions,
- Naïve Bayes (NB) for probabilistic modeling of coded claims events,
- Support Vector Machines (SVM) for distinguishing subtle coding abuse behaviors.

Generally, these algorithms depend on historical cases that have been labeled (e.g., prior audits or confirmed fraud events) and thus, have shown good results in retrospective datasets. However, in

a real-time scenario, supervised models are confronted with problems such as feature freshness, label scarcity, and latency constraints.

There are also unsupervised and semi-supervised techniques to solve these problems, which recognize anomalous or deviated activities in billing without the need for explicitly labeled data. Some of the methods that are employed include:

- Isolation Forest, a technique that can identify the most unusual claim behaviors,
- Local Outlier Factor (LOF), which is able to detect anomalies that are context-dependent,
- Variational Autoencoders (VAEs), that are suitable for reconstructive anomaly detection in high-dimensional data.

Many papers suggest hybrid architectures that use fixed provider features jointly with variable claim-level features. These hybrid models result in higher precision and better performance; however, the publications seldom talk about the requirements of a system that can carry out such models in real-time and on a large scale.

### 3. PROPOSED METHODOLOGY

#### 3.1. High-Level Architecture Overview

The methodology being proposed is based on a modular, real-time architecture that is aimed at functioning within the healthcare claim adjudication environments' strict performance constraints. Fundamentally, the system moves claim events via a real-time ingestion layer, which takes the claim information from EDI submissions, provider portals, or internal systems, and then structures them into event streams. The event streams are delivered to a high-throughput platform like Kafka, Pulsar, or Kinesis that guarantees the reliable, scalable, and ordered delivery of claim events. Then, a real-time feature store adds the context to these events, thus the ML layer can get the provider history, claim patterns, and other features that it needs for making accurate predictions. At the same time, a rule engine performing the checked payer and regulatory logic keeps the code, policy, and benefit verification that are the basis of the process, done in a consistent manner.

The ML inference layer is doing a similar job as the rule engine and thus can be found working alongside each other. They communicate through low-latency gRPC or REST calls generating risk probabilities related to overpayment or billing that behave in an unusual manner. The outputs that result from here are routed to a single risk scoring engine, which uses outcomes from the rule engine, ML predictions, and anomaly indicators to come up with an all-inclusive risk score. A decision engine, therefore, based on this score, determines the final act and communicates back with the car, pending-for-review, or routing to the investigation team as the possible next step. Through an audit and explainability layer, the system keeps full transparency of the process, thus ensuring that every decision made is easy to follow and can be defended. All of them together make up a real-time, integrated system that is at the same time very accurate, fast, and operationally aligned with the existing payer workflows.

**Table 2: End-to-End Architecture and Analytical Components for Real-Time Overpayment Detection**

| Component                        | Description                                                                                        | Key Analytics / Functions                                                                    | Technologies Commonly Used             |
|----------------------------------|----------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|----------------------------------------|
| Claim Ingestion Layer            | Captures incoming claim transactions in real-time from EDI, provider portals, or internal systems. | Cleansing, normalization, duplicate detection.                                               | Kafka, Kinesis, Pulsar                 |
| Real-Time Feature Store          | Supplies enriched provider, member, historical, and contextual features instantly.                 | Peer comparison, cost history lookup, provider behavioral features.                          | Redis, Bigtable, DynamoDB              |
| Rule-Based Engine                | Applies deterministic policy enforcement and medical coding validations.                           | NCCI edits, bundling logic, frequency limits, medical necessity checks.                      | Drools, Optum CES, Custom rule engines |
| Machine Learning Inference Layer | Generates predictive risk scores for overpayment likelihood and anomaly patterns.                  | Supervised (XGBoost/LightGBM), Unsupervised (Isolation Forest/Autoencoders), Hybrid methods. | gRPC/REST microservices, GPUs/CPU      |
| Risk Scoring Engine              | Combines rule outputs, ML scores, anomaly indicators into a unified score.                         | Weighted scoring, confidence indexing, business logic routing.                               | Custom scoring pipelines               |
| Decision Engine                  | Determines whether to auto-pay, pend, or escalate claim review.                                    | Thresholding, clinical routing, financial hold workflows.                                    | Microservices, BPM engines             |
| Explainability & Audit Layer     | Ensures transparency, compliance, and traceability of every decision.                              | SHAP/LIME explanations, rule-trigger logs, full claim trace.                                 | SHAP, LIME, ELK Stack                  |
| Adjudication System Integration  | Incorporates all analytics outputs back into legacy or modern adjudication platforms in <200 ms.   | Payment approvals, pends, flags, alerts.                                                     | API gateways, FHIR/EDI integration     |

### 3.2. Data Requirements

How well real-time payment integrity works is very much dependent on how wide and how good the data supporting it are. Risk assessment is basically done on the basis of claim data which include headers, service lines, diagnoses, procedures, modifiers, and financial fields. Provider data gives more background by providing details of specialties, locations, sanction status, historical utilization patterns, and network characteristics. So, when these datasets are mingled with member demographic and geographic information, the system gets the power to spot the aberrant billing patterns which are relative to peer groups and regional trends. Information on prior authorization,

historical paid claims, and known overpayment cases lead the ML models in identifying the patterns even further by capturing longitudinal patterns and historical outcomes.

The system also utilizes different external datasets apart from payer data in order to achieve the detection accuracy. Clinical coding standards like ICD, CPT, HCPCS, DRGs, and APCs help in making sure that claims are assessed against medically correct structures. A connection with CMS fee schedules and provider exclusion lists gives the context of regulation and compliance. Benefit plan configurations bring the contractual nuance into the risk assessment process, thus, claim analysis will be accounting for the specific coverage rules. Hence, these datasets are like building blocks of a comprehensive, multi-dimensional foundation that enables sophisticated and reliable real-time analytics.

### 3.3. Rule-Based Edits

Rule-based editing continues to be a major factor in achieving payment integrity. This is because it offers a clear, policy-driven validation that can be determined by rules. The rules check if the services are within the coverage, if they have correct codes, if they are medically necessary and if they are not duplicative. By implementing the policies of the payers, medical standards and the requirements of the regulators, the rule engine is the one that actually makes sure that the basic safety features are extended to all claims on a consistent basis. The system evaluates coding combinations, checks the diagnosis-procedure alignment, and looks at the contextual appropriateness of the services such as frequency limits and bundling logic. In addition, duplicate claims and already known upcoding patterns are detected at this level, thus, straightforward errors or policy violations are identified at the initial stage.

Though rule-based editing is of great importance, the method may not be sufficient in detecting fraud schemes that are newly created as well as subtle behavioral anomalies. For this reason, the rules engine has been created in such a way that it collaborates with ML models. It gives the structure and the deterministic logic, and ML handles complex, non-linear risk patterns. The outcome of the rule can be seen in the form of severity indicators or policy flags, which later on will be used for the computation of the final risk score. The combination of these two techniques enables the system to be in line with the policy enforcement which is deterministic while the machine learning is probabilistic, thus, the accuracy is at the maximum level and the providers are not unnecessarily irritated.

### 3.4. Machine Learning Pipeline

The machine learning pipeline is basically the one that implements a data analysis method in the proposed methodology. Feature engineering is to transform the raw data of claims and providers into the variables that are significant and indicate the behaviors, financial risk, coding complexity, and historical patterns. Time-related features, for example, can show provider billing speed, seasonal changes, and recurring anomalies. Peer comparison features can point to the deviations from the normal patterns within the same specialty or region. Historical overpayment markers, anomaly indicators, and cost metrics also help to deepen the feature set. All these engineered features are

updated in real time through streaming pipelines so that they can be as current, accurate, and fully compatible with real-time adjudication as possible.

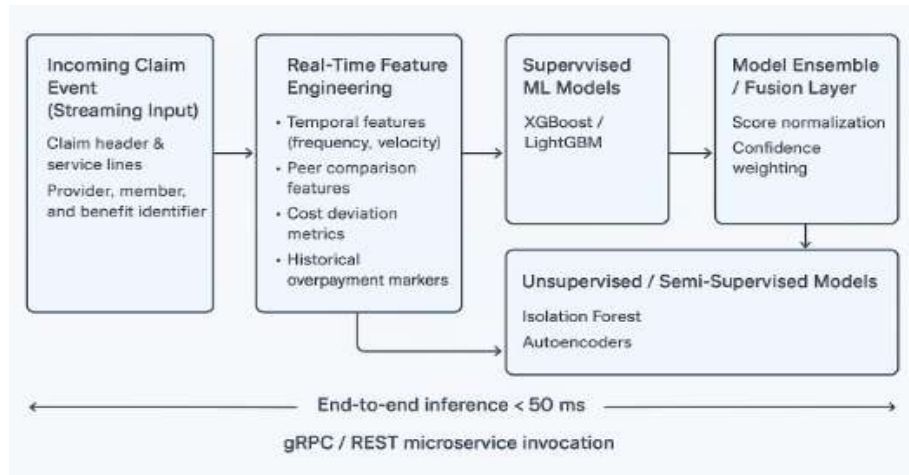


**Fig 1 - Machine Learning Pipeline for Real-Time Risk Prediction**

Modeling is supported by top-notch algorithms like gradient boosting (XGBoost, LightGBM) and ensemble structures that mix supervised learning with unsupervised anomaly detection. This mix of techniques gives the system the ability to spot not only the risk patterns that are known but also those that have never been seen before. The ML models are active through a low-latency inference layer that is set up to give predictions in less than 50 milliseconds, thus allowing the injection to be done directly in the critical adjudication path. Monitoring of canary releases, model versioning, and drift ensures that there is always improvement and stability over the changing claim landscapes. The outcome is a dynamic, high-precision ML pipeline that adjusts to the changing provider behaviors and coding trends.

### 3.5. Risk Scoring Framework

A risk scoring engine uses the rules and ML outputs to combine all analytical signals into one unified score, which can also be understood. To arrive at the risk profile of a claim, the engine considers various factors such as the severity of the rule, the ML-predicted overpayment probability, anomaly scores, confidence intervals, and even historical error patterns by integrating them in a weighted manner. This single score equips the system with the ability to act differently under various risk situations, it can thus determine the presence of a risk through straightforward logical deductions or in a probabilistic way through predictions and also take into account the operational priorities. The business rules and compliance requirements are still at the core of the decision-making logic as ensured by the weighted design of the system while it is still possible for ML insights to a significant extent influence the outcomes.

This composite risk score is what the system uses to make the final routing decision. Generally, low-risk claims are approved automatically, without the need for a human review, thus the payer is

able to keep the throughput at a high level and shorten the waiting times. In the case of medium-risk claims, they may be put on hold for additional documentation, or sent to clinical reviewers or financial analysts. The investigative unit of special cases can receive the high-risk claims, precisely those revealing the most unusual or suspicious patterns, for a thorough examination. In this way, the organization of resources is done in an efficient manner and the effort assigned to review is proportional to the expected financial impact.

### 3.6. Integration with Adjudication Systems

For the methodology to be successful on an operational level, it should be able to work hand in hand with the existing adjudication systems without any problems. The system facilitates communication of both synchronous and asynchronous types, thus adjudication engines are able to make real-time calls to the ML and rules engine without affecting the flow of work. Callouts based on API provide the payers platforms with compatibility and mutual communication without any restrictions, on the other hand, containerized microservices offer the system resilience and scalability. At the same time, high availability configurations aim at uninterrupted operation even during times of system upgrades or peak claim volumes. This integration design allows for a real-time architecture that can co-exist with old systems while at the same time deliver modern analytical capabilities.

Features of explainability are among the components of the integration layer that aims at ensuring transparency and compliance. The outputs comprise model explanations based on SHAP or LIME, rule-triggered reasons, as well as detailed audit logs covering every step of the scoring process. Such a degree of traceability facilitates the organization's readiness for external audits and provider appeals, thus lessening the risk of compliance and aiding the implementation of regulatory expectations. In the end, the integration framework serves as a connection between modern analytics and established adjudication workflows which is technically strong and operationally viable.

## 4. CASE STUDY

### 4.1. Background

#### 4.1.1. Implementation within a Payer Environment

The payer featured in this case study is a national health plan that manages the care of about 20 million members through different insurance products such as commercial, Medicare Advantage, and Medicaid. The organization with over 60 million claims a year was in a situation of great complexity to ensure that the adjudication was done in a timely and accurate manner. The prior works mainly depended on rule-based systems and post-payment audits to locate improper payments which limited the capacity to intervene early and caused inconsistent recovery results. Moreover, these retrospective methods led to operational inefficiencies and provider disruptions, who were often required to respond to documentation requests weeks or even months after services had been delivered.

A detailed examination of claim trends has shown that ambulatory surgical procedures and the categories of durable medical equipment (DME) have been the main contributors to the excessive

payment of the healthcare program. These segments have shown significant variations in the coding, a high frequency of upcoding, and an increased exposure to billing anomalies. The leadership saw the necessity of having a system for the detection of fraud in real-time which could thus enable the intervention at the point of payment. The large volumes of data, various types of claims, and the urgent need for modernized integrity controls, which together described the environment, were the right conditions for the introduction of an AI-driven, real-time prevention solution.

## 4.2. Deployment Architecture

To support real-time detection, the payer adopted a modern architecture centered around cloud-native processing, low-latency components, and scalable microservices. Claim data flowed through Kafka, which served as the backbone of the ingestion pipeline, enabling parallel processing and reliable message distribution. This event-driven structure allowed the system to ingest claim headers and lines in near real-time while maintaining strict ordering for claims with multiple service lines. The streaming capability also provided replay and audit functionality, which became valuable during model testing and regulatory reviews.

The analytical components were deployed on AWS or GCP, where the machine learning pipeline handled training, versioning, drift detection, and automated redeployment. Redis played a key role as a high-speed feature store, providing rapid lookup of provider patterns, cost histories, prior claim outcomes, and contextual variables required for ML inference. The inference layer itself consisted of containerized microservices orchestrated through Kubernetes, ensuring elasticity and high availability. This architecture allowed risk scoring to occur within milliseconds, making it feasible to integrate directly with the payer's adjudication systems.

## 4.3. Operational Workflow

The flow of the operation was such that the adjudication engine of the payer received a claim, after which a synchronous call to the real-time detection pipeline was made. The claim event was grabbed on the fly by Kafka from the streaming and was then normalized and enriched with the help of the additional data that was sourced from provider repositories, historical claims, and benefit configurations. While the claim was on its way through the workflow, the rule engine was running a set of deterministic checks that were related to coding consistency, bundling logic, and clinical appropriateness. The outputs of these rules were the main risk factors that formed the fundamental layer for the scoring process that was more extensive.

At the same time, the ML inference service was using both real-time features and long-term provider behavior patterns to assess the risk of the claim. The resultant model score and anomaly indicators were combined with rule outputs to create a composite risk decision. This decision was, as a rule, the first one to be returned to the adjudication engine and it was done within 150 milliseconds. Claims that had an elevated risk were automatically pending for the investigation, whereas low-risk claims were allowed to go through adjudication without any interruption. This dual-track, low-latency

workflow was the way in which advanced analytics could be at work alongside the legacy systems that had not been upgraded, without the throughput being slowed or provider payments being disrupted.

#### 4.4. Key Performance Indicators

The payer tracked success through an extensive range of KPIs that not only accounted for the financial impact but also included model performance, operational efficiency, and provider experience. Model metrics such as AUC were one of the indicators of very high predictive accuracy which was a source of confidence for risk scoring decisions. Latency numbers showed that the system was able to meet the payer's sub-200 ms adjudication requirement most of the time, including periods of peak traffic. Furthermore, the platform's continuous high throughput capacity was a strong indicator of the scalability of the architecture and its viability for large enterprise-level claim volumes.

Additionally, to the model and system performance, the operational KPIs showed significant improvements. The payer noticed a significant decrease in the overpayment frequency, which served as a confirmation to the effect of intercepting improper claims pre-payment. The provider relations also got better as the number of post-payment audits and documentation requests decreased, thus, a more comfortable reimbursement process was experienced. The reviewer staff also had less work as the reduction of unnecessary spends lowered their workload and gave them the opportunity to engage with high-impact cases. In sum, these KPIs exemplified the extensive benefits of the implementation of an integrated real-time payment integrity solution.

## 5. RESULTS AND DISCUSSION

### 5.1. Quantitative Results

Qualitatively the changes were also big - the implementation delivered huge improvements quantitatively across the financial, operational, and technical dimensions. The 60% decrease in overpayments was the most significant effect, which accounted for the majority of the system's ability to single out the claims with the highest risk even before issuing the payment. Hence, the change from retrospective recovery to proactive prevention lowered financial leakage significantly and, therefore, loss ratios stabilized.

Moreover, the number of cases for manual review has been dramatically lowered as well, by 35%, thus, there are fewer instances of human evaluators receiving the cases. This adjustment has allowed analysts to engage in complex, lucrative investigations rather than spending their time on routine verifications. Technically speaking, the machine learning model has reached an AUC of 0.89, thus it can be considered a very accurate risk stratification tool for claim types of different natures. Also, the real-time pipeline was always working within 120–150 ms intervals even during the times when the loads were over 10,000 claims per second, thus the architecture is scalable and ready for production.

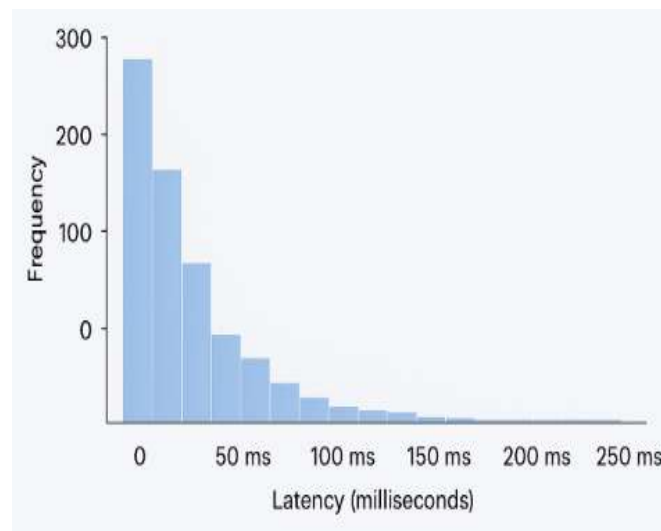


**Fig 2 - Latency Distribution under Peak Load**

## 5.2. Qualitative Findings

Qualitative feedback from internal stakeholders and external providers was consistent in recognizing additional benefits. Providers communicated that they experienced fewer post-payment audits as well as fewer documentation requests, thereby contributing to a more comfortable reimbursement process and enhanced trust in the payer's review system. The lessening of audit abrasion therefore contributed to the strengthening of provider relationships and their alignment with compliance.

Business teams were more confident about the honesty of the payment choices that were made. The use of both rule-based and machine learning-based insights gave more clarity to the risk drivers and lessened the uncertainty that came from manual adjudication. The very first implementation indeed showed a time of high false positives, especially for providers having unconventional but valid billing patterns. It was necessary to keep tuning, iteratively refining features, and having reviewer feedback loops to be able to stabilize the system. This understanding was a confirmation that it was still very important to balance speed, accuracy, and fairness in real-time decisioning.

However, the innovations also unveiled a number of significant issues. Deliberately wiring up the new real-time scoring components with the old adjudication systems necessitated the construction of bespoke interfaces and the exact coordination to keep the adjudication timelines consistent. The challenge of managing model drift turned into a continuous task as the evolution of provider behavior, coding trends, and clinical practices meant that the model had to be retrained and recalibrated from time to time.

Besides that, data quality problems have taken ground, for example, incomplete provider profiles, inconsistent code submissions, and missing historical attributes. These holes sometimes influenced scoring reliability and thus called for strong confirmation, fallback logic, and elevated data

governance. Moreover, the payer was obliged to ensure absolute compliance with the regulations that demanded clear explainability, audit traceability, and defensible rationale for each automated decision.

### 5.3. Comparative Analysis

Benchmarking real-time detection against the traditional retrospective audits made the strategic benefits of the former quite evident. The real-time detection stopped the improper payments from going forward and thus decreased the operational cost approximately by 40–50% as a result of less recovery attempts, appeal processes, and documentation reviews. In most cases, retrospective audits can only recover a fraction of improper payments, whereas real-time controls avert the leakage altogether.

Furthermore, the elimination of post-payment interventions contributed to better provider relations. Providers had fewer retroactive adjustments, fewer disputes, and more trust in the way their claims were evaluated. The change enhanced the payer's operational efficiency and gave the payer a better standing among the network providers. In the end, real-time payment integrity was better not only in terms of financial performance but also in terms of the long-term impact on the ecosystem.

## 6. CONCLUSION AND FUTURE SCOPE

Such real-time overpayment detection at the very instant of the payment is not only technically plausible but is also becoming strategically indispensable for the healthcare payers that are leading the way. By using the advancements in data streaming, low-latency analytics, and machine learning, organizations have the power to turn their payment integrity programs that were once simply reactive, audit-heavy entities into proactive systems that stop financial leakage. The combined architecture – layering rules, ML, anomaly detection, and explainability – shows that high precision and sub-second decisioning can exist together in the current adjudication environments.

Their findings reveal that these architectures yield significant tangible value. In other words, payers have the capacity to bring down overpayments substantially, Manual review workloads can also be lowered, and compliance outcomes improved, while, at the same time, provider trust is further strengthened by way of post-payment disputes and documentation requests minimization. Besides this, real-time prevention is able to give financial predictability and reduce institutional overhead, thus presenting a far better cost-benefit profile than the conventional retrospective audit models. The capacity demonstrated for scalability to millions of claims per day, in turn, confirms the feasibility of real-time intervention as a core capability of an enterprise.

Anyway, the idea of payment integrity in real-time is more of an unfolding horizon rather than a finish line. On one side, there is federated learning, generative AI, predictive provider modeling, blockchain-based claim provenance, and autonomous adjudication systems that are waiting to be harnessed to upscale precision, transparency, and automation in payment integrity. On the other hand, as payers keep on modernizing their infrastructure and governance frameworks, the industry is getting

closer and closer to the future where intelligent, adaptive, and highly automated payment integrity will be the rule, not the exception due to the integration of these innovations.

## REFERENCES

- [1] Kodete, Chandra Shikhi. "A Real Time AI System for Automated Financial Technology Payment Detection and Risk Reduction." *International Journal of Scientific Research in Computer Science Engineering and Information Technology* 7 (2021): 685-710.
- [2] Li, J., D. Aleman, and R. Sikora. "A DATA MINING FRAMEWORK FOR IDENTIFYING CLAIM OVERPAYMENTS FOR THE HEALTH INSURANCE INDUSTRY."
- [3] Rajapaksha, Charith Isuru. "Machine Learning-Driven Anomaly Detection Models for Cloud-Hosted E-Payment Infrastructures." *Journal of Computational Intelligence for Hybrid Cloud and Edge Computing Networks* 6.12 (2022): 1-11.
- [4] Liu, Chang. *Development of prediction model for real-time parking availability for on-street paid parking*. Diss. University of Pittsburgh, 2016.
- [5] Cai, Bill Yang, et al. "Deep learning-based video system for accurate and real-time parking measurement." *IEEE Internet of Things Journal* 6.5 (2019): 7693-7701.
- [6] Young, Matthew M., et al. "Using artificial intelligence to identify administrative errors in unemployment insurance." *Government information quarterly* 39.4 (2022): 101758.
- [7] Abbey, Duane C., ed. *Healthcare payment systems: an introduction*. CRC Press, 2009.
- [8] Kenyon, Robert, et al. "Internet initial claims evaluation." (2003).
- [9] Hoseini, Cyrus. *Leveraging machine learning to identify quality issues in the Medicaid claim adjudication process*. Indiana State University, 2020.
- [10] Matthews, Michelle. *CONSTRUCTION PAYMENT AUTOMATION THROUGH A BIM-BLOCKCHAIN (BIM-Bc) AND SMART CONTRACT FRAMEWORK*. Diss. Faculty of Engineering and Built Environment, University of the Witwatersrand, 2021.
- [11] Buoni, Alessandro. *Fraud detection in the banking sector: a multi-agent approach*. Diss. Turku Centre for Computer Science (TUCS), 2012.
- [12] Conforti, Raffaele. *Managing risk in process-aware information systems*. Diss. Queensland University of Technology, 2014.
- [13] Cole, Curtis L., and Adam Cheriff. "Ambulatory Systems." *Medical Informatics: An Executive Primer* (2015): 115.
- [14] Fowler Jr, Donald L., and Elizabeth Pascuzzi. "Claims Administration." *Essentials of Managed Health Care* (2007): 387-442.
- [15] Parziale, Lydia, et al. *Enabling Real-time Analytics on IBM z Systems Platform*. IBM Redbooks, 2016.