ARTICLE 10

Analysing Expenditure Vouchers of Indian Railways using AI/ML for Anomaly and Fraud Detection

V Gautham Kumar*

Received : 01 June 2025 **Accepted : 12 November 2025**

Abstract

The article highlights the scope for analysing the railway expenditure using Natural Language Processing (NLP) and Artificial intelligence (AI) reasoning on bill descriptions of vouchers to identify anomalies such as misclassification, duplicate payments, non-compliance and possible fraud. The method is used to convert the bill description into an embedding to capture its semantic meaning and then analyse the allocations to understand the common patterns and deviations. When the training dataset is labelled and a model is developed, the model should be able to identify the anomalies in test data. Using the reasoning of the GPT models [1], the system should identify common patterns in allocations associated with bill descriptions and be able to detect anomalies even if the data is not labelled. Use of AI/Machine Learning (ML) in auditing has potential for a full population voucher audit and early anomaly detection in railway expenditure.

Keywords

Artificial Intelligence, Machine Learning, Natural Language Processing, Voucher Audit, Anomaly Detection, Audit Assurance and Indian Railways.

10.1 Introduction

Railway Audit Wing of CAG Office is responsible for ensuring financial propriety in the functioning of the Indian Railways (IR). The Integrated Payroll and Accounting System (IPAS) of IR, developed by CRIS [2] processes various types of bills, both for employees and contractors and maintains expenses for railways. The primary output of IPAS is the generation of accounting vouchers, which form the basis of railway financial records and statements.

*Assistant Audit Office, O/o Principal Director of Audit, South Central Railway.
Email: villurig.scrly@cag.gov.in

[1]GPT is Generative Pre-Trained Transformer, is a family of AI models built by OpenAI. GPT gives AI applications the ability to interpret data and reason.

[2] Center for Railway Information Systems is an organisation under Ministry of Railways for development and maintenance of Railway IT systems.

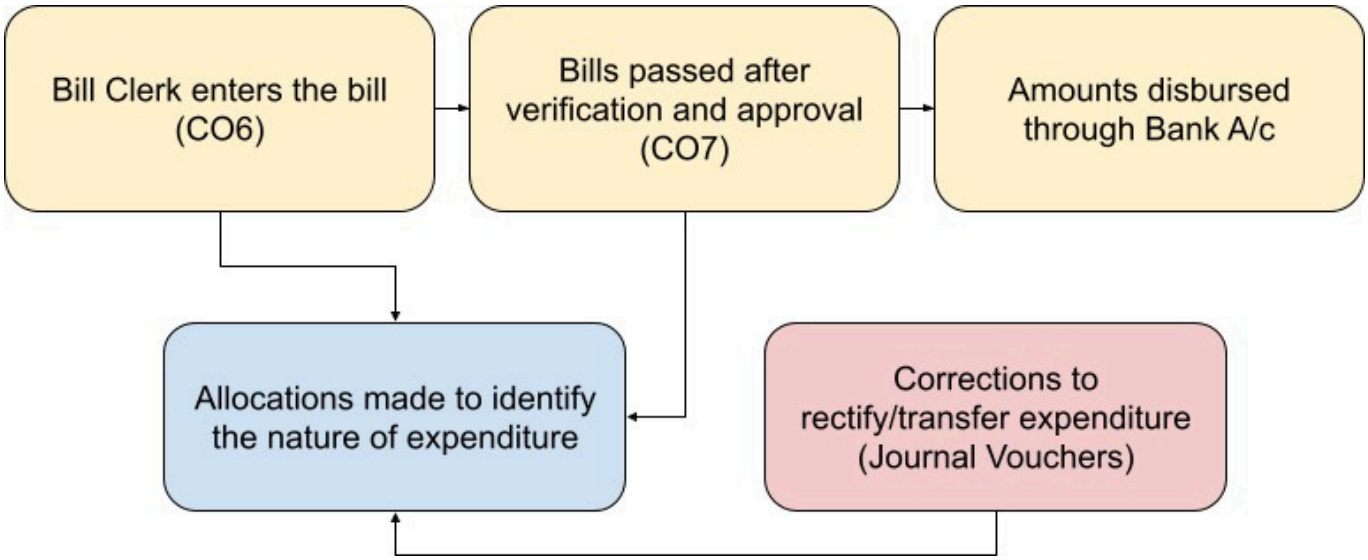
The Railway Audit Wing of CAG has the mandate of auditing the railway expenditure as per the norms and extent of checks prescribed by SMI [3] for different kinds of vouchers such as Contractor Bills, Supplier Bills, Journal Vouchers, etc., The vouchers generated in a month are extracted from IPAS by respective sections of railway audit offices, for selection (judgmental or random) and audit.

10.2 Bill Passing

CO6 [4] vouchers are bills entered in the IPAS application by railway executives' departments. for verification and passing by the Accounts department. These vouchers contain data such as party name and code, bill description, amount, deductions, bank account number and others. Along with this, each passed CO6 voucher also has an allocation code for all affected heads of accounts. The allocation code uniquely identifies the nature and object of expenditure and forms an important identifier to detect misclassification and mistakes in railway accounting. The Accounts department of the Railways, upon verification and subject to the availability of funds, passes the bills for payment through banks (Figure 10.1).

The Journal Vouchers contain information affecting the heads of accounts, for rectification/modification, along with the narration describing the reasons thereof. Journal vouchers are also used for the adjustment of expenditure between divisions in a railway zone or between different railway zones.

Figure 10.1: Flow of bill passing in IPAS



10.3 Data Volume

In the South Central Railway Zone, the number of CO6 and JV entries in IPAS in a financial year is approximately around 2.2 lakh, and the corresponding allocation entries are around 10 - 12.5 lakh. The volume of voucher data being generated in the IPAS application, if effectively put to use in audit, will emerge as a cornerstone in enhancing the impact of audit findings.

[3] Secret Memorandum of Instructions
[4] Cash Order and the number 6 indicates the stage of bill passing.

The Bill Description and Narration fields in CO6 and JV tables are natural language text entries made by the bill entering clerks and are important in understanding the details of expenditure vis-à-vis the primary objective (allocation) of such expenditure.

10.4 Integrated Applications in Indian Railways

IPAS is also integrated with some other railway IT applications, like IRWCMS (Indian Railway Works Contracts Management System), CMS (Crew Management System), etc. Integration provides additional information to the bill passing authorities, like agreement details, security deposit, and performance guarantee to be collected, special conditions, etc., through IRWCMS. The CMS application, which records crew login and logout times, provides information about overtime hours for the calculation of overtime allowance during payroll generation.

10.5 Traditional Audit

Traditional audit practices, though fundamental, face significant limitations in data-intensive environments, reducing the overall efficiency and coverage of the audit process. Audit that is rule-based and manual with limited sample size and subjective in voucher selection, etc., is both time-consuming and voucher dependent, and can pose a serious risk that irregularities in unselected vouchers remain undetected. These methods also cannot handle the large volumes of data being generated in IPAS daily.

10.6 Fragmented Approach

For hundreds of bills passed in IPAS during a month, the sample selection reduces the scope of audit to a few individual vouchers. Audit of these individual voucher(s), if not audited holistically, presents a fragmented view and is hard to consolidate across divisions, zones or financial years. The potential audit risk in the sample of vouchers selected can be identified only if the auditors are experienced and have vast domain experience. For better understanding, consider the following scenarios:

I - In Works Contracts, the contractor submits the bills as and when a percentage of the work is completed. These bills are called Contract On-Account bills. When a particular on-account bill is selected for audit, the bill is audited in isolation for compliance, collection of taxes and levies, etc.

The auditor is deprived of the information about all the previously passed bills (audited or not audited) for the same work to check if the current bill is duplicated or overpayment is being made or if the bill contains the correct bank details or if the supplies mentioned are actual or not.

In South Central Railway, a fraud of ₹ 2.2 crore was found where fake medical bills were generated and passed in IPAS without there being any real supply of medicines.

II – Price Variation Bills (under Price Variation Clause- PVC) are passed against contractor on-account bills to adjust for inflation for the cost of raw materials in long gestation projects (works contracts). When a PVC bill is selected for audit, it is also required to check all the on-account bills on which the PVC claim is being made. This is time consuming and resource-oriented and sometimes the old on-account bills may not be available. Lack of complete information can result in faulty audit assurance in the audit of such bills.

A fraud of ₹ 6.33 crore was noticed in North Frontier Railway while reviewing the amount of PVC bills; there was tampering of the figures in on-account bill. The inflated on-account bills resulted in excess payment.

Apart from these, understanding the semantics of the bill descriptions/narrations of vouchers, which is critical to audit, is resource oriented and a constraint in traditional audit.

10.7 Role of Semantics in Audit

Audit tools like IDEA aid auditors greatly in identifying duplicate records, trends and patterns, data analytics, etc. but does not offer the flexibility to develop custom applications specific to domain data. Also, traditional search enabled by these tools can analyse and identify anomalies in expenditure, only if specific keywords are known to audit. These are fast and accurate but fail to understand the semantic meaning, when different words are used in bills or vouchers.

10.7.1 Understanding Semantics of Voucher Descriptions

Semantics refers to the analysis of word meanings, understand relations between them and go beyond treating the words as just keywords.

Descriptions available in vouchers (bills) are primary to identify the nature of the expenditure. These are natural language text entered by the bill clerks, and similar expenditures may have different ‘words’ used to mean the same. From the data extracted from IPAS, expenditure incurred due to a decree passed by courts, which is of the nature of charged expenditure commonly contained the words as given in Table 10.1.

Table 10.1 Words used to describe expenditure incurred on court decree found in IPAS
· arbitration award · OA (Original Appeal) · LAC (Land acquisition case) · decree award · court award · compensation claim

Semantic search performed on any word from the above table, would give search results matching all other words with inherent same meaning available in the dataset. A similar output from traditional search would require auditors to identify all the available keywords which is laborious and time consuming.

Analysing the inherent meaning (semantic) of textual descriptions available in vouchers is critical in understanding the nature of expenditure and help to detect possible anomalies like misclassifications in expenditure, duplicate bill passing, fraud detection, etc.

10.7.2 Natural Language Processing

Natural Language Processing (NLP) is a field of Artificial Intelligence that uses machine learning concepts to make computers read and understand the human language. NLP can perform a wide range of tasks – text classification, pattern recognition, sentiment analysis, etc.

NLP and semantics in audit (voucher descriptions) for pattern recognition can enable audit to process large volumes of data and provide meaningful insights for anomalies, risk assessment and data-driven audit assurance. NLP also has the ability to learn from the continuously generated data to understand, analyse and reveal any hidden patterns in ‘bill description’, ‘amount’, ‘allocation codes’ for improving audit efficiency and effectiveness.

Models trained on historical data can also monitor the expenditure real time, by identifying any deviation from the learned patterns and thus help auditors. The algorithms not only learn from the primary data – bill descriptions, allocation codes but from also other data available like IP address, device information, time of transactions etc.

10.7.3 System Development

A prototype unsupervised (these algorithms discover patterns without the need for human intervention) AI/ML system is under development with the aim to help auditors in:

- detecting misclassification in allocation of expenditure using semantic similarity;
- detecting duplicate expenditure or bill payments; and
- detecting outliers in expenditure for possible fraud.

10.7.4 Data processing

The expenditure of railways is primarily available in three tables (CO6, JV and Allocation). The CO6 table records the expenditure, which contains the bill amount and bill description, among other information. The classification (object of expenditure) is recorded in the Allocation table which contains information such as Debit/Credit, Voted/Charged and the Allocation code (Structure in Table 11.2). The JV table is used to adjust the expenditure made.

When a bill is passed, it is presumed that all allocation codes are entered correctly by the railway executive, and as such, the data is not labelled.

Data available from IPAS is processed to create a combined dataset containing 'description' and 'allocation code' from two different datasets (CO6 table and JV table). Other metadata like party name, transaction amount, and whether the expenditure is voted or charged, is also added to the dataset.

10.7.5 Create embeddings

The semantic meaning of each bill description, narration, allocation, amount, etc., of CO6 and JV tables is captured using dense vector embeddings using Hugging Face Transformer model. Dense vectors are mathematical objects that represent data in AI/ML. The dense vectors can describe subtle relations and nuances that exist between data and help identify semantic similarities. For the financial years 2023-24 and 2024-25, dense vector embeddings are created for around 2.5 million CO6/JV records.

Table 10.2 Structure for bill passing in IPAS system of Indian Railways

CO6 Table

CO6 Number	CO6 Date	Bill Type	Bill Amount	Party Name	Bill Description
09010125002156	04-08-2025	GEM Bill	62517	Santosh Kumar	Hiring of vehicle
09010125002157	04-08-2025	Pay Order	9675	Florist Pvt Ltd	Purchase of flower bouquets for official functions
09010125002158	04-08-2025	Vehicle Bills	52500	Infra Developers	Pooled vehicle for SDGM Office
09010125002159	04-08-2025	Vehicle Bills	51922	K Nagaraj	hiring of vehicle dv cvo ST

JV Table

JV Number	JV Date	JV Type	Year Month	TC No	Narration
090132416032	05-06-2025	Capital	202506	...	TPI Bill Payment of SCR for the month of OCT-2024 having amount 1196.2 for 0901
090132416065	07-01-2025	Revenue	202501	...	Rectification of work ID, Expenditure booked under PH-1700, CO7 09010424700010

Foreign Key relation from CO6 Table and JV Table to Reference Number of Allocation Table

Allocation Table

Reference Number	Voted/Charged	Debit/Credit	Amount	Allocation	Allocation Description
09010125002156	V	D	62517	92081000	Refund of revenue collected
09010125002156	V	C	62517	00867002	Cheque Paid
09010125002158	V	C	50341	00867002	Cheque Paid
09010125002158	V	C	159	93065200	Other Sundry Receipt
09010125002158	V	C	1000	00002102	Income Tax
09010125002158	V	C	500	00844541	TDS GST
09010125002158	V	C	500	00844542	TDS GST
09010125002158	V	D	52500	00870906	Transfer to Division

10.7.6 Index in Elasticsearch

The CO6, JV and Allocation table data and dense vectors are then stored in Elasticsearch along with other required metadata. Elasticsearch is preferred for its advantage of semantic and vector similarity search, and its "more like this" (MLT) query feature.

10.7.7 Retrieval System

LangChain [5] was integrated into the model as a semantic retrieval layer on Elasticsearch to perform fast semantic queries over millions of vectorised transactions. Langchain facilitates the integration of Large Language Models (LLMs) with external data source and help build applications that respond to user queries based on context.

It provides tools and structures to build and implement RAG pipelines. Retrieval Augmented Generation(RAG) is a technique to improve the LLM responses by extracting information from a custom or domain-specific knowledge base. Such results obtained from RAG will be more relevant, context-aware and accurate.

From the Langchain retriever, the system uses semantic similarity queries to detect inconsistencies in the data, as detailed:

10.7.7.1 Duplicates: If a near identical match exists with the same bill description, or narration, amount and allocation code, a duplicate entry will be flagged to provide insight to the auditor for further examination.

10.7.7.2 Outlier detection: If no similar record exists within a given description, narration, bill amount and allocation code, a possible unusual entry or outlier is flagged by the system for further audit.

10.7.7.3 Misclassification detection: For a given bill description, narration and allocation code, if most similar records belong to another allocation code, the system will flag a possible misclassification.

These checks are performed using the similarity_search_with_score method available in Langchain, and the available data is batch scanned through periodic job scheduling using Celery or as and when new expenditure data of the Railway audit is made available.

10.7.7.4 Reasoning: After obtaining the top n records from the retrieval system, reasoning is done using a GPT model like GPT-4, GPT-4 of OpenAI to explain why a transaction could be an anomaly by summarising similar records and highlighting deviations and to provide justification to the auditors.

[5] LangChain is a popular framework for working with AI, Vectors, and embeddings. It is used to simplify building a variety of AI applications. Elasticsearch can be used with LangChain in three ways: to store and retrieve documents from Elasticsearch, with the help of an LLM like OpenAI, to transform a user's query into a query + filter to retrieve relevant documents from Elasticsearch and for the most flexible way to retrieve documents from Elasticsearch (Source: https://www.elastic.co/search-labs/integrations/langchain, accessed 28 September 2025).

10.8 Conclusion

As data generated in IPAS continues to grow and become more complex, traditional audit practices show limitations in effective audit assurance. AI/ML offers a more effective tool to aid auditors in processing large and interconnected datasets, identifying patterns and detecting anomalies.

Data Availability

There is no new data associated with this article.

Ethics Statement

This research, idea and concept met the ethical guidelines and legal requirements of the country in which it was performed.

Funding

The idea and concept is the original work of the author and there was no funding of any kind or form.

Conflict of Interest

None declared

Acknowledgement

None declared

References

1. Center for Railway Information Systems. (n.d.). Ministry of Railways.
2. Comptroller and Auditor General of India. (n.d.). New Delhi.
3. Caseware. (n.d.). IDEA Data Analysis Software (Version X.X). Caseware International Inc.
4. Elastic. (n.d.). Elasticsearch. Retrieved from https://www.elastic.co/elasticsearch
5. Hugging Face. (n.d.). Home. Retrieved from https://huggingface.co/
6. Hugging Face. (2023, July 18). Getting started with embeddings. Hugging Face Blog. https://huggingface.co/blog/getting-started-with-embeddings1
7. IBM. (n.d.). Natural language processing. Retrieved from https://www.ibm.com/think/topics/natural-language-processing
8. IBM. (n.d.). Unsupervised learning. Retrieved from https://www.ibm.com/think/topics/unsupervised-learning
9. Indian Railways. (2019). RBA 85/2019: Compendium. Research Designs and Standards Organisation (RDSO). https://rdso.indianrailways.gov.in/railwayboard/uploads/directorate/accounts/downloads/Compendium_2019/RBA85_2019_23092019.pdf

10. LangChain. (n.d.). Home. Retrieved from https://www.langchain.com/

11. LangChain. (n.d.). RAG tutorials. Python documentation. Retrieved from https://python.langchain.com/docs/tutorials/rag/

12. Times of India. (2019, July 10). CBI: Rly officials embezzled over Rs 7 cr in med bill fraud. Times of India, Hyderabad. https://timesofindia.indiatimes.com/city/hyderabad/cbi-rly-officials-embezzled-over-rs-7-cr-in-med-bill-fraud/articleshow/70149977.cms

---

Kumar, V. G. (2025). Analysing expenditure vouchers of Indian railways using AI/ML for anomaly and fraud detection. Manthan: Journal of Public Sector Auditing and Accounting of SAI India, 1(1), 74–83.