Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 07, July-2024

Automating Financial Audits using Data Pipelines and AI: A Conceptual Hybrid Approach

Kishor Yadav Kommanaboina
Independent Researcher
The Ohio State Univ Alum

Tejaswini Kumar
Independent Researcher
Columbia Univ Alum

Bhargava Kumar
Independent Researcher
Columbia Univ Alum

***Abstract*— This paper proposed the use of data pipelines and AI in financial audit automation based on integrating both quantitative data with qualitative audit insights. In the era of big data, financial audits have grown increasingly complex, and conventional approaches of handling them cause difficulties. In this research, we describe a novel hybrid conceptual approach that uses robust data pipelines and advanced AI models to combine quantitative financial data with qualitative insights from auditors. The methodology consists of extracting quantitative data from financial records, collecting qualitative data through surveys and interviews, and integrating both data types utilizing data transformation and harmonization methods. After then, AI models are created and trained on this combined information to automate the entire auditing process, from data entry to the creation of the final report. Significant gains in audit efficiency, accuracy, and data quality are anticipated with this strategy. The paper addresses the practical ramifications for auditors and makes recommendations for future lines of research to enhance and validate the hybrid method, which could revolutionize the field of financial auditing.**

***Keywords*— Financial Audit workflow, Data pipelines, Artificial Intelligence (AI), Automated audits, Data Harmonization**

I. INTRODUCTION

The advent of big data and cutting-edge technologies has drastically changed the financial auditing landscape. Traditional auditing systems are finding it more and more challenging to manage the massive amounts of data and complexity associated with modern financial audits because of their reliance on human judgment and manual processes. Combining artificial intelligence (AI) with data pipelines provides a workable solution to these issues, enhancing the efficiency, accuracy, and dependability of financial audits.

The potential of AI in auditing has been the subject of numerous studies, which have highlighted the ability of the technology to automate repetitive processes, spot irregularities, and offer more in-depth analysis of financial data. Wang et al. [1] discussed the inadequacy of conventional auditing techniques for managing enormous datasets, who argue that sophisticated technologies are necessary. While AI can improve audit speed and accuracy, Nizamdinova et al. [2] point out that there are drawbacks as well as potential advantages to integrating AI into the auditing process. With an emphasis on the advancements in anomaly identification and risk assessment, Ivakhnenkov [3] explores how AI is changing the auditing landscape. Fredrick et al. [4] investigate the efficiencies recognized through Intelligent Process Automation (IPA) in auditing, demonstrating how automation can reduce manual effort and errors. Munappy et al.'s study [5] highlights the value and difficulties of managing an organization's data pipeline while emphasizing the necessity of data integration and quality for AI systems to function efficiently. In their discussion of how auditing is affected by the quick advancement of technology, Solanki et al. [6] emphasize the necessity of ongoing technological integration and adaptation. The use of AI in gathering and evaluating financial data is examined by Nunes et al. [7], who highlight how AI has the potential to revolutionize the auditing process.

Previous studies have predominantly concentrated on two parts of auditing: the quantitative side, which involves data pipelines and artificial intelligence, or the qualitative side, which includes human judgment and knowledge. Studies that successfully integrate these two strategies to take advantage of their complementary strengths are scarce, though. This gap offers a great chance to create a comprehensive framework that combines quantitative data with qualitative insights from auditors, processed using reliable data pipelines and cutting-edge AI models.

In this paper, we propose a conceptual hybrid approach that combines qualitative audit insights with quantitative data to automate financial audits using data pipelines and AI. This approach aims to address the existing literature gaps and enhance the overall efficiency, accuracy, and reliability of financial audits.

II. METHODOLOGY

This section illustrates our conceptual hybrid approach to automating financial audits with data pipelines and artificial intelligence. It also includes the technique for data collection and analysis. Workflow automation, AI model building, integration, and data collecting are all part of the technique. We ensure our method will improve the auditing process by utilizing both quantitative financial data and qualitative insights from auditors.

A. Data Collection

1) Quantitative Data Collection:

a) Interviews and Focus groups:

- Participants: In-depth interviews and focus groups with managers, other important stakeholders, and seasoned financial auditors are proposed. The participants would be chosen according to their depth of knowledge and proficiency in financial audits.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 07, July-2024

- Questions: Open-ended questions would be utilized for directing the focus groups and interviews in order to gather in-depth information on the auditing process, its difficulties, and the areas that might profit from automation. Subjects covered would be typical problems, the value of human judgment in audits, and opinions on AI integration.

b) Surveys:

- Design: Surveys are designed for collecting qualitative information from a wider spectrum of auditors. To gather both quantitative data and qualitative insights, the surveys will include a variety of closed-ended and open-ended questions.
- Distribution: Surveys would be distributed through professional auditing networks and organizations to ensure a wide range of responses.

1) Quantitative Data Collection:

c) Fixed Data Sources:

- Financial Records: Balance sheets, income statements, cash flow statements, and other historical financial documents from corporate databases would be studied. These documents offer a dependable and constant supply of quantitative data for training models..
- Transaction Logs: Detailed transaction logs from accounting software would be utilized to capture day-to-day financial activities. These logs are essential for identifying patterns and anomalies in financial data.

d) Dynamic data Sources:

- Market Data: Real-time providers of financial data would be the source for financial market data, including economic indicators and stock prices. To reflect current market conditions, this data require update on a regular basis.
- Regulatory Updates: Regulatory databases would be used to monitor changes to financial regulations and compliance requirements. This guarantees that the audit procedures and AI models continue to adhere to the most recent criteria.

B. Data Integration

e) Data Harmonization:

- Consistency: Harmonizing qualitative and quantitative data would be the first step in the process to guarantee terminology and format uniformity. This entails synchronizing data from various sources, coding qualitative replies, and standardizing financial indicators.
- Transformation: Using coding techniques, qualitative data from surveys and interviews would be converted into organized formats to enable efficient integration with quantitative financial data.

a) Integration tools:

- ETL Tools: The process of automating data integration would involve the use of Extract, Transform, and Load (ETL) tools, which enable the extraction of data from several sources, transform it into a uniform format, and load it into a single repository.

Data Integration Platforms: For managing the combined datasets, advanced data integration tools would be employed. The real-time data processing capabilities of these platforms guarantee the constant updating of dynamic data sources.

B. AI Model Development

2) Model Design and Training

f) Architecture: An AI model architecture would be designed to manage both qualitative and quantitative data. For qualitative data, Natural Language Processing (NLP) models like Bidirectional Encoder Representations from Transformers BERT [8] would extract significant insights from textual data. To analyze and detect patterns in numerical data, machine learning models such as Random Forests [9] and Gradient Boosting Machines (GBM) [10] would be used.

BERT for qualitative data: A language model called BERT (Bidirectional Encoder Representations from Transformers) is based on the Transformer architecture and uses self-attention techniques to process input data efficiently. In contrast to conventional models that process text in a linear manner, either from left to right or from right to left, BERT is bidirectional, meaning it can concurrently take into account the context from both sides. Due to this advantage, BERT can produce more complex and nuanced comprehension of language since, it can comprehend a word's complete context by looking at the words that surround it. BERT is able to capture intricate relationships and dependencies within the text because to the Transformer design, which comprises of numerous layers of feed-forward neural networks and attention heads. BERT performs better than other natural language processing systems in large part because of its architectural design.

Tree-based Models for quantitative data: Two machine learning models that use ensemble learning techniques to improve predictive performance are Gradient Boosting Machines (GBM) and Random Forest. In order to reduce overfitting and maximize generalization, Random Forest constructs several decision trees during the training phase.. Each tree is trained using a distinct subset of the data and characteristics. To provide a final output that is more reliable and accurate than any one tree alone, the model integrates the predictions of all the trees. While GBM creates trees in a sequential manner, focusing more on the misclassified data points in each new tree, the method corrects issues in prior trees. Optimizing gradient descent reduces the loss function, which is how this is accomplished. This is achieved through gradient descent optimization, which minimizes the loss function. Shrinkage and subsampling are two regularization techniques included in GBM to prevent overfitting. In order to build a strong predictive model that captures complex patterns and interactions in the data, Random Forest and GBM both make use of the power of multiple weak learners (trees).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 07, July-2024

G. Feature Engineerings:

- BERT for qualitative data : Lemmatization, tokenization, lowercasing, and punctuation removal are used to standardize the input while preprocessing text data for BERT. BERT-generated contextual embeddings are utilized to extract key features such named entities, themes, and sentiment ratings. These embeddings are able to capture the nuances of words based on their context. Owing to these features, BERT can get deep insights from qualitative data, which is an essential skill for understanding the opinions and choices of auditors.
- Tree based Models for quantitative data : Normalizing and scaling numerical data, dealing with missing values, and encoding categorical variables are all part of feature engineering for tree-based models like Random Forest and Gradient Boosting Machines (GBM). To improve the models' ability to grasp intricate patterns, derived features like financial ratios and interaction terms are developed. In order to maintain the most relevant features and enhance the models' performance for a variety of tasks, feature selection utilizing significance scores and dimensionality reduction is employed.

a) Model Evaluation:

- BERT for qualitative data: Its effectiveness in removing themes, sentiment ratings, and named entities from qualitative data is evaluated using accuracy, precision, recall, and F1-score. Cross-validation and validation datasets are used to guarantee robustness and generalizability. Additionally, qualitative evaluations attest to the significance of the gleaned insights for audit decision-making.
- Tree-based models for quantitative data: The assessment of Random Forest and GBM is based on its capacity to forecast financial results, identify irregularities, and categorize transactions. Metrics including accuracy, precision, recall, F1-score, and AUC-ROC are employed to assess their effectiveness. The relevance of the features used is confirmed using feature importance scores, and cross-validation is utilized to guarantee generalizability.

C. Workflow Automation:

3) Automation Pipeline:

Placement: Throughout the workflow, the automation pipeline would be used in several phases. In order to prepare the data in a format that could be used as model input, the pipeline would first handle data extraction and preprocessing. After model feeding, the pipeline would oversee how model outputs were integrated into the audit process, producing audit reports and identifying irregularities that needed more examination..

Workflow Steps: The entire process—from data entry to the creation of the final report—would be streamlined by the automated auditing procedure. Data extraction, transformation, model inference, and outcome reporting would be important phases.

Integration Points: The interaction between data pipelines and AI models would be detected through the identification of critical integration points. This comprises the points where the models acquire data, the decision-making process uses the model outputs, and the final audit reports are produced.

4) Process Optimization:

Efficiency Gains: Automation of repetitive operations, reduction of human labor, and acceleration of audit procedures would be achieved through optimization of the automation pipeline.

Error Reduction: The implementation of validation checks and error correction procedures within data pipelines will ensure good data quality and accuracy, hence mitigating the chance of errors in the audit reports at the end.

III. CONCLUJSION

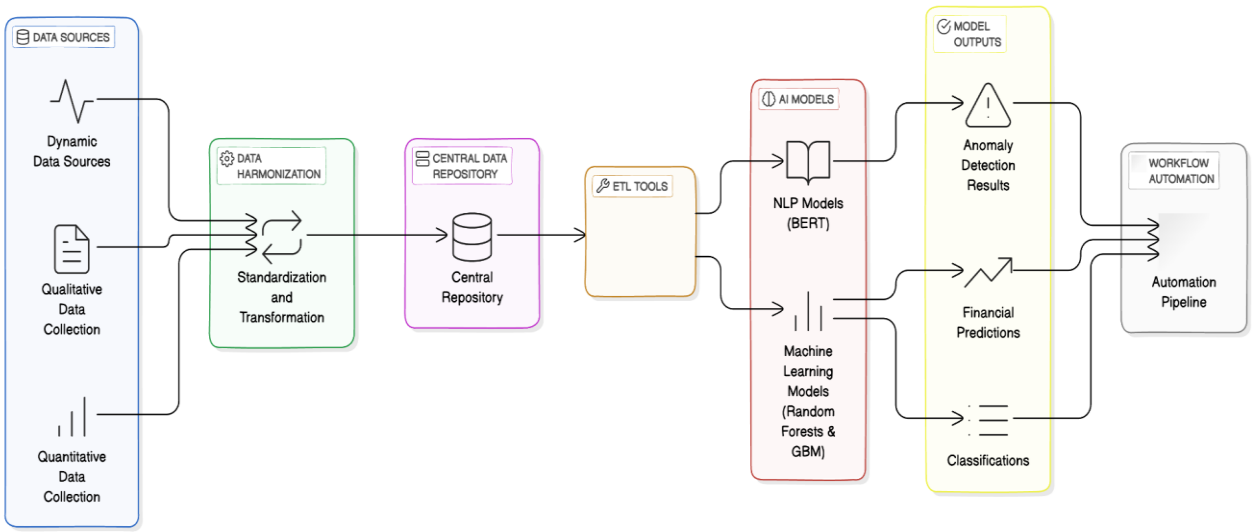

Fig. 1. Data Pipeline Architecture for Automating Financial Audit

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 07, July-2024

This article provides an in-depth approach and data pipeline architecture for leveraging AI and data pipelines to automate financial audits. To improve the effectiveness, precision, and dependability of financial audits, the proposed structure combines qualitative and quantitative data sources. The data flow begins with the collection of various data types, which are integral to the proposed architecture:

Data Sources:

- Qualitative Data Collection: Insights from interviews and surveys.
- Quantitative Data Collection: Historical financial records and transaction logs.
- Dynamic Data Sources: Real-time financial market data and regulatory updates.

The subsequent phase in the process is data harmonization, which standardizes and transforms these disparate data sources to guarantee consistency. Using ETL tools, the harmonized data is loaded, retrieved, and converted from a central repository where it is kept. The processed data is fed into AI Models:

- NLP Models (BERT) handle qualitative data.
- Machine Learning Models (Random Forests, GBM) process quantitative data.

The outputs generated by the AI models are integrated into the workflow automation pipeline and include financial predictions, classifications, and results from anomaly detection. Using this pipeline, audit reports are automatically generated, and abnormalities are marked for additional examination.

IV. INTERPRETATION AND IMPLICATGIONS

The amalgamation of qualitative perspectives and quantitative data facilitates a more comprehensive and precise auditing procedure. advanced AI models can greatly enhance the accuracy of classifications, financial forecasts, and anomaly detection. Automating the audit process reduces manual work and errors, leading to more efficient and successful audits. Regularly updating data and models ensures regulatory compliance and continuously improves the precision of financial audits.

V. LIMITATIONS AND FUTURE SCOPE

Proposed architecture provides a robust The proposed architecture for automating financial audits provides a robust architecture, However, certain limitations that need to be addressed:

- Data Quality: The quality of data obtained from varied sources imposed significant impact on effectiveness of AI models. Moreover, obtaining high quality data is also a challenge.
- Model Interpretability: AI models can be challenging to understand, especially complicated ones like BERT and GBM. In order to ensure transparency in the audit process, future research ought to concentrate on improving these models' interpretability.
- Scalability: For the architecture to manage complicated audit situations and massive data volumes, it must be validated and scaled for performance.
- Real-Time Processing: the timeliness and relevance of audit findings could be significantly enhanced by implementing real-time data processing capabilities.
- Integration with Existing Systems: Seamless integration with existing financial and auditing systems is crucial for practical implementation. Future work should explore ways to achieve this integration efficiently.

In conclusion, the proposed data pipeline architecture represents a significant advancement in the automation of financial audits, leveraging the strengths of both qualitative and quantitative data through sophisticated AI models. Addressing the identified limitations and exploring the suggested future research directions will further enhance the robustness and applicability of this approach in real-world financial auditing scenarios

ACKNOWLEDGMENT

I would like to thank Editor Aakash for their valuable assistance in editing this paper. His attention to detail has helped improve the manuscript.

REFERENCES

[1] H. Zhao and Y. Wang, ‘A Big Data-Driven Financial Auditing Method Using Convolution Neural Network’, IEEE Access, vol. 11, pp. 41492–41502, 2023, doi: 10.1109/ACCESS.2023.3269438.

[2] A. Nizamdinova, A. Kzykeyeva, and A. Arystambayeva, ‘Introduction of artificial intelligence technologies in the organization of auditing activities’, ECONOMIC SERIES OF THE BULLETIN OF THE L.N. GUMILYOV ENU, vol. 143, no. 2, pp. 285–295, 2023, doi: 10.32523/2789-4320-2023-2-285-295.

[3] S. Ivakhnenkov, ‘Artificial intelligence application in auditing’, Scientific Papers NaUKMA. Economics, vol. 8, no. 1, pp. 54–60, Oct. 2023, doi: 10.18523/2519-4739.2023.8.1.54-60.

[4] Frederick Owusu Ampofo, Joseph Elikem Kofi Ziorklui, Nicholas Nyonyoh, and Bernard Owusu Antwi, ‘Integrated predictive analytics in IT audit planning’, Finance & Accounting Research Journal, vol. 6, no. 7, pp. 1291–1309, Jul. 2024, doi: 10.51594/farj.v6i7.1324.

[5] A. R. Munappy, J. Bosch, and H. H. Olsson, ‘Data Pipeline Management in Practice: Challenges and Opportunities’, 2020, pp. 168–184. doi: 10.1007/978-3-030-64148-1_11.

[6] U. Solanki, K. Mehta, and V. K. Shukla, ‘Robotic Process Automation and Audit Quality: A Comprehensive Analysis’, in 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, Mar. 2024, pp. 1–4. doi: 10.1109/ICRITO61523.2024.10522375.

[7] T. Nunes, J. Leite, and I. Pedrosa, ‘Intelligent Process Automation: An Overview over the Future of Auditing’, in 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), IEEE, Jun. 2020, pp. 1–5. doi: 10.23919/CISTI49556.2020.9140969.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019, doi.org/10.48550/arXiv.1810.04805

[9] L. Breiman, ‘Random forests’, Mach Learn, vol. 45, no. 1, 2001, doi: 10.1023/A:1010933404324.

[10] Z. He, D. Lin, T. Lau, and M. Wu, ‘Gradient boosting machine: a survey point zero one technology’, ArXiv, vol. 1908.06951, 2019, doi: 10.48550/arXiv.1908.06951.