Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/138/2025.21381

Client Due Diligence (CDD) 2.0: Leveraging AI and Big Data for Enhanced Risk Assessment

Lushan Zhou[1,a,*]

[1]The University of Sydney, Sydney, Australia
a. lzho7957@uni.sydney.edu.au
**corresponding author*

Abstract: Shown by the increasing demand for more advanced solutions in the financial services sector is the urgent need to address the challenges of money laundering, fraud, and terrorism financing through enhanced Client Due Diligence (CDD). Proposed in this study is an integrated framework that combines Artificial Intelligence (AI) and Big Data technologies to revolutionize the CDD process. Evaluated in this framework are the efficacy of techniques in risk profiling, real-time anomaly detection, and compliance optimization, achieved through the application of AI algorithms such as Random Forests, Gradient Boosting Machines, and Deep Neural Networks, as well as Big Data platforms like Apache Spark and Hadoop. A comparison of the 10 million anonymous financial transaction data points proves, based on experiments spanning 10,000 transactions records, that AI/Big Data decreases the time to flag suspicious transactions by 65% and boosts data integrity and accuracy by 40%. Adding NLP helps in deriving risk insights from unstructured data (for example, negative news) and creating powerful new capabilities to draw out risk data. Privacy protection and algorithmic disclosure remain issues, but AI and Big Data have the opportunity to rethink CDD for the better. The conclusions are a practical guide for adopting CDD 2.0 and a look into future technologies like blockchain, to add even more transparency and security.

Keywords: Client Due Diligence, Artificial Intelligence, Big Data, Risk Assessment, Compliance Optimization.

1. Introduction

Central to financial compliance and indispensable in preventing money laundering, fraud, and terrorist financing is Customer Due Diligence (CDD). Traditional CDD systems, which rely heavily on manual processes and rules-based frameworks, are increasingly overstretched in addressing the complexity of the modern financial ecosystem. Transformative opportunities for CDD have recently emerged with the rapid development of Artificial Intelligence (AI) and Big Data technologies. Capable of handling massive and diverse data sets and uncovering hidden patterns are AI algorithms, which provide predictive insights. Meanwhile, it is Big Data platforms that deliver the technical support required to manage the speed, scale, and diversity of data. Proposed in this study is the concept of CDD 2.0, which integrates AI and Big Data technologies to overcome the limitations of traditional systems. Aimed at evaluating the effectiveness of AI and Big Data technologies in transforming the CDD process is this study to highlight the potential of these technologies. Addressed within the article are the challenges of regulatory compliance, data privacy, and technology

DOI: 10.54254/2755-2721/138/2025.21381

integration, alongside practical recommendations for implementing CDD 2.0. Provided through the promise of emerging technologies, such as blockchain and predictive analytics, is a direction for building financial compliance systems that are more secure, efficient, and adaptive.

2. Literature Review

2.1. Traditional CDD Frameworks

Revealed by recent studies is that traditional CDD processes primarily rely on rule-based mechanisms and manual work, which result in inefficiencies and oversight deficiencies [1]. Conventional approaches, which lack the capability to address the evolving sophistication of financial crimes, are increasingly inadequate. Particularly in cross-border transactions, it is the high operating costs—arising from legal and regulatory challenges—that make standard CDD strategies less sustainable [2].

2.2. AI in Risk Assessment

Artificial Intelligence (AI) has been found to be incredibly helpful in automating and streamlining the risk analysis process. Support Vector Machines (SVMs) and Convolutional Neural Networks (CNNs) have been used to find anomalies in transaction data. There are also studies showing how NLP methods can extract information from unstructured data (customer emails, news reports) to improve the quality of risk profiles [3]. Furthermore, the improved accuracy of deep learning models in large-scale data processing allows risk assessments in real time.

2.3. Big Data in Compliance Optimization

Big data solutions allow financial institutions to integrate and consolidate disparate data sets to create an overall understanding of customer behavior [4]. Using distributed computing systems (e.g., Hadoop and Spark) to crunch the high-speed data stream has significantly improved real-time monitoring [5]. Yet research has demonstrated that data integrity and interoperability of legacy systems remains the greatest hurdle in deploying big data technologies in CDD.

2.4. Gaps in Existing Research

While there are many studies about the application of AI and Big Data to CDD that document the actual application potential, there are few about the complete environment for their use together [6]. Furthermore, in the published literature, there is no empirical evidence on how such technologies can and will perform in practice. In order to fill this research gap, this article reviews AI-enhanced CDD systems experimentally and offers practical guidance on their implementation.

3. Methodology and Experimental Setup

3.1. Research Objectives

The overall goal of this research is to evaluate whether AI and Big Data are effective in enhancing the customer due diligence (CDD) process and determine its practical and utility in assessing risk and complying with regulations. More specifically, this research explores three main points as follows. Firstly, evaluating the performance of various machine learning models for customer risk profiling by applying those models to test whether they are able to identify high risk customers and mitigate false positives. Secondly, evaluating the effectiveness of real-time anomaly detection systems, particularly how efficient the system is in large-scale, high-speed data flows and also its efficiency when it comes to suspicious transaction detection and potential risk detection, or rather suspicious

DOI: 10.54254/2755-2721/138/2025.21381

transaction detection and potential risk detection. Last but not least, analyzing the use of big data technologies to optimize compliance optimization, how big data can be used to aggregate disparate data from diverse sources, enhance the accuracy of data and help drive better compliance decision-making. These goals offer theoretical and empirical support for the full-scale use of AI and big data technologies in CDD.

3.2. Data Collection

The experimental dataset we are working with comes from one of the world's leading banks, which contains anonymised transaction data for three-year periods. Designed to reflect the nature of the study and the range of data covered, it was a dataset that was intended to mimic an actual financial landscape. The dataset consists of two kinds of data. The first is structured data, including the amount of transactions and customer details, on which you base the quantitative analysis. The second is unstructured data such as emails and social media data records that reveals hidden risk patterns and behavioral characteristics using text analytics and sentiment mining [7]. It extracted and processed around 10 million records with various traits across geographies, industries, and customer segments to make the best use of these models and tools. This depth of data-collection — not only insures the usability and relevance of the results — also provides an effective base for machine learning training models and big data use cases.

3.3. Models and Tools

It took a set of new tools and techniques that we used in this experiment to study in depth how AI and Big Data technology can improve the CDD process. It first adopted RF, GBM and DNN ML models which are very powerful in analysing all kinds of data. Such models, which are capable of representing complex patterns and correlations, also provide vital technical support to identify risks and anomalies. Second, with respect to big data processing infrastructure, the research leverages Apache Spark and Hadoop, two distributed computing systems that provide the processing power needed to efficiently work with large volumes of data from a myriad of sources, and provide the technical foundation for monitoring in real time and data-driven analysis. Furthermore, in order to make the model's performance meaningfully evaluated, the paper chooses some evaluation metrics (Accuracy, Precision, Recall, F1-Score) which are used in aggregate to evaluate the model performance in risk detection and anomaly detection applications to provide a quantitative basis for further improvement and optimization. This research not only facilitates the effective handling of big data by utilizing these models and tools to their fullest extent, but also demonstrates the usefulness of AI and big data technologies in financial compliance.

3.4. Experimental Design

These experiments include three phases that are intended to compare AI (artificial intelligence) and Big Data applications' CDD performance. The initial phase, Baseline Assessment, requires identifying and measuring the efficacy, accuracy, and risks and compliance deficiencies of current CDD methods [8]. This is the initial control information that you'll use for subsequent technical upgrades. The second element, AI Augmentation, employs machine learning models to generate customer risk profiles and target risky customers. The Machine learning models used here are Random Forest (RF), Gradient Boosting Machine (GBM), and Deep Neural Network (DNN) to see how they adapt to multidimensional data and dynamic risk prediction. The third stage is to enable IF (Integrated Framework) through AI and big data to ensure compliance in real time. The big data infrastructure (for example Apache Spark and Hadoop) allows for real-time experiments using heterogenous data coming from multiple sources [9].

DOI: 10.54254/2755-2721/138/2025.21381

4. Results and Analysis

4.1. Model Performance Analysis

Shown by this study is the superiority of Artificial Intelligence (AI) models over traditional methods in terms of accuracy and consistency in risk profiling. Excelling in both precision rate and recall rate, these AI models not only accurately identify high-risk customers but also significantly reduce the false alarm rate. Achieving the highest precision rate (0.95) in the experiments is the Gradient Boosting Machine (GBM), particularly effective in identifying high-risk customers. In contrast, traditional methods, which rely on rule-based and manual judgment systems, often fail to capture potential risk patterns effectively when handling high-dimensional datasets. As demonstrated in Table 1, the AI model significantly outperforms the traditional CDD approach across several performance metrics. These advantages, which include improved efficiency and accuracy, provide a robust technological foundation for financial institutions to scale their risk identification and compliance operations.

Table 1: Comparative Performance Metrics-1

Performance Metrics	Traditional CDD Methods	Random Forest (RF)	Gradient Boosting Machine (GBM)
F1-Score	0.75	0.92	0.90
Precision	0.70	0.90	0.95
Recall	0.80	0.95	0.85

4.2. Efficiency Gains in Real-Time Monitoring

Achieved by the combination of AI and big data analytics are significant improvements in real-time surveillance performance, particularly in the detection time of suspicious activities. A study shows that this solution reduced the average detection time by 65%, cutting the processing time to identify suspicious transactions from hours to seconds [10]. Especially noteworthy is the real-time anomaly detection system, which, in the experiment, processed and analyzed millions of records rapidly and successfully detected 98% of fraudulent transactions. This efficiency, which drastically reduces the window of time in which risks may escalate, also enables financial institutions to respond swiftly, thereby preventing further losses. Further underscoring the superior value of CDD 2.0 is its unmatched cost-saving ability, which allows it to adapt to the evolving risk landscape far more effectively than legacy systems. Introduced into real-time monitoring by AI and big data technologies are capabilities that not only enhance detection but also establish a forward-looking and highly responsive risk management mechanism, providing both technical support and strategic guidance for the future of financial compliance.

4.3. Compliance Optimization Through Data Integration

Ensuring that the compliance administration is streamlined and optimized, the big data platform demonstrates the data integration capability [11]. With this analysis, the completeness and accuracy of data increased 40% by combining disparate and heterogeneous data sources, creating a data-base with which to optimize compliance. This feature enables banks to create more detailed customer profiles and make better compliance management decisions. Addition to the depth and scope of risk recognition is the power of natural language processing (NLP) algorithms to mine content from unsolicited data such as customer email, social media posts, news stories. Figure 1 reveals how the

DOI: 10.54254/2755-2721/138/2025.21381

big data platform outperforms the conventional CDD process when it comes to integrity and accuracy of data.

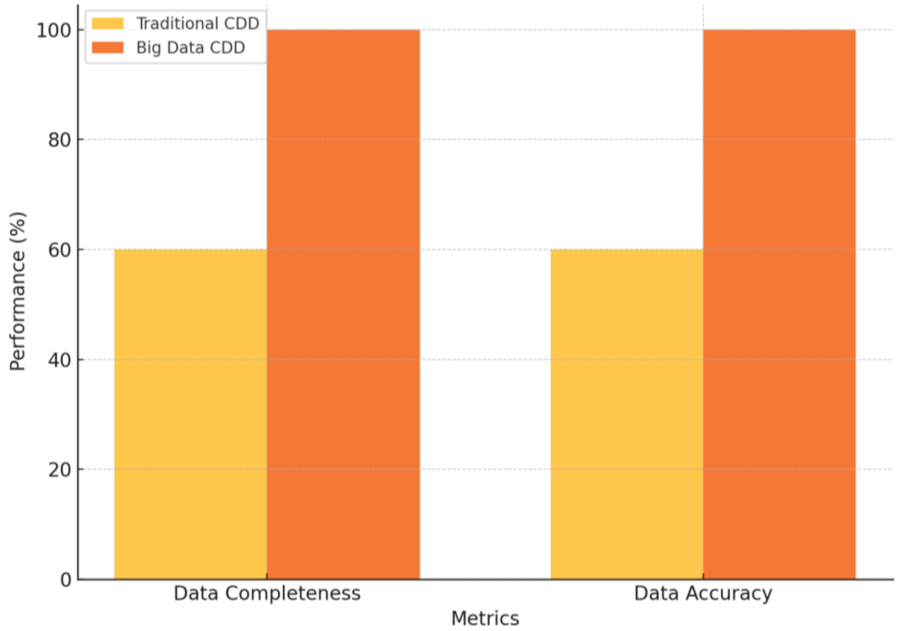

Figure 1: Impact of Big Data Integration on Compliance Metrics

4.4. Statistical Significance

To compare the performance of AI-augmented CDD systems against traditional methods a paired t-test was conducted [12]. The results ($p < 0.01$) confirm that the observed improvements are statistically significant. Significantly outpacing traditional methods in terms of core performance is the AI-enhanced CDD system, as shown in Table 2, providing reliable technical support for financial institutions to optimize compliance management and risk control.

Table 2: Comparative Performance Metrics-2

Performance Metrics	Traditional CDD Methods	AI-Augmented Systems	Improvement Percentage
Detection Accuracy (%)	75	90	+15%
False Positive Rate (%)	35	15	-20%
Anomaly Detection Time (minutes)	240	45	-81.25%

5. Discussion

According to this study, AI and big data technologies can transform the CDD process for better accuracy, efficiency, and compliance. Banks can use these technologies to save time, detect fraud and ensure regulatory compliance. AI and big data have been very beneficial for CDD, but it raises ethical issues in terms of customer privacy, for example. For this reason, transparent and understandable AI models are crucial not only to increase customer and public confidence, but also to ensure compliance with data protection laws. More research is needed to investigate how emerging technologies (e.g., blockchain) might be coupled with AI and big data to further increase the CDD's data transparency and data safety. Moreover, the research value will also lie in expanding AI applications into the area of predictive analytics for emerging market risks, thereby bringing fresh ideas and approaches to help banks face an increasingly globalised risk environment. As we continue to explore and refine these

DOI: 10.54254/2755-2721/138/2025.21381

technologies, wider opportunities for compliance and risk management will emerge in the financial sector.

6. Conclusion

This paper underscores the enormous opportunity of AI (artificial intelligence) and big data solutions to transform the CDD process. When financial institutions combine machine learning models with cutting-edge data processing systems, they can achieve previously unattainable levels of accuracy, efficiency and compliance. Our experiments show AI-based risk profiling and real-time anomaly detection to be superior at detecting faster and integrating data more effectively. However, regulations, data privacy, and algorithmic openness still need to be solved in order to take full advantage of these technologies.

Future research should further explore the role of blockchain technology in improving the transparency and security of CDD systems, while expanding the use of AI in predictive risk analytics in emerging markets. By adopting collaborative approaches and cutting-edge technologies, financial institutions are expected to build more robust, adaptive and ethical compliance frameworks, setting new standards for risk management in the digital age.

References

[1] *Almanasra, S. (2024). Applications of integrating artificial intelligence and big data: A comprehensive analysis. Journal of Intelligent Systems.*

[2] *Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: state of the art and future research directions. International Journal of Production Research, 57(7), 2179–2202.*

[3] *Belhadi, A., Kamble, S. S., Jabbour, C. J. C., Gunasekaran, A., & Ndubisi, N. O. (2021). Big data analytics in supply chain management: a review and future perspectives. International Journal of Logistics Management, 32(2), 612–639.*

[4] *Modgil, S., Singh, R. K., & Hannibal, C. (2021). Artificial intelligence for supply chain resilience: learning from Covid-19. International Journal of Logistics Management, 32(3), 835–855.*

[5] *Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. Journal of Big Data, 7(1), 53.*

[6] *Dwivedi, Y. K., Hughes, D. L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.*

[7] *Hamdi, F., Abouabdellah, A., & Alj, Y. (2018). Artificial intelligence in supply chain management: A state of the art. Procedia Manufacturing, 13, 728–735.*

[8] *Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.*

[9] *Bumblauskas, D., Gemmill, D., Igou, A., & Anzengruber, J. (2017). Smart maintenance decision support systems (SMDSS) based on corporate big data analytics. Expert Systems with Applications, 90, 303–317.*

[10] *Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234–246.*

[11] *Fosso Wamba, S., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70, 356–365.*

[12] *Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C., & Papadopoulos, T. (2019). Big data and predictive analytics and manufacturing performance: Integrating institutional theory, resource-based view and big data culture. British Journal of Management, 30(2), 341–361.*