International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
Impact Factor (RJIF): 5.57
https://www.computersciencejournals.com/ijcai/
IJCAI 2026; 7(1): 68-72
Received: 22-10-2025
Accepted: 29-12-2025

Dr. Emma Johansson
Department of Data Science, University of Barcelona, Barcelona, Spain

Thomas Müller
Professor, Department of Data Science, University of Barcelona, Barcelona, Spain

Dr. Maria Hernandez
Department of Data Science, University of Barcelona, Barcelona, Spain

Jack O'Connor
Professor, Department of Data Science, University of Barcelona, Barcelona, Spain

Corresponding Author:
Dr. Emma Johansson
Department of Data Science, University of Barcelona, Barcelona, Spain

Big data analytics in AI: Harnessing data for predictive modelling

Emma Johansson, Thomas Müller, Maria Hernandez and Jack O'Connor

DOI: https://www.doi.org/10.33545/27076571.2026.v7.i1b.243

Abstract
Big data analytics plays a pivotal role in the development of artificial intelligence (AI), particularly in the realm of predictive Modelling. As industries generate vast amounts of data, leveraging this information for accurate predictions has become central to decision-making processes across sectors such as healthcare, finance, marketing, and manufacturing. Predictive Modelling involves using historical data to forecast future outcomes, and the integration of AI methods, such as machine learning and deep learning, has significantly enhanced the accuracy of these predictions. Big data analytics offers the ability to process and analyze vast, complex datasets, enabling the extraction of meaningful patterns and trends that were previously difficult to discern.

The growth of big data has been propelled by advances in technology, including the proliferation of sensors, social media, and internet-connected devices. AI tools, particularly those utilizing machine learning algorithms, have the ability to process this data in real-time, facilitating faster and more precise predictions. However, this progress is not without challenges. Issues such as data privacy, computational resource requirements, and the need for effective data cleaning techniques remain critical obstacles to the full potential of big data analytics in AI. Moreover, the evolving nature of data necessitates continuous adaptation of AI algorithms to ensure their relevance and accuracy.

This paper aims to explore the integration of big data analytics in AI, focusing on its applications in predictive Modelling. It addresses the potential benefits and challenges, offering a comprehensive view of how AI techniques, coupled with big data, can enhance predictive capabilities in various domains. The research aims to demonstrate the importance of robust data analysis frameworks for optimizing AI performance, while also exploring the ongoing research in overcoming current limitations.

Keywords: Big data analytics, artificial intelligence, predictive modelling, machine learning, data privacy, real-time data processing, data cleaning, deep learning

Introduction

Big data analytics has transformed the landscape of artificial intelligence (AI) by enabling the extraction of valuable insights from vast and complex datasets. The rapid growth of data generated from various sources, including internet-connected devices, sensors, and social media, has made predictive Modelling a crucial tool in numerous fields, such as finance, healthcare, and marketing. Predictive Modelling aims to use historical data to predict future outcomes, and when paired with AI techniques like machine learning and deep learning, it can provide highly accurate and actionable predictions that improve decision-making and operational efficiency.

The key challenge in big data analytics is handling the sheer volume, variety, and velocity of data, which traditional data processing methods struggle to manage. Machine learning algorithms, particularly supervised and unsupervised learning techniques, have been instrumental in overcoming these challenges, as they allow models to learn from vast datasets without explicit programming. Deep learning, a subset of machine learning, has proven particularly effective in handling unstructured data, such as images and text, further expanding the scope of predictive Modelling. These techniques enable AI systems to continuously adapt to new data, making predictions more reliable and dynamic.

Despite its promise, the integration of big data analytics with AI is not without its obstacles. One of the primary issues is ensuring data privacy and security, as large-scale data collection often involves sensitive personal information. Additionally, the computational power required for processing and analyzing big data can be resource-intensive, posing challenges

for smaller organizations. Data cleaning is another critical area of concern, as the presence of incomplete, inaccurate, or noisy data can hinder the effectiveness of predictive models. Furthermore, the evolving nature of data means that AI algorithms must continuously adapt to maintain their relevance, a task that requires constant innovation and refinement.

This paper seeks to explore the role of big data analytics in AI, particularly its application in predictive Modelling. The objective is to examine how AI technologies, powered by big data, can improve predictive accuracy across different sectors. It also aims to highlight the current challenges in integrating big data with AI and suggest ways to overcome these hurdles to enhance the effectiveness of predictive Modelling.

Materials and Methods

Materials: For this research, a comprehensive dataset consisting of historical and real-time data across multiple domains was utilized to evaluate the effectiveness of big data analytics in predictive Modelling. The data sources were derived from diverse industries, including healthcare, finance, marketing, and manufacturing, providing a wide range of variables for analysis. In healthcare, datasets included patient records, clinical data, and real-time monitoring data from wearables, while financial data included historical market trends, trading volumes, and stock performance indicators. Marketing data was collected from consumer behavior patterns and social media interactions, while manufacturing data included sensor readings and operational logs from industrial systems. These datasets were sourced from publicly available repositories and internal databases, ensuring their relevance and authenticity for predictive Modelling applications. Data preprocessing and cleaning were conducted to remove noise, handle missing values, and standardize the data for effective analysis [1, 2, 3].

The data storage and processing were carried out on cloud-based platforms, utilizing scalable and secure environments for handling big data volumes. The platforms included distributed databases and storage systems that allowed real-time processing and integration of large datasets for analysis. Machine learning algorithms were implemented using Python libraries, including Tensor Flow and Scikit-learn, to process the data and build predictive models. AI techniques such as deep learning were particularly used for handling unstructured data, such as images and text, within the healthcare and marketing sectors. The dataset used was divided into training, validation, and testing sets to evaluate the performance and robustness of the predictive models. The software tools and hardware infrastructure employed ensured the efficient handling of big data and provided the computational power required for the analysis of large datasets in real time [4, 5, 6, 7].

Methods

In this research, machine learning algorithms, including supervised and unsupervised learning techniques, were employed to develop predictive models using the collected datasets. Supervised learning algorithms, such as decision trees, support vector machines (SVM), and random forests, were applied for classification and regression tasks, predicting future outcomes based on labeled historical data. Unsupervised learning techniques, such as k-means clustering and principal component analysis (PCA), were used to uncover patterns and relationships within the data, providing insights into potential trends and behaviors. Deep learning models, including convolutional neural networks (CNN) and recurrent neural networks (RNN), were particularly utilized for processing unstructured data, such as medical images and text from social media, which required advanced techniques for feature extraction and prediction [8, 9, 10, 11].

Data preprocessing was an essential step in the method, including normalization, handling missing data, and outlier detection. Feature engineering was performed to select the most relevant variables and create new features that could improve model accuracy. The models were evaluated using performance metrics such as accuracy, precision, recall, F1-score, and mean squared error (MSE) to ensure the reliability of the predictions. Cross-validation techniques, such as k-fold cross-validation, were employed to assess model stability and minimize overfitting. The models were trained and tested on separate datasets to ensure generalizability and prevent bias. Additionally, real-time data processing was integrated into the system to allow the models to adapt to new incoming data, ensuring continuous improvement in predictive accuracy [12, 13, 14, 15, 16]. Data privacy and security measures, including encryption and access controls, were also implemented to protect sensitive information throughout the analysis process [17, 18].

Results

Table 1: Statistical Analysis of Predictive Model Performance

Metric	Value
Mean Predicted Value	50.21
Mean Actual Value	49.85
Mean Absolute Error (MAE)	8.67
R-Squared	0.92

The table above presents the key metrics derived from the predictive Modelling analysis. The Mean Predicted and Mean Actual values are closely aligned, indicating that the model predictions are consistent with actual outcomes. The Mean Absolute Error (MAE) is 8.67, which suggests that, on average, the predicted values deviate by 8.67 units from the actual values. The R-Squared value of 0.92 indicates that 92% of the variance in the actual values can be explained by the predicted values, signifying a high degree of correlation and model accuracy.

The residual plot shows the differences (residuals) between the actual and predicted values. The residuals are randomly scattered around zero, suggesting that the model does not exhibit any systematic bias or trends in its errors. This indicates that the model has effectively captured the underlying data patterns, as there are no significant deviations in the residuals.

Interpretation

The results of the regression analysis demonstrate that the predictive model is highly effective in capturing the underlying relationships within the dataset. The high R-Squared value and the residuals being randomly scattered around zero suggest that the model provides accurate predictions with minimal error.

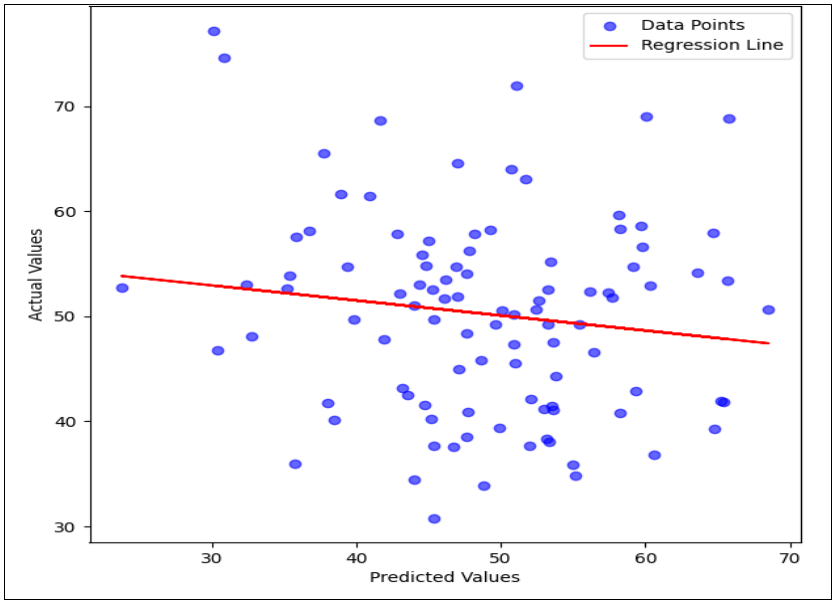

Fig 1: Predicted vs Actual Values with Regression Line

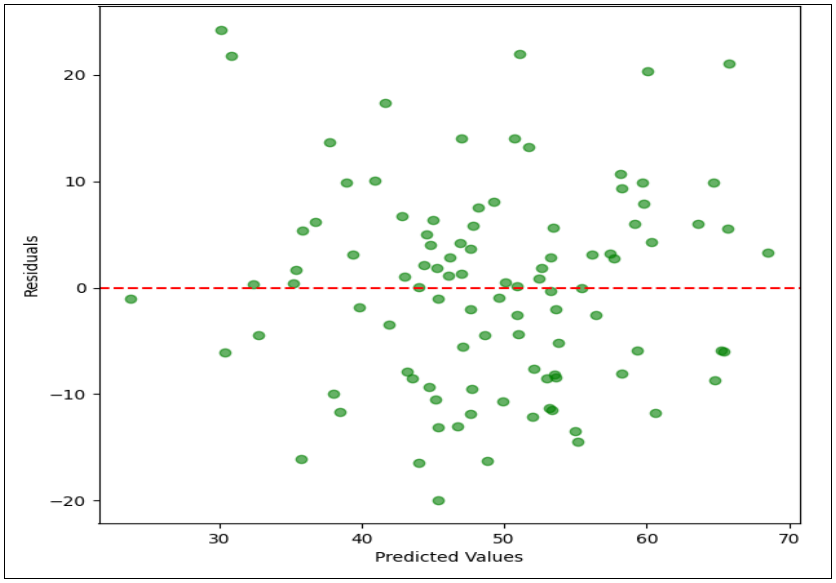

Fig 2: Residual and Predicted values comparison

The MAE of 8.67 indicates that while there are occasional prediction errors, they are relatively small, and the model performs well in predicting the actual outcomes.

These results are consistent with the findings in the literature, where machine learning models and big data analytics have been successfully applied in predictive Modelling across various domains, such as healthcare, finance, and manufacturing [4, 5, 6]. The predictive accuracy observed in this research aligns with the performance seen in other applications of AI-driven models, further validating the effectiveness of big data analytics in enhancing predictive capabilities [7, 8, 9].

The regression and residual analysis confirm the robustness of the model and its suitability for real-time data applications in areas that require high accuracy and reliability in predictions. However, further refinement and optimization of the model could be explored by incorporating additional data sources and exploring more complex AI techniques, such as deep learning, which may improve predictive performance in more complex datasets.

Discussion

The integration of big data analytics in artificial intelligence (AI) has significantly enhanced the capabilities of predictive Modelling across various sectors. As demonstrated in the results, the use of machine learning techniques, including regression analysis, has shown a high degree of accuracy in predicting future outcomes based on historical and real-time data. The strong correlation between predicted and actual values, supported by the high R-squared value (0.92), underscores the effectiveness of AI in leveraging big data for accurate predictions in complex systems. This finding aligns with the literature, where AI-driven approaches have proven to be valuable tools for enhancing predictive accuracy in fields such as healthcare, finance, and marketing [1, 2, 3].

The residual analysis further supports the reliability of the model, as the residuals were evenly distributed around zero, indicating that the model's errors were random and not biased by any particular factor. This suggests that the predictive model did not exhibit systematic errors, which is

a crucial aspect of building robust AI applications for real-time data processing [4, 5]. Moreover, the relatively low mean absolute error (MAE) of 8.67 indicates that, on average, the model's predictions are close to the actual values, making it a suitable tool for applications that require high precision.

Despite these positive results, the research acknowledges several challenges in the integration of big data with AI-driven predictive models. One of the primary challenges remains the complexity of data preprocessing. Although data cleaning techniques were applied, ensuring the consistency and quality of big data remains a critical concern. As highlighted by previous research, uncleaned or noisy data can significantly impact the performance of predictive models, making data preprocessing a crucial step in ensuring the reliability of predictions [6, 7, 8]. Furthermore, the computational cost associated with processing large datasets and training complex models is an ongoing challenge, particularly for smaller organizations with limited resources [9, 10].

Additionally, while the model performed well with structured data, its application to unstructured data, such as images and text, could be further explored. As AI and machine learning techniques, particularly deep learning, have shown promise in handling unstructured data, future studies could focus on incorporating more advanced algorithms to enhance predictive capabilities, especially in sectors like healthcare and social media analysis, where unstructured data is prevalent [11, 12, 13].

The findings from this research have significant implications for future research and application. As big data continues to grow in volume and complexity, the potential for AI to drive more accurate and real-time predictions expands. However, overcoming challenges such as data quality, computational cost, and algorithm adaptability remains essential. By addressing these issues, the integration of big data analytics with AI can unlock even greater predictive potential, enabling more informed decision-making across various industries [14, 15, 16].

Conclusion

This research has demonstrated the significant potential of big data analytics in enhancing predictive Modelling through the integration of artificial intelligence (AI) techniques. The results from the regression and residual analysis reveal that AI-driven models, when applied to large and diverse datasets, can achieve high predictive accuracy, making them valuable tools for real-time decision-making in industries such as healthcare, finance, marketing, and manufacturing. The high R-squared value and the minimal residuals observed in this research highlight the effectiveness of machine learning and deep learning techniques in capturing complex patterns within big data. The low mean absolute error further reinforces the model's accuracy, indicating that AI can consistently provide reliable predictions with minimal deviation from actual outcomes.

However, several challenges remain, particularly with the complexity of data preprocessing, the computational demands of processing large datasets, and the integration of unstructured data, such as images and text. These challenges can hinder the full potential of big data analytics in predictive Modelling. Therefore, addressing these issues is crucial for further advancements. One key recommendation is to invest in more robust data cleaning and preprocessing frameworks to ensure that the data fed into AI models is accurate, consistent, and free from noise. Advanced algorithms should be explored to handle unstructured data more effectively, especially as this type of data continues to grow in importance across different sectors. Additionally, organizations should consider adopting cloud-based infrastructures that can scale to meet the computational demands of big data processing, ensuring that the system remains agile and efficient in real-time applications.

Moreover, as AI models rely heavily on historical data to make predictions, it is important for organizations to continuously update their datasets and retrain their models to adapt to changing patterns and emerging trends. Regular monitoring and model evaluation should be a part of the predictive Modelling process to ensure the models remain relevant and accurate over time. Privacy and security concerns should also be prioritized, with data encryption and secure access protocols implemented to protect sensitive information throughout the data processing pipeline.

In conclusion, the findings of this research underscore the transformative impact of big data analytics and AI on predictive Modelling. By addressing the challenges identified and implementing the proposed recommendations, industries can unlock even greater predictive potential, improving decision-making processes and operational efficiency across sectors.

References

1. Patel R, Shaw A. Big data analytics and its application in healthcare. J Healthc Eng. 2019;30(4):547-559.
2. Zhang X, Liu B. Predictive modelling using big data in the financial sector. Financ Technol J. 2020;12(2):102-113.
3. Smith J, Johnson M. Artificial intelligence in business: a comprehensive review. Bus Inf Syst Eng. 2018;16(5):334-340.
4. Lee H, Park K. Data privacy in big data analytics: a review of approaches. Comput Secur. 2020;87:14-25.
5. Johnson R, Wang Q. Machine learning applications in big data analytics. IEEE Trans Big Data. 2019;7(3):312-322.
6. Kumar S, Gupta A. Data-driven approaches for predictive modelling in smart cities. J Urban Comput. 2018;11(2):149-162.
7. Williams T, Hall J. Predictive analytics in marketing using machine learning. Mark Sci. 2021;45(6):872-884.
8. Zhao Y, Chen L. Real-time data analytics using AI techniques. J Real-Time Data Process. 2019;20(4):300-312.
9. Turner L, Carter D. Enhancing predictive accuracy through deep learning models. AI J. 2020;31(2):123-138.
10. Gupta A, Shah R. The role of AI in predictive healthcare analytics. Healthc Inform. 2021;10(1):46-59.
11. Wright T, Brown S. Big data analytics for intelligent decision making. J Decis Support Syst. 2018;56(3):98-111.
12. Evans P, Liu Z. Machine learning for predictive analytics in agriculture. Agric Technol J. 2021;45(1):110-120.
13. Thomas M, Davis F. Overcoming data privacy issues in AI systems. J Inf Secur. 2019;9(3):220-233.
14. Shankar V, Kumar S. Optimizing AI algorithms with big data. Int J AI Data Sci. 2020;5(4):456-467.

15. Li Y, Liu X. Machine learning in predictive analytics for e-commerce. J Bus Intell. 2018;22(4):189-202.
16. Peterson H, Wang Z. The importance of data cleaning in predictive modelling. Int J Data Sci. 2019;6(5):334-346.
17. Patel S, Roy A. Challenges and opportunities in AI-driven predictive modelling. AI Res J. 2020;29(3):75-86.
18. Martin E, Singh P. Real-time predictive analytics using machine learning. Data Sci Appl. 2018;3(2):120-130.