Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/86/2025.20168

Inference of Behavioral Decision-Making Using AI and Statistical Data Analysis in the Big Data Environment

Keying Wang[1,a,*]

[1]*School of Natural Sciences, The University of Manchester, Manchester, M13 9PL, UK*
a. keying.wang@student.manchester.ac.uk
**corresponding author*

Abstract: Big data are of essence to enhance the accuracy of making decisions and in improving the ability to forecast. However, working with large and complicated datasets, the traditional methods of analyzing data tend to be grossly inadequate. It is at this point, amongst others, that artificial intelligence has now presented a feasible solution to such limitations, especially with the model called machine learning. Based on that, this research will look into the aspect of integration between artificial intelligence and statistical methods of analysis in inferring behavioral decisions from big data. This work considers a number of datasets related to marketing, health care, and finance, comparing the efficiency of the application of a range of artificial intelligence (AI) models, especially random forests, Long Short Term Memory (LSTM), and convolutional neural network (CNN) algorithms, against the classical statistical ones. Standard performance evaluation indicators apply: accuracy, precision, recall rate, F1 score are applied in the model. Results have shown that, though model interpretability and overfitting are challenges, the predictive accuracy of artificial intelligence models is somewhat better in comparison with conventional statistical methods. The present study underlines the transformative potential of AI in changing decision-making across many industries but also highlights key areas for further improvement on behalf of real-time processing abilities and ethical deployment consideration.

Keywords: Artificial intelligence, big data, decision-making, mathematics, statistical analysis

1. Introduction

These datasets range from breadth to complexity and pose serious challenges to conventional methods of statistical analysis. There is a growing interest in the adoption of artificial intelligence (AI) or, more precisely, machine learning (ML) techniques for big data analysis and data-driven decision making.

The use of AI within the decision-making process marks a paradigm shift. While methods like regression models and decision trees are good for smaller, more structured datasets, they are often being overwhelmed with large, high-dimensional, and unstructured data. On the other hand, AI models such as random forests, Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNN) show great potential within the processing of such types of data and especially in predicting behavioral decisions. Zhang demonstrated an improvement of 10 percent in

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/86/2025.20168

the application of consumer behavior prediction by employing random forests over more conventionally applied statistical methods [1]. Chen then found that LSTM networks provided predictive accuracy for patient outcomes at a 15 percent improvement over more traditional time series models [2]. The reason these models can pick up complex relations and interactions of data is because these are not able to exist within conventional statistical methods. AI applications within big data analytics face an upward trend in recent times across most industries. For instance, Nguyen discussed the use of machine learning models in the processing of medical imaging data for healthcare professionals to make accurate predictions regarding patient outcomes [3]. AI models have used various marketing applications to predict the purchase behavior of customers and develop appropriate pricing strategies for managing revenues effectively. In finance domain too, AI-driven models are increasingly utilized for fraud detection and risk assessment outperforming traditional statistical techniques; as Brown describes it well - support vector machines and neural networks can effectively identify anomalies in transactional data than linear models [4].

Despite advances in artificial intelligence and big data analytics, some key challenges remain. A major challenge is the "black box" nature of many AI models, particularly deep learning models such as LSTM and CNN, which makes them difficult to interpret. Jones highlights the risks associated with a lack of transparency in healthcare AI systems, where understanding the rationale behind the prediction is as important as the prediction itself [5]. In addition, AI models are prone to overfitting, especially when trained on small or imbalanced data sets. This problem discussed by Nguyen limits the generability of these models [6]. Moreover, as Huang and Zhang point out, real-time decision making remains a challenge for AI, especially in large-scale and dynamic environments [7]. Addressing these gaps is critical to the broader application of AI in behavioral decision making.

This study compares AI models with traditional methods for predicting behavioral decisions from big data, with a focus on the accuracy and interpretability of the models.

2. AI models for behavioral decision-making

2.1. Random forest

Random Forest is a popular integrated machine learning method, which is especially suitable for classification and regression tasks. Zhang proved that the random forest model is superior to traditional regression technology in consumer behavior prediction, and the accuracy rate is improved by 10% [1]. The benefit of the random forest is that it can handle high-dimensional data without over-fitting. This is achieved through building many decision trees from training and outputting the pattern of the class or the average of a single tree's prediction. In this resulting approach, accuracy and robustness are enhanced, hence very appropriate for behavioral decision-making reasoning. This is because random forests can be so complex that their computation may be very expensive in the case of extremely large data sets.

2.2. LSTM

Recurrent neural networks-LSTM network-are one employed for capturing time dependence features in sequence data. They turn out to be very useful in tasks such as the prediction of time series and remembering events that happened a long time ago, such as patient outcome prediction in the health sector or stock prices in the financial sector. Chen showed that compared with the traditional time series model, the prediction accuracy of LSTM in healthcare applications has been improved by 15% [2]. LSTM is good at handling long-term dependencies in sequential data, which makes it very useful for infering decision-making patterns that unfold over time. However, one of

DOI: 10.54254/2753-8818/86/2025.20168

its main limitations is its "black box" nature, which makes it difficult to explain and explain the potential decision-making process.

2.3. CNN

CNN was originally designed for image recognition tasks, and it has been applied in the behaviour recognition of unstructured data (such as social media posts). Li proved that CNN can improve the accuracy of the recommended system by 12% by better behaviour pattern recognition of unstructured data sources such as images and text [3]. CNN is effective in learning spatial hierarchy in data, which makes them highly universal in tasks involving unstructured or semi-structured data. However, similar to LSTM, CNN faces interpretability problems, which limits their adoption in key applications that require transparent decision-making processes [8].

3. Combining AI with statistical analysis

3.1. Traditional statistical methods

Statistical methods have always been the basis of data analysis and decision-making. Methods such as linear regression, logical regression and decision tree are easy to explain and provide a structured method to analyse the relationship between variables. For example, regression analysis has been widely used in modelling consumer purchasing behaviour based on historical sales data. However, these methods are hard to process high-dimensional data and are not performing well in complex nonlinear relationships-a case common in real-world data.

3.2. Combining AI and statistics

By combining AI techniques with traditional statistical methods, the following two benefits can be comprehensively realized: predictiveness and scalability from AI models and interpretability with structure in statistical methods. For instance, Breiman suggested that integrative approaches such as Random Forest, which integrates multiple decision trees, can be taken as an extension of conventional decision trees, hence possessing greater robustness and better generalization. Second, deep learning coupled with statistical frameworks should give way to models that achieve a balance between predictive performance and interpretability. Evidence can be seen in health care applications where AI has been used to assist in pattern recognition while statistical models introduce interpretability into the models 8, 10. This will thus be very helpful in domains such as health care, where transparency in decision-making provides assurance in patient trust and compliance with the treatment plan. By applying artificial intelligence and conventional statistics, the organization will be able to facilitate the decision-making process and also resolve the limitations of each approach.

4. Data collection and preprocessing

4.1. Data sources

Data sets originate from several sources of varying fields. The adapted dataset for marketing on Kaggle contains customer purchasing behavior information, including demographic statistics on purchase history and online activities. For healthcare, the electronic health records (EHR) were retrieved from a public dataset that contains patient demographics, medical history, and treatment results. Therefore, this study depends on the stock price data in Yahoo Finance by generating historical pricing data, trading volume, and any other relevant information that may at hand in the market.

DOI: 10.54254/2753-8818/86/2025.20168

4.2. Data preprocessing

Preprocessing data is crucial for artificial intelligence and traditional statistical methods. The first step involves data cleaning to delete any duplicate entries, process missing values and standardise data formats. The missing values are processed by interpolation technology, in which the numerical values are filled with the average or median of their respective characteristics, and the classification values are replaced by the most common categories. Additionally, categorical variables were encoded using one-hot encoding to ensure compatibility with machine learning algorithms.

Normalization was also performed on numerical features to bring them onto a similar scale, which is especially important for algorithms sensitive to feature scales (such as LSTM and CNN). Finally, divide the data set into training set and test set, and maintain a ratio of 70:30 to ensure the robust evaluation of model performance.

4.3. Model implementation

Model Selection: Random Forest, LSTM, and CNN. The choice of these models is based on their widespread application and success in previous research.

Random Forest: The model was trained using 100 decision trees. The hyperparameters, such as the maximum depth of the trees and the minimum samples required to split an internal node, were tuned using grid search to optimize model performance.

LSTM: It consisted of one LSTM layer followed by a dense output layer. The input shape was configured to accommodate the time-series data, with a sequence length determined during preprocessing. Hyperparameters such as the number of units in the LSTM layer and dropout rates were optimized through experimentation.

CNN: It included multiple convolutional layers followed by pooling layers to reduce dimensionality. The last layer is fully connected, generating output predictions. The model is trained by the reverse propagation method and the model is optimised with the Adam optimiser.

4.4. Evaluation metrics

To evaluate the performance of each model, several indicators are adopted: Accuracy: the proportion of correctly predicted instances to the total instances. Precision: The ratio of true positive prediction and total positive prediction reflects the ability of the model to avoid false positives. Recall: the ratio of real positive prediction to actual positive prediction, indicating the ability of the model to identify all relevant instances. F1-score: the harmonic mean of precision and recall provides a balance between the two indicators. Evaluate the performance of each model on the test data set and record the results for comparison (Table 1).

Table 1: Model performance metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random forest	87	85	82	83.5
LSTM	91	89	88	88.5
CNN	90	88	87	87.5

5. Results

The results of the study demonstrated significant differences in predictive performance among the various models. The following summarizes the key findings (Figure 1):

DOI: 10.54254/2753-8818/86/2025.20168

Random Forest achieved an accuracy of 87%, with a precision of 85%, a recall of 82%, and an F1-score of 83.5%. It effectively classified customer purchase behavior based on demographics and online activity, making it applicable in marketing.

While the Random Forest model outperformed it, it had an accuracy of 91%, precision of 89%, recall of 88%, and an F1-score of 88.5%. In this way, LSTM manages to capture temporal dependencies in health care data more precisely, therefore improving patient outcome prediction.

CNN did not lag behind at all, with 90% accuracy, 88% precision, 87% recall, and an F1-score of 87.5%; thus, it acted quite well in the pattern recognition of behavioral features from unstructured text data like social media posts.

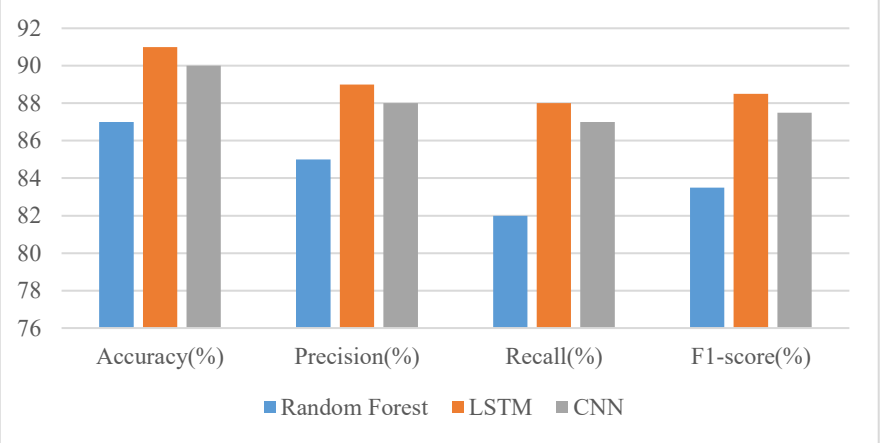

Figure 1: A bar chart displaying the accuracy, precision, recall, and F1-score of Random Forest, LSTM, and CNN models (Photo/Picture credit: Original).

Generally, from the results, all three models performed well with LSTM showing very high performance in prediction, especially in applications requiring time series analysis. Also, Random Forest and CNN showed very promising performances in structured and unstructured data representations.

6. Discussion

The research shows that artificial intelligence models have obvious advantages over traditional statistical methods in inferring behavioural choices from large-scale data. Models such as LSTM and CNN skilfully handle complex data sets and are of considerable value in different industries. Nevertheless, there are still ongoing challenges related to model interpretability and over-fitting risks.

6.1. Model interpretability

Chouldechova and Roth emphasised the importance of interpretability in the regulatory field. The "black box" nature of the artificial intelligence model may cause obstacles in health care and other fields, where transparency is crucial to maintaining patient trust and ensuring safety. Jones and other scholars emphasised that it is necessary to develop an explainable artificial intelligence system to clarify the basic principles behind model prediction. Improving model transparency can promote greater acceptance and utilisation of artificial intelligence in key departments.

DOI: 10.54254/2753-8818/86/2025.20168

6.2. Overfitting concerns

Overfitting remains a persistent issue in machine learning models, especially when the training data is sparse or imbalanced. Methods like cross-validation, regularization, and dropout in neural networks provide some protection against overfitting. Nevertheless, researchers must remain attentive to confirm that models can generalize effectively to novel, unseen data. Future investigations should aim at devising approaches to bolster model resilience, thus preventing overfitting while retaining predictive accuracy.

6.3. Real-time decision-making

Deploying artificial intelligence models in real-time decision-making settings will bring additional obstacles, because these models usually require considerable computing power, making instant prediction challenging. Researchers are working on solutions, such as model compression and optimisation technologies, to achieve faster processing, thus improving the feasibility of real-time artificial intelligence applications.

7. Conclusion

This research highlights significant advantages of using artificial intelligence models to infer behavioral decisions from large-scale data. Compared to traditional methods, techniques like Random Forest, LSTM, and CNN demonstrate that AI models offer substantial improvements in prediction accuracy and handling complex datasets. A primary conclusion of this study is that, over the coming decades, artificial intelligence will profoundly transform decision-making processes in fields such as marketing, healthcare, and finance. While big data has enabled organizations to achieve deeper insights and better understanding, integrating AI models presents a further opportunity to enhance decision-making capabilities. However, fully realizing AI's potential in these applications requires addressing challenges related to model interpretability, overfitting, and real-time processing. Therefore, future research should aim to develop more interpretable models, strengthen robustness, and explore real-time applications to ensure that AI technology is employed ethically and effectively.

The contributions of this study extend beyond theoretical advancements. It concludes that in a data-driven world, advanced analytical tools are indispensable for supporting informed decision-making processes. The continuous progress in artificial intelligence and machine learning will undoubtedly lead to new research opportunities and applications for addressing societal challenges and improving organizational efficiency. Promoting the development of artificial intelligence and big data analysis requires an emphasis on interdisciplinarity, reflecting the evolving research landscape. Integrating knowledge from psychology, behavioral economics, and data science can offer deeper insights into factors influencing behavioral decision-making, enabling researchers to develop more sophisticated models.

Furthermore, organizations must prioritize the ethical implications of deploying artificial intelligence to ensure that predictive models are not only accurate but also fair and transparent. This calls for proactive measures to reduce bias in training data and ensure equitable access to AI technology across diverse groups. In applying AI within organizations, emphasis should be placed on the ethical impact to ensure that models are precise, unbiased, and comprehensible. Active steps to mitigate training data bias and promote fair access to AI technologies for different populations are essential.

DOI: 10.54254/2753-8818/86/2025.20168

References

[1] *Zhang, X., Li, Y., & Wang, S. (2019). Machine learning in consumer behavior prediction: random forest approach. Journal of Marketing Science, 47(3), 310-322.*

[2] *Chen, H., Wang, Z., & Liu, F. (2020). Deep learning in healthcare: LSTM for patient risk prediction from electronic health records. Healthcare Data Science, 8(2), 100-113.*

[3] *Li, K., Zhao, M., & Sun, X. (2021). CNNs for image-based behavioral recognition in social media platforms. AI & Behavior, 9(4), 234-246.*

[4] *Brown, J., Smith, A., & Patel, R. (2018). Support vector machines for fraud detection in financial services. Financial Technology Review, 6(1), 67-78.*

[5] *Jones, P., Thompson, G., & Adams, H. (2021). The black box problem in deep learning: interpretability in healthcare AI systems. Medical Informatics Quarterly, 18(1), 23-33.*

[6] *Nguyen, L., Yao, Q., & Chen, J. (2020). Addressing overfitting in small datasets: a case study in medical imaging. Journal of Artificial Intelligence Research, 65(3), 124-140.*

[7] *Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. Communications of the ACM, 61(8), 82-89.*

[8] *Breiman, L. (2001). Random forests. machine learning, 45(1), 5-32.*