Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/108/2025.20929

Comparative Analysis of Traditional Machine Learning and Deep Learning Methods: Performance on Datasets of Varying Complexity

Kaitao Jiang[1,a,*]

[1]*Computer Science, University of Warwick, Coventry, CV4 7AL, United Kingdom*
a. Kaitao.Jiang@warwick.ac.uk
**corresponding author*

Abstract: This study presents a comparative analysis of traditional machine learning and deep learning methods, evaluating their performance on datasets of varying complexity. Traditional methods such as Random Forests and Naive Bayes exhibit high accuracy and computational efficiency for low-complexity tasks, making them suitable for real-time applications. In contrast, deep learning techniques, including Convolutional Neural Networks (CNNs) and Autoencoders, excel in high-complexity tasks such as image and speech processing but require significant computational resources and longer training times. By analyzing their respective strengths and limitations, this research provides insights into selecting the appropriate algorithm based on dataset complexity and task requirements. The findings highlight opportunities for hybrid models to combine the benefits of both approaches, addressing computational efficiency and accuracy. Future research directions include enhancing deep learning model interpretability and optimizing preprocessing techniques to improve performance in data-scarce scenarios.

Keywords: Machine Learning, Deep Learning, Model Performance, Data Preprocessing.

1. Introduction

Machine learning has emerged as a transformative technology in today's data-driven world, providing powerful tools to analyze complex datasets and solve practical problems [1, 2]. It involves a range of techniques that enable computers to learn from data and make predictions, spanning traditional algorithms and deep learning methods [3]. Traditional algorithms, such as Random Forests and Naive Bayes classifiers, typically perform well on structured data [4, 5]. In contrast, deep learning techniques like Convolutional Neural Networks are good at handling high-dimensional and intricate data [6].

As datasets become increasingly complex and voluminous, the limitations of traditional algorithms become more pronounced. These limitations highlight the growing prominence of deep learning methods, which are better equipped to handle the challenges posed by high-dimensional and heterogeneous data [7]. As data volume grows, selecting the most suitable algorithm becomes crucial to optimize predictive accuracy. Consequently, researchers and engineers are faced with the challenge of improving accuracy while ensuring efficient model performance.

DOI: 10.54254/2755-2721/108/2025.20929

The goal of this study is to compare the performance of traditional algorithms, including Random Forests, Principal Component Analysis, and Naive Bayes, with deep learning techniques, such as Convolutional Neural Networks and Autoencoders, in processing complex data and enhancing model accuracy. By evaluating the efficacy of these methods, we aim to offer practical insights for selecting appropriate algorithms in real-world applications, thereby optimizing both accuracy and efficiency [8]. Understanding how these different approaches perform under various algorithms is key for model optimization.

2. Literature Review

Machine learning algorithms are pivotal in enhancing model accuracy and have been widely adopted across various domains. This section provides an overview of key machine learning techniques and their impact on model performance, classified into traditional machine learning algorithms, deep learning methods, dimensionality reduction techniques, and ensemble approaches.

2.1. Traditional Machine Learning Algorithms

Traditional machine learning methods have wide applications in data analysis and classification tasks. In particular, Random Forest (RF) and Naive Bayes (NB) classifiers typically exhibit high accuracy and low computational cost on low-complexity datasets. Random Forests, introduced by Breiman, represents an ensemble learning technique that creates multiple decision trees during training and produces an output based on the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This approach enhances predictive accuracy and helps control overfitting by averaging the results to reduce model variance. Although Random Forests effectively handle high-dimensional data and are resilient to overfitting, they can be computationally demanding and difficult to interpret due to their complexity. Random Forests are particularly effective for high-dimensional and structured datasets, excelling in classification and regression problems. However, their complexity presents challenges in terms of computational cost and interpretability. They are beneficial for classification and regression tasks involving structured data [9].

Naive Bayes classifiers are probabilistic models based on Bayes' theorem, assuming the independence of predictors [5]. Despite the simplicity of this assumption, in various real-world scenarios, Naive Bayes models often perform remarkably well due to their computational efficiency and ability to handle smaller datasets effectively. Their advantages include computational efficiency and strong performance with smaller datasets. However, the independence assumption may not always hold, potentially impacting accuracy. Naive Bayes classifiers are widely used in areas such as text classification, spam detection, and sentiment analysis [10].

2.2. Deep Learning Methods

As the complexity and scale of datasets continue to grow, traditional machine-learning methods gradually reveal their limitations. Deep learning methods, including CNNs and autoencoders, automatically extract high-level features from data through multi-layer neural networks, enabling them to capture complex patterns that are difficult for traditional approaches to identify.

Convolutional Neural Networks are a specialized form of deep learning architectures designed to handle data with grid-like structures, such as images [6]. CNNs utilize convolution layers to extract local patterns, pooling layers to reduce dimensionality, and fully connected layers to consolidate learned features. Their strength lies in capturing local patterns, making them highly effective for image and speech recognition tasks [11]. However, CNNs require large quantities of training data and computational resources, and the hyperparameter tuning process can be complex.

DOI: 10.54254/2755-2721/108/2025.20929

Autoencoders are unsupervised neural networks that aim to learn efficient representations of input data by training to minimize reconstruction error [12]. They include an encoder and a decoder. Autoencoders are particularly useful for feature extraction, dimensionality reduction, and denoising. They are prone to overfitting if not adequately regularized, necessitating careful design of network architecture and parameter tuning. Their applications include anomaly detection, image compression, and generative modeling [13].

2.3. Dimensionality Reduction Techniques

Dimensionality reduction is a key step in data preprocessing, particularly for high-dimensional datasets, as it simplifies models, mitigates overfitting, and enhances computational efficiency. Principal Component Analysis is a statistical approach that transforms a set of correlated variables into a new set of uncorrelated variables known as principal components, ranked by the variance they capture from the original data [14]. PCA can project the data into a lower-dimensional space to simplify models to reduce computational overhead and enhance model performance by mitigating noise and redundancy to reduce overfitting risk [15].

While PCA is restricted to linear transformations and therefore limited in capturing nonlinear patterns, autoencoders address this limitation by leveraging neural networks to model complex, nonlinear relationships. In contrast, autoencoders utilize neural networks to model nonlinear relationships and thus probably offer better performance for complex datasets [12].

2.4. Ensemble Methods and Deep Learning

In some tasks, combining different machine learning algorithms can lead to better performance. By integrating traditional machine learning methods with deep learning techniques, researchers have developed innovative models that balance computational efficiency and predictive accuracy. Random Forests demonstrate how ensemble techniques enhance model accuracy by combining multiple models' predictions to reduce variance and improve generalization [4]. This strategy takes advantage of group prediction, which tends to be more reliable than those from individual models.

CNNs' layered configuration allows them to extract increasingly abstract features at each stage, which is critical for identifying intricate patterns in data. This hierarchical feature extraction significantly contributes to their high accuracy in classification and recognition tasks [16].

3. Development and Evolution of the Field

The field of machine learning has undergone significant development and evolution over the past few decades. Early research predominantly focuses on statistical methods and basic algorithms such as linear regression, support vector machines, and decision trees. As the volume of data increased and computational power improved, the complexity of algorithms gradually rose. One notable milestone was the introduction of ensemble learning methods, such as Random Forests, which significantly enhanced the accuracy and stability of models.

In recent years, deep learning methods have revolutionized the landscape of machine learning research. Convolutional Neural Networks (CNN) have made groundbreaking advances in image processing, while Autoencoders have provided new perspectives for unsupervised learning and feature learning. Although deep learning methods excel in handling complex datasets, they require substantial computational resources and longer training times, limiting their widespread adoption, especially in resource-constrained environments.

DOI: 10.54254/2755-2721/108/2025.20929

4. Challenges and Applications

Traditional machine learning methods, such as Random Forest and Naive Bayes, perform well on low-complexity datasets, offering high accuracy and computational efficiency. These methods are widely applied in fields like medical diagnosis, financial forecasting, and e-commerce recommendation systems. For example, Random Forest can be effectively used for customer behavior analysis and risk prediction [4]. Compared to deep learning methods, traditional machine learning methods generally offer higher computational efficiency, making them suitable for real-time systems that require quick feedback. The Naive Bayes model is also widely used in tasks like text classification and spam filtering [17].

Deep learning methods (such as Convolutional Neural Networks (CNN) and Autoencoders) have made significant progress in high-complexity tasks such as image processing, speech recognition, and natural language processing. Convolutional Neural Networks (CNNs), for instance, have found extensive applications in fields such as medical image analysis, autonomous vehicles, and video surveillance, where their hierarchical learning enables them to capture intricate data patterns [16]. Deep learning methods automatically learn data features, eliminating the need for manual feature extraction. This is particularly effective in tasks like image recognition and speech recognition.

5. Discussion

Traditional machine learning methods (e.g., Random Forest, Naive Bayes) excel in handling low-complexity datasets. They offer many advantages like high accuracy and good computational efficiency and can provide quick and effective predictions in many practical applications. Random Forest, introduced by Breiman, is particularly effective in solving high-dimensional data problems [18]. Deep learning methods (such as CNN and Autoencoders) perform exceptionally well in high-complexity tasks like image and speech processing. However, they come with high computational costs and long training times. Furthermore, deep learning models often lack interpretability, which can be a barrier to their application in domains such as medical diagnostics [3, 16]. When selecting algorithms, we should consider the complexity of the dataset, the task requirements, and available computational resources. For simple classification tasks, traditional machine learning methods are often preferable due to their efficiency and reliability. Conversely, deep learning is better suited for complex tasks like image classification or speech recognition, where the hierarchical feature extraction capabilities of CNNs can be leveraged effectively. Future research can explore hybrid models to combine the advantages of both methods in different tasks [19]. Data preprocessing is critical to model performance, Whether traditional machine learning or deep learning. Poor-quality data, whether incomplete or noisy, can significantly degrade model effectiveness. Future research should prioritize optimizing data preprocessing and cleaning techniques to improve model generalization and performance, particularly in scenarios involving limited or unstructured datasets [20].

6. Conclusion

This paper discusses the performance of traditional machine learning and deep learning methods on datasets of varying complexity. Traditional machine learning methods perform well on low-complexity tasks due to their computational efficiency and accuracy. In contrast, deep learning methods exhibit stronger learning abilities for high-complexity tasks. Nevertheless, deep learning methods have higher computational costs and training times. We also found that the choice of algorithm should be based on the specific task requirements and the characteristics of the dataset. Future research could focus on hybrid and optimized algorithms, exploring how to combine traditional machine learning and deep learning methods to balance computational efficiency and

DOI: 10.54254/2755-2721/108/2025.20929

accuracy. Moreover, the continued improvement of computational capabilities may further expand the applicability of deep learning methods, particularly in large-scale and high-dimensional datasets. Addressing the interpretability issue of deep learning models is another critical area for future exploration, as enhanced transparency could foster greater trust and wider adoption across sensitive domains like healthcare. Additionally, data preprocessing and feature engineering remain pivotal to model performance. Furthermore, data preprocessing and feature engineering remain crucial factors influencing model performance. Future studies should explore how effective preprocessing can enhance model generalization. This is particularly important in scenarios with limited data, which warrants further attention. These advancements are essential for overcoming the challenges of modern machine learning and unlocking its full potential.

References

[1] *Mitchell, T. M. (1997). Machine learning. McGraw-Hill.*
[2] *Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.*
[3] *Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.*
[4] *Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324*
[5] *McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization (pp. 41–48).*
[6] *LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791*
[7] *Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755*
[8] *Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160(1), 3–24.*
[9] *Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.*
[10] *Zhang, H. (2004). The optimality of naive Bayes. AAAI 2004 Conference on Artificial Intelligence*
[11] *Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (Vol. 25, pp. 1097–1105).*
[12] *Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507. https://doi.org/10.1126/science.1127647*
[13] *Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11, 3371–3408.*
[14] *Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.*
[15] *Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9*
[16] *LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539*
[17] *Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 29–36. https://doi.org/10.1145/860435.860444*
[18] *He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90*
[19] *Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 1251-1258. https://doi.org/10.1109/CVPR.2017.190*
[20] *Chung, J., & Zohdy, M. A. (2020). Data preprocessing for machine learning: Improving predictive accuracy. IEEE Access, 8, 11322-11337. https://doi.org/10.1109/ACCESS.2020.2972421*