Performance Comparison of Convolutional Neural Network and Long Short-Term Memory for the Classification of Handwritten Digits

Toyobo, Oluwatobi Joel[1]; Olabiyisi, Stephen Olatunde[2]; Ismaila, Wasiu Oladimeji[3]; Oyedele, Adebayo Olalere[4]

[1,2,3]Department of Computer Science, Ladoke Akintola University of Technology, Ogbomoso, Oyo state Nigeria

[4]Department of Information systems, University of Portsmouth, United Kingdom

DOI: https://doi.org/10.51584/IJRIAS.2025.10060041

Received: 25 April 2025; Accepted: 13 May 2025; Published: 04 July 2025

ABSTRACT

Handwritten digit recognition, a task in computer vision, is critical for applications such as postal automation, banking, and digitization of forms. Traditional approaches have leveraged statistical models, but the rise of deep learning, particularly Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM), has revolutionized the field. However, a comprehensive performance comparison of CNN and LSTM architectures in the context of handwritten digit classification remains underexplored. This study aimed to address this gap by evaluating and comparing CNN and LSTM models for the classification of handwritten digits. Two machine learning models – Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) were trained with the preprocessed handwritten digit dataset. The CNN model was designed with multiple convolutional and pooling layers, along with dropout for regularization. The LSTM model was designed with LSTM layers to capture sequential patterns in the data, followed by a dense layer for classification. The models were implemented in python, evaluated and compared based on accuracy, precision, recall and F1-score. The evaluation and comparison results indicate that CNN achieved 99.31% accuracy, 99.0% precision, 99.0% recall, and a 99.0% F1-score, while LSTM achieved 98.90% accuracy, 99.0% precision, 99.0% recall, and a 99.0% F1-score. The results demonstrated that CNN outperformed LSTM in terms of accuracy and misclassification errors, making it the optimal choice for image-based handwritten digit recognition. This finding underscores the efficiency of CNN in addressing challenges related to digit recognition, contributing to the advancement of automated digit classification systems and improving the accuracy of image-based classification tasks.

Keyword: Performance comparison, Handwritten digits, Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM).

INTRODUCTION

The task of handwriting recognition is a cornerstone challenge in both computer vision and image recognition realms. Crafting algorithms and models that can accurately decipher and classify handwritten characters or symbols remains an ongoing hurdle in today's technology landscape. The versatility of this technology spans a broad spectrum, encompassing critical functions such as the digitization of historical documents and empowering smart devices to interpret and understand user-generated input (Ahlawat *et al.*, 2020).

Central to this technological advancement is the profound impact of Convolutional Neural Networks (CNNs) (Zebari *et al.*, 2022) a revolutionary technique inspired by the human brain's visual recognition process. Similar to how humans teach children to recognize objects by exposing them to numerous images, CNNs operate on the premise of automatically extracting hierarchical features from raw pixel data. This unique ability has positioned CNNs as a linchpin in the realm of deep learning, especially in tasks that rely on visual inputs, such as images or handwritten text.

Deep learning, a subset of machine learning, has demonstrated exceptional performance in image recognition tasks. Convolutional Neural Networks (CNNs), in particular, have become the de facto standard for image classification due to their ability to automatically learn hierarchical features (Krizhevsky *et al.,* 2012 and Adeniran *et al.,* 2025). Several studies have explored the application of deep learning to handwritten digit recognition. For instance, (LeCun *et al.,* 1998) introduced the LeNet-5 architecture, a pioneering CNN model that achieved remarkable accuracy on the MNIST dataset, a benchmark for handwritten digit classification. Subsequent research has focused on improving CNN architectures, exploring different hyperparameters, and investigating the impact of various data augmentation techniques (Simard *et al.,* 2003).

Some deep learning techniques have gained attention in recent years, and they perform noticeably better than shallow neural networks in the area of picture or digits categorization (Le Roux *et al.,* 2008 and Olojede *et al.,* 2025). Compared to shallow neural networks, deep learning techniques are made up of several layers that progressively extract increasingly complex and invariant properties from the raw input pictures (Le Roux and Bengio 2010). Deep neural network training has been made feasible by the proliferation of large-scale data sets and more potent computer environments, which has resulted in the widespread use of deep learning techniques.

In the quest to contribute to knowledge, this study aims to evaluate and compare the handwritten digits models (CNN and LSTM) to classify them base on their predefined categories with the objectives to;

1. acquire the dataset containing labeled images of handwritten digits and preprocess the dataset to ensure consistency and improve model performance.
2. implement and Train Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) models for the classification of the acquire preprocessed handwritten dataset using python.
3. compare the models' performance using classification accuracy, precision, Recall, and F1-score.

Related Work

Handwriting recognition has already achieved impressive results using shallow networks (Choudhary *et al.,* 2015). Many papers have been published with research detailing new techniques for the classification of handwritten numerals, characters and words. (Pham et al., 2014) applied a regularization method of dropout to improve the performance of recurrent neural networks (RNNs) in recognizing unconstrained handwriting. The author reported improvement in RNN performance with significant reduction in the character error rate (CER) and word error rate (WER).

According to Rastegari *et al.,* 2016, image classification system based on a structure of a Convolutional Neural Network (CNN). The training was performed such that a balanced number of face images and non-face images were used for training by deriving additional face images from the face images data. The image classification system employs the biscale Convolutional Neural Network with 120 trained data and the auto-stage training achieves 81.6% detection rate with only six false positives on Face Detection Data Set and Benchmark (FDDB), where the current state of the art achieves about 80% detection rate with 50 false positives.

Ahalawat *et al.,* 2020 presented that traditional systems of handwriting recognition have relied on handcrafted features and a large amount of prior knowledge. Training an Optical character recognition (OCR) system based on these prerequisites is a challenging task. Research in handwriting recognition field is focused around deep learning techniques and has achieved breakthrough performance in the last few years. Still, the rapid growth in the amount of handwritten data and the availability of massive processing power demands improvement in Recognition accuracy and deserves further investigation.

Nasim *et al.,* 2020 presented that convolutional neural networks (CNNs) are very effective in perceiving the structure of handwritten characters/words in ways that help in automatic extraction of distinct features and make CNN the most suitable approach for solving handwriting recognition problems.

Vidhale *et al.,* 2021 focused on creating an efficient algorithm to recognize handwritten digits from scanned user inputs. We compared different algorithms, adjusting hidden layers and epochs, to see which one yielded the highest accuracy in classifying these digits.

Xiao et al., 2022 research highlighted the complexity and limitations of traditional algorithms for recognizing handwritten digits, particularly with large databases. Our study explored an alternative approach using extension engineering, constructing classical and extensional domains for handwritten digits. By calculating correlation degrees between input digits and standard digits, we effectively classified and recognized handwritten digits, demonstrating the potential of extension engineering in this field.

METHODOLOGY

In the process of comparing the performance evaluation of Convolutional Neural Network and long short-term memory for the classification of handwriting digits, the following steps were involved;

i. Data Acquisition: The Handwritten digits dataset was acquired from Modified National Institute of Standard and Technology (MNIST) which consist of 70,000 grayscale images of handwritten digits (0-9), each represented as a 28 x 28-pixel grid with intensity values ranging from 0 to 255. The MNIST dataset was utilized for image classification, serving as the primary dataset for model training and evaluation. MNIST, a well-known benchmark dataset in computer vision and machine learning, has 70,000 handwritten digits from 0 to 9 in grayscale images

ii. Data Pre-processing: The Handwritten digits dataset involved reshaping each image to a 28x28x1 format for compatibility with CNNs, normalizing pixel values to a 0-1 range, and applying one-hot encoding to class labels for multi-class classification. Additionally, data augmentation techniques such as random rotations, shifts, and zooms were implemented to enhance the model's generalization ability. Preprocessing the MNIST dataset is a vital step in preparing the data for optimal model training. The first duty is to ensure that all photographs are in the correct format, specifically resized to 28x28 pixels if they are not already. The MNIST dataset supplies images in this format, however in circumstances when data may be sourced from elsewhere, scaling maintains consistency in input size across the entire dataset. As shown in Figure 3.3, this phase is crucial, as most neural networks require set input dimensions, and irregular image sizes could lead to problems during model training.

The preprocessing pipeline for the MNIST dataset involves several key steps: reshaping, normalization, and one-hot encoding. These steps ensure that the data is in the correct format and scale, enabling the deep learning models to learn efficiently and perform optimally.

a) Reshaping: Each image was reshaped from its original 28x28 pixel format to a 28x28x1 format to include a single channel (grayscale) and make it compatible with the CNN architecture.

b) Normalization: The pixel values were scaled from the original range of 0-255 to a normalized range of 0-1, helping to stabilize and speed up the training process.

c) One-Hot Encoding: The class labels (digits 0-9) were one-hot encoded to represent the classification output in a format suitable for multi-class classification.

d) Data Augmentation: To improve the models generalization ability, data augmentation techniques such as random rotations, shifts, and zooms were applied to the training dataset.

iii. Model Development: The dataset was trained and tested on the selected two algorithms Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) Algorithm.

iv. Model Evaluation: Evaluate the performance of the models using classification accuracy, precision, recall and F1- score metrics. Both models were assessed based on key performance metrics, including classification accuracy, precision, recall, F1-score, and overall efficiency in handling the classification task. These metrics were selected to comprehensively evaluate the strengths and limitations of each model, offering insights into their practical applicability for image-based tasks. The CNN model, known for its ability to capture spatial hierarchies of features, is contrasted with the LSTM model, which excels in sequential data processing.

RESULTS AND DISCUSSIONS

As shown in table 1, the comparative analysis between the CNN and LSTM models highlights their respective strengths and overall performance. Both models demonstrated exceptional results, with classification accuracies of 99.31% and 98.90% for CNN and LSTM, respectively, while the difference in accuracy appears marginal, it underscores CNN's superior ability to handle image-based tasks effectively. CNN’s optimization for recognizing spatial hierarchies allows it to capture intricate patterns in the MNIST dataset, contributing to its slight performance edge. Furthermore, both models achieved identical precision, recall, and F1-scores, each at 99.0%, reflecting their remarkable balance in correctly identifying and predicting digit classes. This consistency demonstrates that both architectures are highly capable of generalizing across the dataset. However, the CNN model's overall efficiency and suitability for static image processing make it the preferred choice for such tasks.

In addition to its higher accuracy, CNN offers computational advantages that further solidify its position as the more efficient model for image classification. Its ability to leverage convolutional layers for feature extraction ensures faster and more accurate recognition of patterns compared to the sequential nature of LSTM. On the other hand, the LSTM model, while effective, is better suited for sequential or temporal data due to its recurrent architecture. This design can inadvertently lead to longer processing time and slightly reduced accuracy when applied to static image data MNIST. The CNN model's structure allows it to avoid these pitfalls, providing a better fit for tasks where spatial relationships are critical. These differences highlight the importance of selecting model architectures based on the specific characteristics of the data being analyzed. Despite the LSTM models slight drawbacks in this context, its competitive performance indicates its versatility across varied datasets.

Overall, the comparison demonstrates the robust capabilities of both CNN and LSTM models for handwritten digit recognition, with CNN emerging as the optimal choice for the given dataset. While LSTM maintains a strong performance with metrics nearly identical to CNN's, its architecture is inherently limited for tasks requiring spatial feature analysis. CNN, conversely, excels in this regard and benefits from reduced computational overhead and enhanced accuracy. The results highlight the distinct advantages of leveraging CNN for image-based machine learning tasks while acknowledging LSTM's potential in handling sequential data. Both models exhibit a shared strength in their ability to generalize across classes, achieving identical precision, recall, and F1-scores. These findings emphasize the importance of aligning model architecture with dataset properties to achieve optimal results. Future research could explore hybrid architectures that combine the spatial processing power of CNN with the sequential data handling of LSTM to address diverse classification challenges.

Table 1. Comparative Analysis

Metrics	CNN	LSTM
Classification Accuracy	99.3%	98.90%
Precision	99.0%	99.0%
Recall	99.0%	99.0%
F1-Score	99.0%	99.0%

Although the precision, recall, and F1-score for both models were equal 99.0%, the slight edge in accuracy highlights CNN's effectiveness in reducing misclassifications. Additionally, CNN’s computational efficiency and faster inference times make it a practical choice for real-time or large-scale applications. LSTM’s marginally lower accuracy and longer processing times indicate its limitations in handling image data. However, this should not undermine the model's potential in its intended domain of sequential or temporal data processing. In this case, CNN's architecture provided a more reliable and efficient solution for the task at hand.

Overall, the results reinforce the importance of aligning the model's architecture with the data's intrinsic properties. The success of CNN on the MNIST dataset serves as a testament to the effectiveness of convolutional networks in image classification tasks. Meanwhile, the LSTM model, despite being less effective here, remains a powerful tool for problems involving sequential data. These findings highlight the need for careful consideration of data characteristics during model selection to optimize performance and efficiency.

As shown in Figures 3 and 4, the analysis of the results from the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models highlights their unique strengths in handling the MNIST dataset of handwritten digit images.

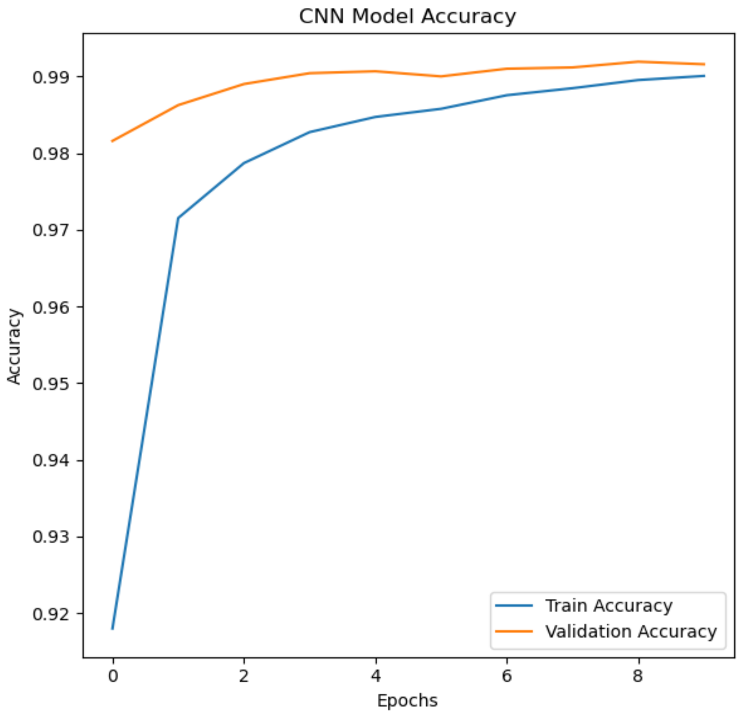

Figure 3. CNN Model Accuracy

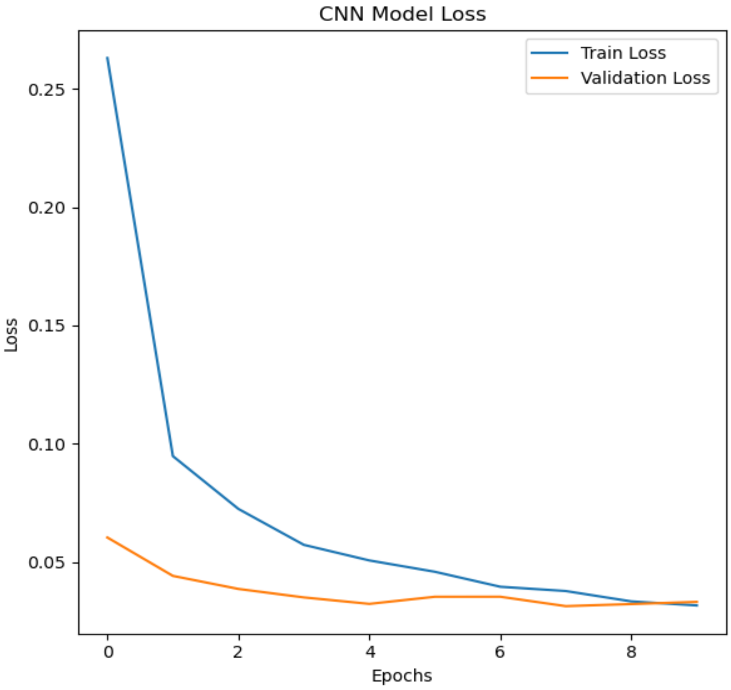

Figure 4. CNN Model Loss

The CNN model excelled with an overall accuracy of 99.31% and an F1 score of 99.0%, demonstrating its ability to effectively process spatial patterns and static image features. As shown in its confusion matrix, the

CNN consistently achieved high precision and recall across all digit classes, minimizing misclassification errors. Notable strengths include its accurate recognition of digits with complex shapes such as "8" and "9," which are often challenging in other models. Misclassification rates were minimal, with only isolated errors, such as confusing "4" with "9," reflecting its robust feature extraction capabilities. The CNN's performance demonstrates its computational efficiency and architecture design, which are highly suitable for tasks involving static spatial data. This exceptional accuracy supports the use of CNN in digit recognition tasks, where image-specific patterns are crucial.

As shown in Figure 5 and 6, the LSTM model, designed primarily for sequential data, also performed commendably on the MNIST dataset, achieving an accuracy of 98.90% and an F1-score of 99.0%. Despite being inherently optimized for time-series data, the LSTM adapted well to image data, though it exhibited slightly higher misclassification rates compared to CNN.

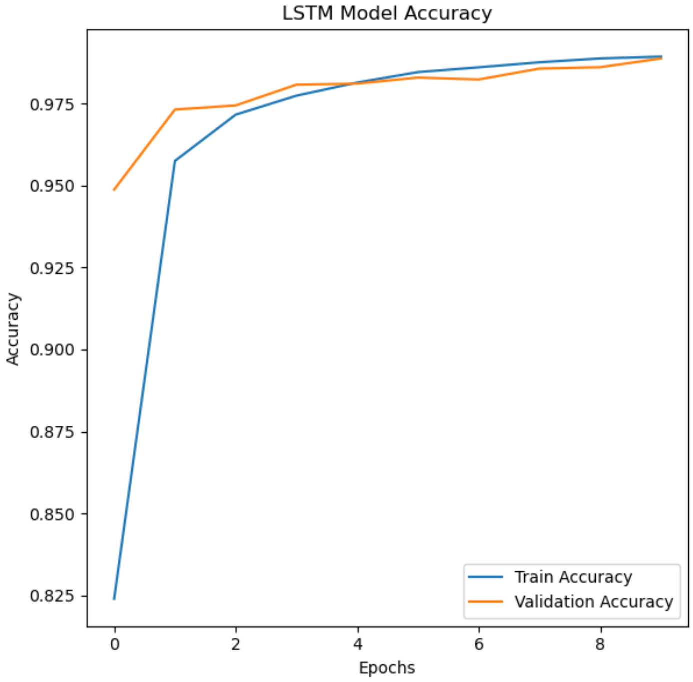

Figure 5. LSTM Model Accuracy

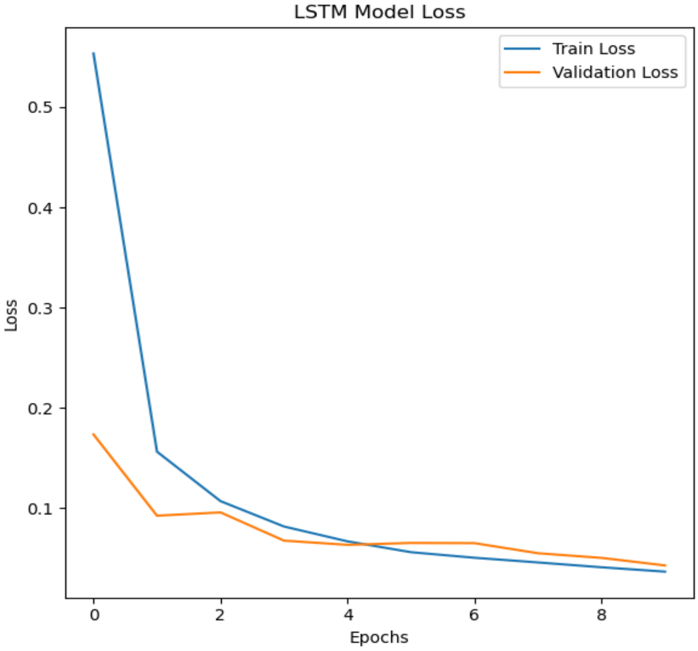

Figure 6. LSTM Model Loss

CONCLUSION

This research has evaluated and compared the performance of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks in handwritten digit recognition tasks using the MNIST dataset. The results demonstrated that CNN outperforms LSTM in tasks requiring spatial feature extraction, achieving higher accuracy, precision, and efficiency. The CNN model's ability to process static image data is evident in its robust performance across all digit classes, as observed in its confusion matrix and overall metrics. In contrast, while LSTM achieved respectable accuracy and F1 scores, it exhibited slightly higher misclassification rates, primarily due to its sequential architecture, which is less suited for static image data.

The study further revealed that the CNN model's architecture is specifically designed to capture intricate spatial patterns, making it more effective for image-based tasks. Its ability to minimize misclassification rates, especially in digits with complex or ambiguous shapes, underscores its superiority in tasks like handwritten digit recognition. On the other hand, LSTM's adaptability and learning dynamics highlight its versatility, though it remains less ideal for static spatial data. These findings emphasize the critical importance of selecting the appropriate model architecture based on the characteristics of the dataset and the nature of the task. Overall, the results provide clear evidence that CNN is the optimal choice for static spatial data tasks like handwritten digit recognition, while LSTM is better suited for sequential.

REFERENCES

1. Adeniran, Rachel Ihunanya., Olabiyisi, Stephen Olatunde., Ismaila, Wasiu Oladimeji., Oyedele, Adebayo Olalere., Olagbemiro, Catherine Olatorera (2025), Performance Evaluation of Long Short-Term Memory and Autoregressive Integrated Moving Average Time Series Models for Stock Price Prediction, International Journal of Latest Technology in Engineering, Management & Applied Science 14(2), 192-199. https://doi.org/10.51583/IJLTEMAS.2025.1402003
2. Ahlawat, S.; Rishi, R. A genetic algorithm based feature selection for handwritten digit recognition. Recent Pat. Comput. Sci. 2019, 12, 304–316.
3. Ahlawat S., and Choudhary A. (2020). Hybrid CNN-SVM Classifier for Handwritten Digit Recognition. Procedia Computer Science, 167, 2554–2560. https://doi.org/10.1016/j.procs.2020.03.309.
4. Alvear-Sandoval R.; Figueiras-Vidal, A. On building ensembles of stacked denoising auto-encoding classifiers and their further improvement. Inf. Fusion 2018, 39, 41–52.
5. Bengio Y. (2009). Learning deep architectures for ai. Foundations and Trendsë in Machine Learning, 2(1), 1–127. Doi:10.1561/2200000006. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al. (2007). Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems, 19, 153.
6. Bengio Y, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, 2009.
7. Bengio Y, Simard P, and Frasconi P, "Learning long-term dependencies with gradient descent is difficult," IEEE transactions on neural networks, vol. 5, no. 2, pp. 157–166, 1994.
8. Choudhary A.; Ahlawat S.; Rishi R. A neural approach to cursive handwritten character recognition using features extracted from binarization technique. Complex Syst. Model. Control Intell. Soft Comput. 2015, 319, 745–771.
9. Demir C.; Alpaydin E. Cost-conscious classifier ensembles. Pattern Recognit. Lett. 2005, 26, 2206–2214.
10. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. 1980, 36, 193–202.
11. Fukushima K and Miyake S, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in Competition and cooperation in neural nets: Springer, 1982, pp. 267-285.
12. Goltsev A and Gritsenko V, "Investigation of efficient features for image recognition by neural networks," Neural Networks, vol. 28, pp. 15–23, 2012.

13. Hubel D and Wiesel T, "Aberrant visual projections in the Siamese cat," The Journal of physiology, vol. 218, no. 1, pp. 33- 62, 1971.
14. Huimin Xiao, Chen Liu, (2022), "Handwriting Digit Recognition Based on Extension Engineering", IEEE, 2915-2918.
15. Krizhevsky A.; Sutskever I.; Hinton G.E. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 2012, 25, 1097–1105.
16. LeCun Y, "LeNet-5, convolutional neural networks," URL: http://yann. lecun. com/exdb/lenet, vol. 20, 2015.
17. LeCun Y, Bottou L, Bengio Y, and Haffner P, "Gradientbased learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
18. LeCun Y, Bengio Y, and Hinton G, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.
19. Le Roux N and Bengio Y, "Deep belief networks are compact universal approximators," Neural Computation, vol. 22, no. 8, pp. 2192–2207, 2010.
20. Olojede, Ojo Abraham, Stephen Olatunde Olabiyisi, Oluwaseun. O Alo (2025), Performance Evaluation of Some Machine Learning Models for Music Genre Classification. International Journal of Latest Technology in Engineering Management & Applied Science, 14(2), 18-24. https://doi.org/10.51583/IJLTEMAS.2025.1402003
21. Savita Ahlawat, Amit Choudhary, (2020), "Improved Handwritten Digit Recognition Using Convolutional Neural Networks (CNN)", IEEE sensor Journal, 01-18.
22. Xiang L.; Li Y.; Hao W.; Yang P.; Shen X. Reversible natural language watermarking using synonym substitution and arithmetic coding. Comput. Mater. Contin. 2018, 55, 541–559.
23. Zebari D. A., Haron H., Sulaiman D. M., Yusoff, Y., and Othman, M. N. M. (2022, December). CNN-based Deep Transfer Learning Approach for Detecting Breast Cancer in Mammogram Images. In 2022 IEEE 10th Conference on Systems, Process & Control (ICSPC) (pp. 256-261). IEEE.
24. Zeng D.; Dai Y.; Li F.; Sherratt R.S.; Wang J. Adversarial learning for distant supervised relation extraction. Comput. Mater. Contin. 2018.