Handwritten Math Symbol Recognition Based on Multiple Machine Learning Algorithms: A Comparative Study

Zhihao Xu

SILC Business School, Shanghai University, Shanghai, 201800, China
1586199252@shu.edu.cn

Abstract. The primary focus of this research paper is addressing the difficulty of identifying handwritten mathematical symbols, which holds significant importance in diverse fields including education, scientific research, and data analysis. The recognition of these symbols is challenging due to their diverse appearance and inconsistencies in individual handwriting styles. Previous research has mainly focused on recognizing handwritten numerals, leaving a research gap in recognizing a wider range of math symbols. To bridge this gap, this study proposes a machine learning approach using the random forest algorithm. The approach utilizes a meticulously collected dataset from Kaggle. The dataset undergoes preprocessing steps including grayscale conversion, average pooling, and resizing to enhance recognition accuracy. The study implemented and evaluated K-Nearest Neighbors (KNN), decision tree, and random forest. The results demonstrate that the random forest model outperforms the other models, achieving a macro average accuracy of 0.99321 and a weighted average accuracy of 0.99412. The ensemble nature of the random forest algorithm contributes to its superior performance in handwritten math symbol recognition. The results support the hypothesis that the random forest model is highly effective in recognizing handwritten math symbols. The findings emphasize the significance of recognizing math symbols in developing intelligent systems for mathematical analysis and understanding and its vast potential for future application.

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_67

Keywords: handwritten math symbol recognition, random forest, machine learning

1 Introduction

Mathematical symbols play a pivotal role in conveying mathematical concepts, operations, and relationships within mathematical expressions. These symbols encompass a wide range of elements, including arithmetic operators' equality signs parentheses, square roots, integration symbols etc. Each symbol holds a specific meaning and contributes to the overall structure and understanding of mathematical expressions. Accurate recognition of mathematical symbols holds paramount importance across various applications. In the field of education, intelligent tutoring systems heavily rely on symbol recognition to provide personalized guidance and feedback to students. In scientific research, mathematical symbol recognition aids in analyzing and interpreting complex mathematical equations and formulas, enabling researchers to make accurate conclusions and discoveries. In data analysis, recognizing mathematical symbols is crucial for processing mathematical expressions and performing mathematical operations, facilitating advanced calculations and modeling.

The recognition of mathematical symbols presents unique challenges due to their diverse appearance and the inconsistencies observed in individual handwriting styles. Handwritten mathematical symbols can exhibit variations in size, shape, slant, and style, making it challenging to develop accurate recognition algorithms. Moreover, symbols with similar shapes or recurring characters. Recognizing characters and symbols can be challenging for Optical Character Recognition (OCR) due to the diverse array of writing styles and the presence of different symbols and recurring characters.

Due to the extensive utilization of handwriting and mathematical content in diverse human interactions, the identification of handwritten mathematical symbols has become highly significant and applicable in practical contexts. Xie et al. introduced an improved convolutional neural network, inspired by AlexNet, specifically designed for recognizing handwritten digits [1]. Their approach incorporated the Inceptionresnet module, replacing Conv3 and Conv4 layers, to improve feature extraction. Additionally, they employed Batch Normalization (BN) for faster convergence and prevention of overfitting. By reducing the number of convolutional kernels, they achieved accelerated training. Experimental results on the MNIST dataset showcased a remarkable detection accuracy of 0.9966, affirming the effectiveness of their algorithm. Deng and Zhang developed an improved LeNet-5 model based on convolutional neural networks which has a simpler structure and thus achieves higher classification efficiency in handwritten digit recognition [2]. Their model with the new structure has the advantages of good robustness and strong generalization ability, and lower false recognition rate. In addition, many researchers have utilized machine learning models such as Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks for handwritten mathematical symbol recognition, achieving notable results [3-5].

Previous research efforts have primarily focused on recognizing handwritten numerals, resulting in significant advancements in the field. However, the recognition of math symbols demands a broader scope, encompassing symbols such as arithmetic operators, equality signs, parentheses, square roots, and integration symbols. Despite some related work addressing a subset of math symbols, research that encompasses a wider range of symbols still deserves more attention.

This study aims at bridging the existing research gap in handwritten math symbol recognition. To address this challenge, the random forest algorithm has been utilized in this paper, which is a powerful machine learning technique known for its ability to handle complex and non-linear data. The approach builds upon a meticulously collected dataset sourced from Kaggle, which encompasses a diverse range of handwritten math symbols, incorporating variations in writing styles. Preliminary findings from this study have shown promising accuracy rates, with the proposed model achieving an impressive accuracy of 99% on the dataset. By specifically addressing handwritten math symbol recognition, this research not only demonstrates the feasibility and effectiveness of the approach but also emphasizes the significance of recognizing math symbols as a crucial component in the development of intelligent systems for mathematical analysis and understanding.

2 Method

2.1 Dataset Preparation

The Handwritten Mathematical Symbols dataset used in this study was sourced from Kaggle [6], which includes a total of 375,974 instances divided into 82 categories. Each instance represents a grayscale image of dimensions 45 pixels by 45 pixels. The dataset comprises fundamental Greek alphabet characters, English alphanumeric symbols, mathematical operators, set operators, a selection of essential predefined mathematical functions, and various mathematical symbols. Visual representations of the raw images are depicted in Fig. 1, wherein the rightmost images specifically illustrate the existential symbol.

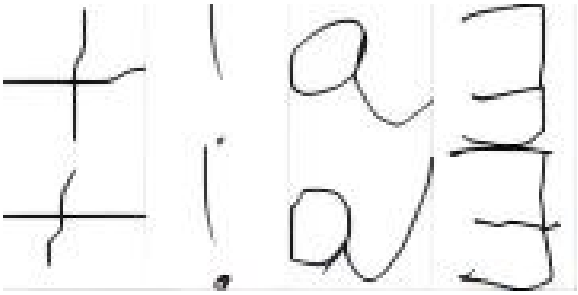

Fig. 1. Visualizations of raw images [6].

In terms of the data preprocessing stage, the grayscale values of the images were obtained, and a 3x3 average pooling operation was applied to transform the images into a resolution of 15 pixels x 15 pixels. This resizing technique aims to reduce the computational complexity and memory requirements while preserving essential features of the handwritten mathematical symbols. By averaging the pixel values within each 3x3 neighborhood, the pooling operation condenses the information and effectively captures the dominant characteristics of the symbols [7]. Fig. 2 is some visualizations of symbols after average pooling. The resulting smaller-sized images facilitate efficient processing and analysis, allowing for improved model training and recognition accuracy. This preprocessing step plays a crucial role in enhancing the overall performance of the machine learning algorithms applied to this dataset. The data were flattened into one dimension in the end.

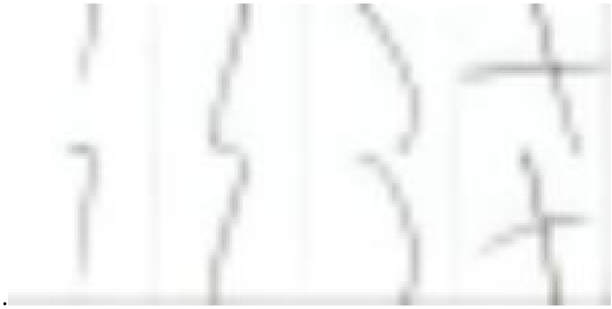

Fig. 2. Visualizations of average-pooled images (Photo/Picture credit: Original).

2.2 Machine Learning Models

KNN. K-Nearest Neighbors (KNN) is a widely used machine learning and pattern recognition algorithm utilized for supervised classification tasks. It is a non-parametric method that relies on the concept of proximity. In KNN, the classification of an unlabeled data point is determined by considering the majority vote of its K closest neighbors in the feature space. This algorithm is flexible and capable of handling both binary and multi-class classification problems. It does not assume any underlying probability distribution of the data, making it robust to varying data distributions. KNN has remained a popular and intuitive classification algorithm for decades, known for its ease of implementation and interpretability. It has applications in various fields, including image recognition, recommender systems, and anomaly detection, and has shown promising results in decision-making processes due to its concept of proximity [8, 9].

Decision Tree. The Decision Tree is a well-known and extensively applied supervised learning algorithm in the fields of machine learning and data mining. It creates a tree-like structure that represents decisions and their potential outcomes, enabling the classification and regression of data by evaluating logical conditions in a sequential manner. The process of building a decision tree involves iteratively dividing the data using chosen attributes and their respective splitting criteria. At each internal node, a feature is selected based on certain criteria to split the data into two or more subsets. This process is repeated recursively for each subset, creating branches and leaf nodes in the tree structure [10]. The leaf nodes correspond to the final decision or prediction. Decision trees can handle nonlinear relationships and interactions among features without requiring explicit feature engineering. On the other side, decision trees may suffer from overfitting [11].

Random forest. Random Forest is a highly flexible and potent machine learning algorithm that builds an ensemble of decision trees by generating multiple random subsets from the original training data [12]. Each decision tree within the forest is trained on a distinct subset of the data and employs a random subset of features for each split. This randomness introduces diversity and helps to reduce the correlation between individual trees. This ensemble method has several advantages over traditional decision trees, including: 1) Robustness to outliers and noisy data. 2) Ability to handle both categorical and numerical features. 3) Provision of feature importance measures. 4) Ability to estimate out-of-bag (OOB) error. These advantages make random forests a popular choice for a variety of machine learning tasks, including classification, regression, and anomaly detection.

Implementation details. In implementation of this study, the KNN model is configured with a k-value of 10, while the decision tree model utilizes a maximum depth of 120 and a minimum number of samples required to split a node set to 10. Additionally, the random forest model employs 50 decision trees with a maximum depth of 120. The training and testing data were split in a ratio of 7:3. To evaluate the performance, this study used weighted average accuracy, macro average accuracy, recall rate, and F1-score as the evaluation metrics.

3 Results and Discussion

The findings presented in Table 1 demonstrate that the Random Forest model exhibits superior performance across all evaluation metrics when compared to both the KNN and Decision Tree models. The macro average accuracy, weighted average accuracy, F1-score, and recall rate are employed to examine their suitability for accurate classification tasks. According to the four evaluation indexes, the Decision Tree model exhibits satisfactory performance, surpassing the KNN model across all metrics. The Random Forest model surpassed the performance of both the KNN and Decision Tree models, achieving a maximum macro average accuracy of 0.99321 and a weighted average accuracy of 0.99412.

Table 1. The performance of different models in the handwritten math symbols dataset.

	Performance

Model	Macro Average Accuracy	Weighted Average Accuracy	F1-score	Recall-rate
KNN	0.79835	0.84562	0.69838	0.65919
Decision Tree	0.83280	0.91927	0.82610	0.82168
Random Forest	0.99321	0.99412	0.98264	0.97776

The Random Forest model outperforms other models in handwritten math symbol recognition possibly due to several key factors. Firstly, as an ensemble learning method, Random Forest combines multiple decision trees to make predictions, reducing overfitting and improving generalization. This enables the model to effectively capture complex relationships and patterns in handwritten symbols, resulting in higher accuracy. Furthermore, Random Forest handles imbalanced classes by adjusting weights or utilizing sampling techniques, ensuring proper classification across all symbol categories. In the future, more advanced models such as convolutional neural networks can be considered to further improve the performance of the model due to their satisfactory performance on other tasks [13, 14].

4 Conclusion

In conclusion, this research primarily concentrated on employing machine learning algorithms for the recognition of handwritten mathematical symbols. The outcomes illustrated the efficacy of the approach, as the random forest model achieved a remarkable accuracy rate of 99.3% on the dataset. The comparison with other models, namely K-Nearest Neighbors and Decision Tree, showed that the random forest model outperformed them in terms of accuracy, highlighting its suitability for this task. The random forest model's ensemble learning approach proved beneficial in capturing complex relationships and patterns within the handwritten symbols, leading to improved recognition accuracy. The findings emphasize the significance of accurately recognizing math symbols in the development of intelligent systems for mathematical analysis and understanding. This research has practical implications in education, scientific research, and data analysis, where symbol recognition plays a vital role.

Future research directions could focus on exploring enhanced preprocessing techniques, feature engineering, and deep learning approaches to further improve recognition accuracy. Additionally, collecting larger and more diverse datasets would contribute to the generalizability and robustness of recognition systems. These advancements hold the potential to enhance the performance of symbol recognition systems and enable their application in real-world scenarios.

References

1. Xie, D., Li, L., Miao, C.: Handwritten numeral recognition based on improved AlexNet convolutional neural network. J. Hebei Univ. Eng. (Nat. Sci. Ed.) 38(4), 102–106 (2021).
2. Deng, C., Zhang, J.: Handwritten digit recognition based on improved LeNet-5 model. Information and Communication, (01):109-112 (2018).
3. Zong, C., Zhang, Y., Shi, D.: Research on handwritten digit recognition based on CNN under PyTorch. Computer and Digital Engineering, 49(06):1107-1112 (2021).
4. Jiang, R., Gulijiamali, M., Ailina: Handwritten digit recognition based on Long Short-Term Memory Neural Network. J. Comput. Technol. Dev. 30(2), 94–97 (2020).
5. Zhang, T., Mouchère, H., Viard-Gaudin, C.: A tree-BLSTM-based recognition system for online handwritten mathematical expressions. Neural Comput & Applic 32, 4689–4708 (2020).
6. Xainano. Handwritten Mathematical Symbols. Kaggle, 2023. Available: https://www.kaggle.com/datasets/xainano/handwrittenmathsymbols. Accessed: May 28 (2023).
7. Pandi, S.S., Senthilselvi, A., Gitanjali, J., ArivuSelvan, K., Gopal, J., Vellingiri, J.: Rice plant disease classification using dilated convolutional neural network with global average pooling. Ecological Modelling, 474, 110166 (2022).
8. Guo, K., Ai, J.: FLANN-Based Improved KNN Medical Classification Algorithm. Comput. Mod. No.324(08), 25-29+35 (2022).
9. Ding, J., Cheng, H. D., Xian, M., Zhang, Y., Xu, F.: Local-weighted Citation-kNN algorithm for breast ultrasound image classification. Optik, 126(24), 5188–5193 (2015).
10. Li, S., Zhou, H., Fang, M.: Research on GPU-Based Image Supervised Classification Algorithm. Comput. Sci. 45(S1), 143-145+170 (2018).
11. Charbuty, B., Abdulazeez, A.: Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(1), 20–28. https://doi.org/10.38094/jastt20165 (2021).
12. Breiman, L.: Random Forests. Machine Learning 45, 5–3.

https://doi.org/10.1023/A:1010933404324 (2001).

13. Yu, Q., Wang, J., Jin, Z., et al.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72: 103323, (2022),
14. Salehi, A. W., Khan, S., Gupta, G., et al.: A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. Sustainability, 15(7): 5930, (2023).