# Automated classification of data: machine learning approach

Devang Vasani
*Master of Electrical and Computer Engineering*
*Carleton University*
Ottawa, Canada
devnaghasmukhbhaivas@cmail.carleton.ca

**Abstract -- This literature review examines machine learning's role in handwritten digit recognition and medical image classification. Evaluating algorithms such as CNN, KNN, SVM, and BP, the study utilizes MNIST and medical datasets, showcasing CNN's superior performance. The articles use MNIST and medical datasets to evaluate algorithms like CNN, KNN, SVM, and BP, indicating CNN's superior performance.**

**Keywords -- CNN, KNN, SVM, BP, learning rate, MNIST, learning algorithms, learning attributes, Image recognition, digit recognition, Image Preprocessing Techniques, Computed tomography (CT), data-preprocessing.**

## I. Introduction

The application of machine learning in our daily lives has expanded dramatically in this era of digitalization and automation. In the age of computers, handling handwritten numbers is a crucial part of information management. For many uses, including banking and finance systems, postal sorting, and education (grading multiple -choice questions with numerical answer), the ability to recognise hand-drawn numbers is crucial. Moreover, on the medical field, the development of computer-aided diagnosis systems also heavily depends on the automatic classification. This literature review evaluates empirical analysis of research presented by experts in this field.

## II. Challenges

While recognizing handwritten digits automatically, the diversity in ways of writing making it difficult to recognise numbers in handwritten text. On the other hand, accurate classification of medical images into suitable classes is hampered by the spatial complexity and variety of anatomy.

## III. Objective

The primary objective of data classification problem is to build a classifier that classify data accurately according to their predefined classes or labels. There are many machine learning algorithms has been developed for data classification problems. The main objective of this literature review is to compare the accuracy of all available machine learning algorithms to choose the best one for data classification problems.

## IV. Dataset

For handwritten digit reorganization, researchers used MNIST dataset which contain more than 60k images. This dataset contains images of 0 to 9 digits. All the images are in 28*28 pixel and is in grey scale. Each pixel has a value in the range of 0 to 255. Data for other research paper has been collected from CT scan images gathered from the Picture Archiving and Communication System (PACS) of the National Institutes of Health Clinical Centre. In another research paper data is taken from rom National Institute of Diabetes and Digestive and Kidney Diseases which was made available online at University of California. And for Anatomy-specific classification of medical images, data has been generated from 1675 patients in the hospital PACS. Through a CT scan of those patients, researchers captured medical images of various bodily parts which are in form of 2D version of medical images (32 * 32 pixels).

## V. Literature

In this research [1] the recognition of handwritten digits highlights the importance of using machine learning methods, in particular Convolutional Neural Networks (CNN), for classifying digit images from the Modified National Institute of Standard and Technology (MNIST) dataset. Paper suggested model makes use of a CNN architecture featuring many convolutional layers and max-pooling functions give best accuracy. They specifically use convolution layer that extract feature from input data. they use ReLU as activation function with maxpooling for down sampling. Also, to ensure model does not get overfit researchers use dropout layer. In this study, they also explain how each layers works in CNN and which layer use at which position to achieve highly accurate model. They also draw accuracy and loss function on graph to track and conclude how accuracy and loss function change with the number of epochs.

Researchers compared the accuracy of CNN, RNN, and SVM on the MNIST dataset in this study [2], splitting the data into ten thousand for testing and sixty thousand for training. The paper claim that the recommended CNN method is performed better then current methods for handwritten digit recognition. And to prove that, Researcher fit that data in every model and find accuracy of every model to choose which model perform well while testing. While training not only model but, he also changes

number of nodes, number of hidden layers, epochs, activation functions to found which model perform well at which attributes. To conclude, The CNN model performs well with 10 epochs, according to their comparison of CNN accuracy with varying epochs, sizes, numbers of layers, and activation functions. Additionally, they discovered that SVM had the best training accuracy but the lowest testing accuracy among of any model which indicate that SVM perform better for current dataset but, when we test that model on different dataset, it performs worst. Researchers also build confusion matrix to see how many data is mislabelled.

Researchers compared the accuracy of KNN, SVM, and BP in this study [3]. In this paper researcher divide this problem into 5 different parts which are, 1) sample collection 2) Data cleaning or preprocessing 3) Feature selection 4) Build classifier based on features we choose 5) Decision making based on classifier. Wenfei Liu divided the research into 70% and 30% training and testing, respectively. Additionally, the data collection is first converted into a 28 x 28-pixel bmp format image to make reading each handwritten digit. After that, 5000 training samples and 1000 test samples are randomly selected. They found that KNN perform worst while CNN perform best on MNIST dataset. Paper also gives overview about how each algorithm works. For example, KNN basically works on geometric measurement principle, in BP to minimize error, model use steepest descent optimization method to track feedback values in backward propagation, for CNN it Use convolutional layer, pooling layer, fully connected and Softmax layer to build a network which classify handwritten digits into their respective classes. And SVM mostly perform well for binary classification.

In this research [4], various paper has been discussed related to handwritten digit reorganization. To test different algorithms, Researcher explained evaluation method like recall, F-measure, precision, accuracy. In this paper they also try to improve model accuracy by implementing some features extraction techniques or by combining two algorithms, or by employing various methods. This paper analyses 8 supervised machine learning algorithms namely SVM, CNN, KNN, Random forest, Gradient Boosting Classification, logistic regression, naïve Bayes, and decision tree. After testing, they build confusion matrix for analysing data and through that, they found that CNN delivers an overall 98% accuracy for all labels, while other algorithms generate very good results as well. For example, Random Forest provides 97% accuracy in the train dataset and 97% accuracy in the test dataset. Support vector machines also perform well in the poly kernel after applying dimensionality reduction and hyperparameter tuning. They also mentioned that there is no over fitting in model since model perform well on test set also.

The research paper "Anatomy-Specific Classification of Medical Images Using Deep Convolutional Nets" by Roth et al. [5] discusses the important problem of automated human anatomy classification in medical images, especially computed tomography (CT) images. In this study, researchers collect data of 4,298 distinct axial 2D key-images and generated data from CT scans of 1675 patients. In this paper, researcher explain Data augmentation strategies prior to training and testing to increase the quality of the dataset and prevent overfitting of the model, which leads to better classification performance and lowers error by 3.7%. Also, while evaluating anatomy-specific classification, researchers used 80% of the data to train a multiclass classifier. ACU is also increased after augmentation which implies that the model is doing a great job at classifying and sorting cases. We can say that model is better at producing precise predictions. Moreover, it is not possible to achieve perfect 1 ACU values. so, 0.998 ACU value indicate our classifier is very next to perfect classifier.

In this research [6], researcher, build classifier which classify patient diabetic or not using various factors like age, Number of times pregnant, Plasma glucose concentration, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), Class variable (0 or 1).initially researcher analyse 9 different algorithms before applying than on real dataset. This paper suggest, when developing the model, the we should gave greater weight to which feature to choose. Researcher also compare different learning attributes on different learning algorithms and at last they are prioritizing precision, and accurate classification over model-building speed. Also, they highlight that a learning algorithm's performance best on one dataset does not necessarily translate to equivalent performance across datasets with different attributes. Also, in order to build good classifier ML algorithms must have higher precision, accuracy, and minimal error also model-building time is not that relevant but, it whould be better if it's low.

## VI. CONCLUSION

For any kind of image data, CNN is performed best while classifying data. Time is also a factor to consider while building any model. Furthermore, various methods give a spectrum of accuracy on various datasets so, no algorithm is universally better than others.

## VII. FUTURE LEARNING

Handwritten digit reorganization [4] produces accurate results for the English language; in the future, a model that produces accurate results for all languages must be developed. Additionally, as there can be very little margin for mistake in the healthcare industry, efforts are made to boost accuracy going forward [5][6].

## REFERENCES


[1] E. R. G, S. M, A. R. G, S. D, T. Keerthi and R. S. R, "MNIST Handwritten Digit Recognition using Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 768-772, doi: 10.1109/ICACITE53722.2022.9823806.

[2] A. K. Agrawal, A. K. Shrivas and V. k. Awasthi, "A Robust Model for Handwritten Digit Recognition using Machine and Deep Learning Technique," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-4, doi: 10.1109/INCET51464.2021.9456118.

[3] W. Liu, J. Wei and Q. Meng, "Comparisions on KNN, SVM, BP and the CNN for Handwritten Digit Recognition," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications( AEECA), Dalian, China, 2020, pp. 587-590, doi: 10.1109/AEECA49918.2020.9213482.

[4] R. C. Joshi, V. R. Patel and A. Goyal, "Evaluation of Supervised Machine Learning Models for Handwritten Digit Recognition," 2022 4th

International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 378-383, doi: 10.1109/ICAC3N56670.2022.10074292.

[5] H. R. Roth et al., "Anatomy-specific classification of medical images using deep convolutional nets," 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, USA, 2015, pp. 101-104, doi: 10.1109/ISBI.2015.7163826.

[6] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.