SSRG International Journal of Computer Science and Engineering
ISSN: 2348 – 8387 /doi:10.14445/23488387/IJCSE-V7I10P102

Volume 7 Issue 10, 6-12, October 2020

Building Data Mining Classification Model For Pixilated Digit Recognition System

Ziweritin, Stanley.[1], Ukegbu, C. C.[2] and Ezeorah, E. U.[3]
[1]*Department of Estate Management and Valuation, Akanu Ibiam Federal Polytechnic, Unwana-Afikpo, Ebonyi State, Nigeria*
[2,3]*Department of Computer Science, Akanu Ibiam Federal Polytechnic, Unwana-Afikpo, Ebonyi State, Nigeria*

Abstract *- Recognition is one of the major areas that have attracted the attention of different researchers, which can be applied in every sphere of life as technology advances. There are several problem domains in adopting data mining classification models with the rise in an exponential growth of structured and unstructured data. High metrics of success rate has not been recorded despite the existence and usefulness of data mining classification models in practice. Especially in the areas of testing and training of classifiers to recognize digits on pixilated images like the existing methods which are not efficient and encouraging in terms of speed and accuracy. Because of the segmented colour grid arrangements or formation of some digits. Therefore; we adopted the proposed model to overcome the challenges facing the methods on pixilated digit recognition systems. The aim is to build an efficient pixilated digit recognition system using neural network and support vector machine data mining classification models which can recognize digits within the range of 0-to-9 inclusively from pixilated or raster images. The system was successfully trained and tested in comparison to ascertain 94% and 99% accuracy level for support vector machine and neural network models respectively using Python programming language.*

Keywords: *Neural network, data mining, support vector machine, pixilated digit, recognition*

I. INTRODUCTION

In recent years, data mining classification methods have been gaining professional interest in the world of research for finding optimal solutions to different machine learning problems; namely: supervised, unsupervised and reinforcement learning[1]. Some of these well-known and widely used methods are support vector machine(SVM), Logistic regression(LR), decision tree(DT) neural network(NN), K-means Clustering, Multi Linear Regression(MLR), Associated Analysis(AA), Ensemble Models(EM), machine learning(ML) algorithms, etc., which can automatically extract relevant features from input data. Classification can be seen as a general problem in the field of data mining classification in which set of categories or groups are classified based on some observed features. But despite the existence and the usefulness data mining classification models in practice, much success has not been recorded in applying these methods for solving real problems in testing and training phases of data mining classifiers[13]. Digit recognition system can be seen as a method of training machines to recognizing a digit presented on papers, bank cheques, pixilated images, etc. The decision of classification is based on the training dataset containing some observable features with unknown and known categories of membership for unsupervised methods[11].

The concept of image processing in pattern recognition is to classify objects such as images, signal waveform or any other measurements needed to be grouped or classified into a number of categories depending on the use. The pixilated digit recognition system consists of preprocessing, feature extraction, training, testing and recognition[9].

This paper is divided into sections as followings: Section II presents a brief review of some of the previous approaches to the study area and the gap in exploring the proposed model; Section III, introduces the materials and methods which the different methods adopted and materials used for developing the model; Section IV, focuses on the results and detailed discussion of results; Section V presents the conclusion to the paper.

II. Related Works

A handwritten English alphabets recognition system was developed using Artificial Neural Networks(ANN) with binary pixels of the alphabets. The dataset of their proposed model was sourced from scanned documents, cleaned and smoothed. The characters of the English alphabets were formulated with 25 regular grids of cells. Scaling and thinning of segments for characters were done to obtain skeletal patterns and transformed into binary values presented on the pixels[10].

A decision tree classification model was proposed with 42,000 images as a sample on standard Kaggle dataset for Handwritten digit recognition system[15]. But the results were inefficient and not encouraging in terms of accuracy as such it could not recognize some of the digits correctly by the decision tree classifier because of the similarity of handwriting styles between some digits and produced low accuracy value of 75%.

Neural network(NN) classification model was adopted to identify sounds of English numerals from zero to nine(0-9) as samples from different people comprise of male and female. The model was tested with fifty(50) voice signals of male and female and produced approximately 82% accuracy level. It was quite satisfactory, but could not work well with the large volume of voice signals, vowels, word types of

input and sentences[12].

A deep machine learning model on the handwritten digit recognition system using a convolution artificial neural network(CANN) was developed to identify distorted data. The error margin of their learning model was set to 0.004(1e-4) in comparing the test and validation accuracy against the number of iterations at different stages on a multi-core CPU and general GPU to checkmate the computational time[14]. From their results, they concluded that the computational time and accuracy level of GPU decreased exponentially and slowly respectively as compared to CPU while the performance ration of GPU: CPU($\frac{GPU}{CPU}$) was found to be 30:1($\frac{30}{1}$). A framework of CNN and multi-level fusion techniques were developed for different classifiers to identify and classify handwritten images using MNIST Dataset and produced 98% accuracy level[8]. A new approach to recognize offline handwritten numerals using Multi-layer Perceptron(MLP) neural network and support vector machine(SVM) classifiers were proposed with 1200 samples of isolated images through the process of modified Hough transformation and four-view project profiles techniques. And produced an accuracy value of 93.12% and 72.5% for SVM and MLP neural network classifiers respectively but required more training time to work well[11].

An assembly of a neural network, decision tree, random forest and K-NN data mining classification models were presented on the handwritten digit recognition system[16]. The model was trained and tested with the MNIST dataset, which was quite satisfied with an overall prediction accuracy of 80%. But could not recognize some of the digits such as 1, 3, 8 in its pixilated form and also predicted digit Nine(9) in place of digit Zero(0) and verse-versa. A neural network(NN) model with the help of the back-propagation algorithm implemented on the handwritten character recognition system[4]. This model was trained and tested with the MNIST dataset loaded from the python sklearn library. The results were quite efficient but time-consuming because it required more training time and other resources to work well. The use of SVM data mining classification model was proposed and implemented on the handwritten digit recognition system with samples of images.

The model was trained and tested with training and testing dataset[6]. But the accuracy level of their proposed model was too low with the MNIST dataset. At the same time, an improved artificial neural network(ANN) on digit recognition system was developed with some adjusted parameters in controlling the NN objective function[5]. But produced low accuracy in prediction because of the absence of convolution neural networks(CNN). The use of Hill climbing algorithm was presented with a handwritten character recognition system for both lower and upper case letters[2]. The results of their model produced 93% accuracy on upper case letters and less than 93% on lower case letter. But required more training time and was unable to recognize or identify some of the handwritten characters. Handwritten digit recognition system using the principal component analysis(PCA) and single layer NN with image dataset was adopted[17]. The results of their proposed system produced 98.39% accuracy rate but required more time to perform well as required and space. The use of SVM and deep machine learning technique was adopted on large scale image recognition system with the "Hoda" Farsi handwritten digit dataset[7]. The performance of their proposed model was efficient. Still, the SVM was less accurate than the CNN because it required more training time to perform better as such, the overall accuracy level of their model was inefficient and not encouraging.

III. Materials and methods

It is important to express the fact that efforts have been made in the area of the digit recognition system. We intend to build on the areas that we have identified some lapses. In the existing methods, we were able to identify and narrow our study in recognizing pixilated digits ranging from 0-to-9 on images using the NN and SVM data mining classification models. The experimental dataset we intend using in designing the model is all promising in terms of accuracy, precision, and time complexity.

Pixilated digits presented on the images are all imported from the python sklearn built-in data library with a segmented grid of cells feed to the classifier. The gamma value of the SVM is set to scale, random states of being 101 and its tolerance level to 0.001. The NN adjustable alpha variable set to $1e^{-5}$ with ten hidden layers, random states set to 1, learning rate to 0.001, shuffle to be true for 200 iterations. The output of classifiers is used to test for the system validation after training.

Figure 1: The design of pixilated digit[18]

In the first phase of the proposed model; images are presented using a regular grid of cells and stored in a grey format fed into the preprocessing phase. The output of preprocessing is fed to the segmentation and feature extraction phases and presented with different coloured pixel arrangements to form digits ranging from 0-to-9, as shown in figure 1 and figure 2.

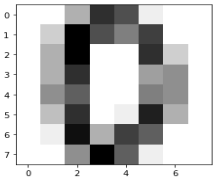


Figure 2: Digit 0 on the pixilated image from the proposed system

In this model, 8*8 =64 pixels of input features are feed to the classifier and trained with a total of 64 pixels and the output of the classifier used to test the classifiers for validation.

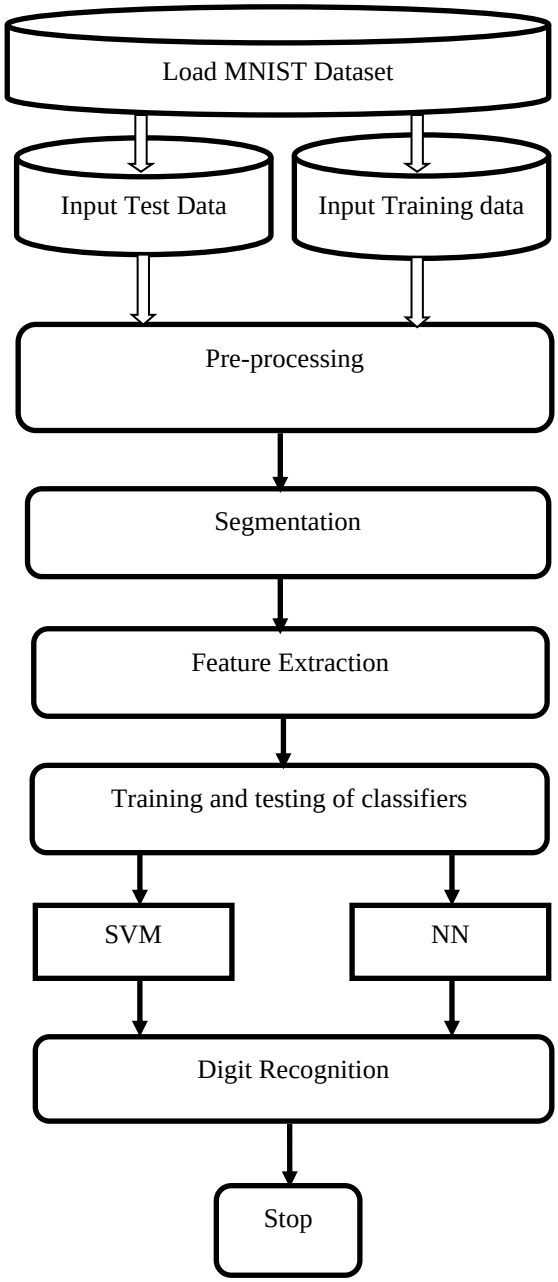


Figure 3: The Proposed System Architecture

A. Source of dataset

The data used by this model is sourced from the Modified National Institute of Standard and Technology(MNIST) experimental dataset contained in python sklearn library with offline pixilated digits ranging from 0-9 using the digits=datasets.load_digits() command. It's one of the largest database libraries in python commonly used for training and testing of different data mining classification models for images processing systems.

***B. Input training and testing data*:** The input dataset uploaded from the MNIST database of the proposed system is divided in into two sets at this stage; training and testing dataset. The training and testing dataset is shown in figure 3 have been loaded and passed to the preprocessing component to be formatted by some preliminary and classification processes in a top-down fashion; namely: preprocessing, segmentation, feature extraction and training of classifiers.

***C. Pre-processing**:* This is a preliminary process that involves different classification processes with the aim of formatting and feeding data obtained from MNIST dataset into the model. The data formatted can easily be broken down to smaller and equal pixels based on the adopted method of analysis. The comparison of knowledge-based and classified data patterns are carried out effectively in the analysis phase of preprocessing. The accuracy in the digit recognition system can be improved by preprocessing the raw data to remove noise. Thus, preprocessing of the raw data is deemed to be necessary before training and testing. It has to be processed in a way that it is suitable for the system to understand.

***D. Segmentation**:* is the process of partitioning an image containing digit into multiple regions(pixels of equal size) and extracting a meaningful region known as the region of interest(ROI which may vary from one digit to another.

E. Feature extraction: is a mandatory technique for any application involving image preprocessing. Feature Extraction is a process of extraction and generation of features to assist the task of object classification. This phase is critical because the quality of the features influences the classification task in adopting data mining classification models. The extended set of features is stored as a vector called the feature vector. The classifier takes the feature vector as input and performs the classification.

a) Classification is the systematic arrangement in groups or categories according to established criteria or observed feature. The classification stage is one of the key decision-making processes by recognition systems, and it uses the features extracted in the previous stage(preprocessing). The feature vector is denoted as X where $X = (f_1, f_2,..., f_d)$ where f denotes features and d is the number of features extracted from digits based on the comparison of feature vector characters as been efficiently classified into appropriate class for recognition.

F. Training and testing of the proposed system classifiers

Data from the future extraction component is feed to the training and testing stages of the proposed model. The total dataset is divided into two parts; the training dataset is 50%, and the test dataset is 50% of the total items respectively. To train the model, we used the training dataset and to evaluate the performance of the model; we used the test dataset. The training and test have 32 items for each dataset using python as a simulation language, and the trained system was further used for recognition. Generally, the performance of the classifiers depends on some factors such as the nature of data, the complexity of the problem and the nature of the learning algorithm.

G. Digit recognition

The data and its label are fitted to the model to learn from it. We do this by using the fit method to pass our training set to the proposed model. The sample digit is displayed as output and predicted with the help of re-shape method from a row vector and shaping it to form 8 by 8 matrix of pixels or grid of cells. We then created a figure after training and testing of the proposed model with a label for recognition, which will be displayed after prediction.

We intend to output the actual label of the sample gotten from the output. An extra feature is added to the current certainty of recognition/predicting system. In this feature, we are considering all digits in the range of the test dataset using the predict method to calculate the success rate of our model on the current training data.

a) Proposed system Methodology

In this study, we adopted the Object-Oriented Analysis and Design Methodology(OOADM). The object-oriented approach combines data and processes known as methods into single entities called objects. The stages for object-oriented design can be identified as: (a)—definition of the context of the system as static or dynamic. (b). Designing system architecture by partitioning the system into layers, and each layer is decomposed to form the subsystems.

(c). Identification of the objects in the system. The objects identified are grouped into classes. (d). The building of design models. (e).Design of object interfaces. The methodology breaks down the components of the proposed system based on the objects that surround the system and using the object components to build the new system around the identified objects. The new system will have relationships, activities and even dependences around the identified objects. In other to achieve a well-organized system, section of activities will be categorized into classes in a way that will make each group of activities and the processes easier to implement[10].

b) Neural network data mining classification model

The concept of a neural network was developed based on the weight connection between neurons[10]. A biological neuron in the human brain is a cell that consists of a nucleus, axon and dendrite branches. The axon is a transmitter that transmits signals to other neurons while dendrite as a receiver that receives signals from other neurons. The neural network has three main layers; input layer, an output layer and hidden layers. During the learning process, the weights can be adjusted to satisfy the input and output conditions. This is a well-known classification and prediction algorithm because it has a high tolerance rate to noise and can also be used to classify unseen data patterns[19]. We intend to adjust the weights by adding input bias and output bias to control the output function. The proposed neural network(NN) model uses set of neurons(activation function) in layers that are processed sequentially comprises of 64 pixels of inputs with ten hidden layers to display a single digit with the help of multi-layer perceptron neural network(MLPNN). The MLPNN model is used to estimate classification in data mining. There are neurons in the hidden layer that contain nonlinear activation function. All the inputs of i_1, i_2,...,i_{n-1} i_n, in dimensions of the data are multiplied by weights w_1, w_2, ..., w_{n-1}, w_n respectively before reaching the neuron, and the result is primarily collected in the linear processing unit. The output of the linear processing unit is passed to the output layer through the activation function in the nonlinear processing unit.

The error between the output value and the actual values are used to update the weights with the concept of slope minimization in providing a convergent approach to the goal in the updating process. Updated values of each weight(Δw) is recorded. We adopted the neural network activation function with the help of the backpropagation algorithm bellowed [3].

$A_j(\bar{x}, \bar{w}) = \sum_{i=0}^{n} x_j, w_{ji}$ Equation(5)

The output function uses the sigmoid function

$O_j(\bar{x}, \bar{w}) = \frac{1}{[1+e^{j(\bar{x},\bar{w})}]}$ Equation(6)

We defined the error function for the output of each neuron as:

$E_j(\bar{x}, \bar{w}, d) = \sum(O_j(\bar{x}, \bar{w}) - d_j)^2$ Equation(7)

The weights are all adjusted with gradient descendant

$\Delta w_{ji} = -\eta(\frac{\partial E}{\partial w_{ij}})$ Equation(8)

Where x is the inputs, w_{ij} are the weights O_j $(\bar{x}, \bar{w})$ are the actual outputs, d_j are the expected outputs and η is the learning rate

c) The SVM data mining classification model

The SVM classification method is one of the best data mining classification model being adopted to improve the recognition accuracy, memory management and training time of classifiers.

The model selection for the training of the SVM involves the following:

(a). Selecting the parameter C, kernel function and any other kernel parameter required. (b). The use of an appropriate algorithm to obtain α^2 and support vectors. (c). Compute the threshold value "b" using the support vectors.

Model selection in SVM is the process of selecting and adjusting of the kernel and its parameters. This help in minimizing some of the generalized errors or performance measures such as the k-fold cross-validation or leave-one-out(LOO) estimates in the SVM model.

IV. Results and discussion

We discussed different data mining classification models and adopted NN and SVM data mining classifiers. The NN was implemented with sixteen input neurons and ten hidden layers to produce single-output(A digit).

A. Training time variation of NN and SVM

The training time variation of the proposed model is the time taken to train our model with the training dataset to recognize digits from 0-to-9 on pixilated imaged, as shown below in table 1.

Table 1: The variation of recognition accuracy on pixilated digits in percentage

Label	SVM	NN
0	0.10501	1.49109
1	0.066	0.92605
2	0.066	0.95605
3	0.065	0.92205
4	0.066	0.93705
5	0.0661	0.94705
6	0.066	1.06906
7	0.09101	1.0110
8	0.067	0.96906
9	0.065	1.10206

Table 1 shows the training time variation for the proposed model. The training time of SVM lies between 0.065, 0.066, 0.067 to 0.10501 seconds while the NN falls between 0.92605, 0.93705 to 1.49109 seconds. The NN classifier required more training time compared to the SVM classifier. Therefore, in terms of time complexity (speed); the SVM is better than the NN.

Label	SVM	NN
0	0.001	0.03
1	0.001	0.0
2	0.0	0.0
3	0.001	0.0
4	0.0	0.0
5	0.001	0.0
6	0.001	0.001
7	0.0	0.001
8	0.001	0.0
9	0.0	0.001

Table 2; shows the validation time variation of the proposed model after training which changes depending on the digit. The validation time of SVM ranges from 0.0 to 0.001 second while the NN ranges from 0.0, 0.01 to 0.03 seconds. Therefore, the test validation time of NN is higher than the SVM model in general.

Figure 4: Data mining pie plot

Figure 4 shows the pie plots of NN and SVM classifiers with the respective percentages for the selected data mining classification methods in recognizing digits on the pixilated image. The SVM occupied 48.6% and NN 51.4% of the pie plot, which shows that the NN occupied a greater part of the pie plot compared to the SVM.

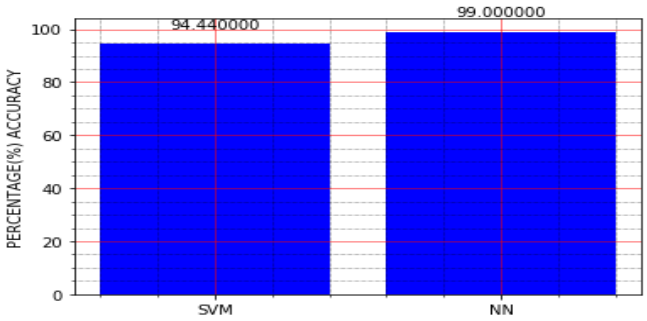


Figure 5: The metrics of accuracy for NN and SVM classifiers

Figure 5 shows the metric accuracy in terms of percentage for the selected data mining classification methods(SVM and NN). The NN recorded 99.00% success rate as the highest compared to the SVM with 94% metrics of accuracy.

Label	SVM	NN
0	94.44	100
1	94.44	99.00
2	94.44	99.00
3	94.44	99.00
4	94.44	100
5	94.44	98.00
6	94.44	100
7	94.44	100
8	94.44	100
9	94.44	100

Table 3 shows the variation in recognition accuracy for each pixilated digit ranging from zero(0)-to-nine(9) in the order of arrangement. The results of SVM produced constant recognition accuracy level of 94.44% while the NN displayed a varying predicted accuracy value that falls between 98.00%-to-100% depending on the digit.

b. A comparative analysis of the prediction accuracy for both models

Table 4 shows the comparative analysis of the recognition accuracy obtained from the data mining classification

models; NN and SVM measured in percentage.

Data Mining Method	Prediction Accuracy	Error
SVM	94.44%	5.56
NN	98.00%	2.00

From the results shown in table 4; it is evident that the predicted percentage accuracy and an error value of the proposed NN model is 98.00% and ±2% which performed better than the SVM with 94.00% and 5.56% error values respectively.

Table 5: The overall performance of Proposed system

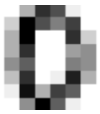
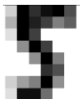
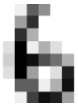
label	Pixilated Digit	Overall Performance
0		97.22%
1		97.22%
2		97.22%
3		97.22%
4		97.22%
5		97.22%
6		97.22%
7		84.22 %
8		93.7%
9		

Table 5 shows the Output digits from the MNIST dataset and overall performance in terms of percentage of the proposed system model for each pixilated digit. The overall percentage accuracy for digit 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9 is recorded as 97.22%, 96.72%, 96.7%, 96.2%, 96.7%, 97.2%, 96.7% and 93.7% respectively. Therefore, the metrics of accuracy falls in the range of 93.7% to 97.22% for all digits. The accuracy of the proposed model is the ratio of correct classifications(TP+TN) from the overall number of cases given in the equation below as:

$$\text{Accuracy}=\frac{\text{The total number of correct classification(TP+TN)}}{\text{The overall number of cases(TP+TN+FN+FP)}}$$

Equation(9)

Where TN represents the true negative, FP is false positive, TP is truly positive, and FN is false-negative cases.

V. CONCLUSION

The existing methods of operation in practice have shown not being capable of performing effectively and efficiently as expected in organizations that deal with the recognition system(RS). However, the metrics of accuracy for SVM was the same irrespective of the digit but varied in training and validation time for each digit. But the recognition accuracy and training time of NN changes for each digit which depend mainly on the complexity of the model and problem at hand. From the experimental results, we conclude that the NN model performed better than the SVM in terms of prediction accuracy and speed in recognizing digits on pixilated images.

REFERENCES

[1] M. J. Aditi, and T. kinjal, "*A Survey on Digit Recognition using Deep Learning*", International Journal of Novel Research and Development(IJNRD), ISSN: 2456-4184 Vol. 3 issue 4, April 2018, pp. 112-118

[2] P. Ayush, and S. S. Chauhan, "*A Literature Survey on Handwritten Character Recognition*", International Journal of Comp. Science and Information Technologies, Vol. 7 Issue 1, *January - February* 2016. *pp.* 1-5.

[3] V. Chandola, A. Banerjee, and V. Kumar, "*Anomaly detection: a survey*," ACM Computing Surveys (CSUR), Vol. 41 issues 3, pp. 15, 2009

[4] C. C. Dan, U. Meier, L. M. Gambardella, and J. Schmidhuber, (2010) "*Deep big simple neural Nets Excel On Handwritten Digit Recognition*", MIT Press, 56-876.

[5] L. Deng, "*The MNIST Database of Handwritten Digits images for Machine Learning Research*", MIT Press, pp. 46-876, 2012

[6] B. Hyeran, and L. Seong-whan, "*A survey on pattern recognition applications of support vector machines*", International Journal of Pattern Recognition and Artificial Intelligence (AI) Vol. 17, issue 3, pp. 459–486, 2003

[7] S. Karen, and Z. Andrew, "*Very deep convolutional networks for large-scale image recognition*", arXiv preprint arXiv: 1409.1556, 2014.

[8] H. Kaur, S. K. Wasan, "*Empirical study on applications of data mining techniques in healthcare*". Journal of Computer Science(JCS). Vol. 2, issue 2, pp. 194–200, 2006

[9] I. J. Kim, and X. Xie, "*Handwritten Hangul recognition using deep convolutional neural networks*". International Journal on Document Analysis and Recognition (IJDAR), Vol. 18, issue 1, pp. 1–13, 2014

[10] F. Lauer, C. Suen, and G. Bloch, "*A trainable feature extractor for handwritten digit recognition*", Pattern Recognition, Vol. 40,

issue 6, pp.1816–1824, 2007

[11] G. Mamta, P. Muktsar, A. Deepika, and P. Muktsar, (2013), "*A Novel Approach to Recognize the offline Handwritten Numerals using MLP and SVM Classifiers"*, International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345,4(7), 952-958.

[12] Y. Perwej, and A. Chaturvedi, "*Neural Networks for Handwritten English Alphabet Recognition*". Journal of Natural Sciences and Engineering(JNSE), Vol. 2, pp.67-89, 2011

[13] E. Salvador, J. C. B., Maria, G. M. Jorge, and Z. M. Francisco, "*Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models",* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, issue 4, pp.45-78, 2014

[14] P. Sarma, S. Sarmah, M. P. Bhuyan, K. Hore, and P. P. Das, *"Automatic Spoken Digit Recognition Using Artificial Neural Network",* International Journal of Science and Technology(IJST), Vol. 8, issue 12, ISBN 2277-8616, pp.1400-1404, 2019

[15] M. Shashank, D. Malathi, and K. Senthilkumar, "*Digit Recognition using Deep Learning*", International Journal of Pure and Applied Mathematics, ISSN: 1314-3395, Vol. 118, issue 22, pp. 95-301, 2018.

[16] C. Shengfeng, G. Yuwen, W. Lee, and A. Rabia, "*Offline Handwritten Digits Recognition Using Machine learning*", International Conference on Industrial Engineering and Operations Management Washington DC, USA, IEOM Society International, pp.275-286, 2018.

[17] A. A. Tsehay, and S. N. Pramod, "*Hand-written Digits Recognition with Decision Tree Classification: a Machine Learning Approach", International Journal of Electrical and Computer Engineering (IJECE), Vol. 9, issue 5, ISSN: 2088-8708, DOI: 10.11591,* pp.4446-4451, 2019

[18] S. Vinneet, and P. L. Sunil, "*Digits recognition using single-layer neural Network with principal component analysis*", Computer Science and Engineering (APWC on CSE), Asia-Pacific World Congress IEEE, 4-5, 2014.

[19] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, Buchan, I., Keane, J.: *"Comparing data mining methods with logistic regression in childhood obesity prediction*". Inf. Syst. Front, Vol. 11, issue 4, pp.449–460, 2009.