Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 9. No. 1, 2021

Article Citation Format
Nwile, B. N., Bennett, E. O. & Nwaibu, N. D. (2021): Pixilated Digit Recognition System Using Data Mining Classifiers. Journal of Digital Innovations & Contemporary Research In Science., Engineering & Technology.
Vol. 9, No. 1. Pp 49-56
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V9N1P3

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received:22nd April, 2021
Review Type: Blind
Final Acceptance:: 28th May, 2021

Pixilated Digit Recognition System Using Support Vector Machine and Logistic Regression Classifiers

[1]Nwile, B. N., [2]Bennett, E. O. & [3]Nwaibu, N. D.
[1,2.3]Department of Computer Science
Rivers State University
Port Harcourt, Nigeria
[1]**E-mails:** nwilebeauty@gmail.com,[2]bennett.okoni@ust.edu.ng

ABSTRACT

Data mining is an increasing and popular tool for dealing with large dataset as an interdisciplinary subfield of computer science used by data professionals. It is a computational method of discovering patterns in large datasets employing techniques of artificial intelligence. The overall goal is to extract information from dataset and transform it into an understandable structure for further use. But the large data collected at times is full of errors, noisy and have missing and false values. Such huge amount of data required data mining techniques to automatically discover hidden patterns in making decisions. The goal is to predict target class for each case in the dataset in which for most of the existing systems, classifiers are unable to recognize selected digits because of the segmented color grid arrangements of some digits. This paper adopts Support Vector Machine and Logistic Regression with some varying hyper-parameter values related to the design specifications in solving digit recognition problem. The system was successfully tested with Support Vector Machine and Logistic Regression techniques which results into 99.00% and 94.44% accuracy respectively.

Keyword: Data mining, Support vector machine, Logistic regression, recognition

1. INTRODUCTION

In recent years, data mining techniques have been gaining significant interest as tools for dealing with large datasets as a subfield of computer science used by data professionals[1],[2],[3].It is a computational method of discovering patterns in large datasets employing techniques of artificial intelligence for solving real life problems[4].It can be used to discover hidden patters from already existing data[5]. The exponential growth of data availability attracts the attention of professional in adopting data mining techniques. Aditi and Kinjal[6],identified the following data mining techniques that can be used to extract relevant features, namely: Support Vector Machine(SVM), Logistics Regression(LR), Decision Tree(DT) Neural Network(NN), K-means Clustering, Multi Linear Regression(MLR), Associated Analysis(AA), Ensemble Models(EM) and etc. Salvadoret *et al.*[7], described digit recognition system as the concept of training machines to recognizing or identifies a digit presented on papers, pixilated images and etc. Laureret *al.*[8]&Sarma *et al.*[9], defined digit/image processing in pattern recognition as the grouping of objects into classes or categories based on a particular criteria.

In digit recognition most the existing system classifiers are unable to recognize selected types of pixilated digits after training stage because of its complexity. Some could not do well with testing datasets because of the segmented color grid formation for some digits which results into over-fitting problem. Also, the existing methods of operation lack the merit of producing accurate and reliable results. The paper intends to develop an efficient pixilated Digit Recognition System using LR and SVM data Mining Classifiers. This will help predict accurately the target class for each case in the dataset. The proposed system classifiers will be trained and tested using the MNIST dataset obtained from python sklearn library.

2. RELATED WORKS

Perwej and Chaturvedi[10], proposed a model for recognizing hand written pixels of alphabets sourced from scanned and smoothed copies of documents using neural network. This was necessary to obtain skeletal patterns with better result. The alphabets were placed on regular grid of cells and converted into binary numbers as input which resulted to 83.5% accuracy level after training with the testing dataset. Ziweritin,*et al*.[11]; developed a NN and SVM classifiers trained and tested using the MNIST dataset to recognize pixilated digits.

The recognition accuracy of NN recorded 99% which was higher than the SVM classifier with 94% metrics of accuracy. There was variation in training recognition accuracy because of the different segmented color of digit formation for pixilated digits ranging from 0 to 9. The performance of NN classifier was low Srivastava*et al*.[12]; developed a feed forward NN model and multi-class SVM classifier to solve the problem of English characters recognition system on images.

This was done with the concept of image template matching as a technique to divide images into regular grid of cells and later matched to image templates. The classifiers were trained with image dataset to learn all the image features. The recognition model only worked well with some selected type digits because of time constraints and resulted in low accuracy level. A Random Forest(RF) based model was developed using Regression and classification classes in predicting grayscale digits embedded on images[13]. The RF-classifier was trained and tested using MNIST dataset with some fine-tuned hyper-parameter(n_estimators, training_size, error rateetc) values to improve the performance of the combined trees in the forest space. The RF-classification model produced 99% recognition accuracy higher than the RF-regressor class with 90% success rate which was low in performance. A CNN model was adopted with some categorical variables presented and mapped into integer values using one hot-encoding technique[14].

The proposed model was trained with 784-input neurons/features, 2-hodden layers and single output layer. The softmax NN activation function was employed for the output layers and the overall recognition accuracy was measured to be 93% by the multilayered perception neural network. A handwritten digit/character recognition system model was developed using Multi-Layered Perception(MLP), convolution neural network and SVM models[15];. The model was trained with input from documents and photographs containing intelligent handwritten text/digits. The CNN was recorded to be the best compared to SVM and MPL in terms of training time and recognition accuracy. The performance metrics was below average and not promising which required more training time and dataset to work well.

3. SYSTEM DESIGN

This paper focuses on SVM and LR because both can be used to solve multiclass classification problem as it relates to the proposed system. And can handle the problem of outliers efficiently with better accuracy. Pixilated digits are all invoked from the python sklearn built-in data library with segmented regular grid of cells feed to both classifiers.

Digit Dataset: The digit-dataset was obtained from the Modified National Institute of Standard and Technology (MNIST). The python scikit learn library was invoked and contained about 1797 pixilated images with 8 by 8 gridded cells. The experimental dataset consists of offline pixilated digits ranging from 0 to 9. The MNIST dataset contained pixilated digits commonly used for training and testing of different data mining classification models for images processing systems.

Data acquisition

The MNIST dataset was generated from python scikit-learn python library containing digits ranging from 0, 1, 2, 3, 4, 5, 6, 7, 8and 9.The data generated contained pixilated digits embedded on images with different color formation and arrangement. And the total dataset was divided into 50% training and 50% testing set.

Data preprocessing

The pre-processing of the raw data is necessary for the training images to reduce threshold value of the pixilated image from the MNIST database. The digits ranging from 0 to 9 are super imposed on regular grid of equal cells called pixels. This contained 8-rows by 8-columns of equal sizes to produce a total of 64 pixels of input forming different colors with 1,797 images[16]. The digit was read as input, converted into RGB image and stored in the form of array with different colored arrangement.

Logistic regression (LR) classifier

We created a regressor class to classify patterns and recognize digits on pixilated images. We passed-in some parameter values such step size, maximum number of duration and class of interest using a binary classifier stored as membership variables[17]. The class of interest was past to the model ranging from digit 0 to 9 with the concept of one versus all. We created 10-labels ranging from 0 to 9, a Boolean variable called Y for each class and introduced a new column called X as bias term to all digits.

Support Vector Machines (SVM) Classifier

The SVM is one of the simplest and more preferred data mining classifier used by data professionals because it can produce better and high accuracy with less computational error[18]. The SVM uses two main concepts namely; hypothesis space and the loss function in finding an "optimal" hyper-plane as a solution to any learning problem[19]. The SVM is memory efficient and uses subsets of training data points in the support vectors called decision function. The simplest formulation of SVM is the linear one, where the hyper-plane lies on the space of the input data[20]. The SVM estimator was defined on the training dataset to recognize pixilated digits formation and the estimator was tested to interpret the testing set. The visualization was done using matplotlib library in python. A SVM classifier was created using the linear kernel and trained with the pre-processed training data to make predictions.

Algorithm 1: Logistic Regression

```
Step   Processes involved
 1     Start
 2     Initialize residuals
 3     Compute weight values
 4     Compute value for adjustable response
 5     Search and update directions
 6     Compute the length for each step
 7     Display approximated solution
 8     Update residuals
```

Algorithm 2: Support vector machine(SVM)

```
Step   Processes involved
 1     Start
 2     Find candidate_SV with closest pair from classification (SV=>support vector)
 3     If there are violating points then
 4             Find violating_points
 5     Compute the candidate_SV= candidate_SV + voilating_points)
 6             If there is any $\alpha_p$< 0 due to the addition of c to S that gives negative then
 7     Candidate_SV = candidate_SV/p
 8                     Repeat module to prune all the data points
 9     end_if
10     end_if
```

4. RESULTS AND DISCUSSION

We are discussing in detail about the results of SVM and LR obtained from the proposed system training and testing dataset as presented below in figures and tables. The pie chart, bar charts, tables and graphs are used for presentation of results and discussion.

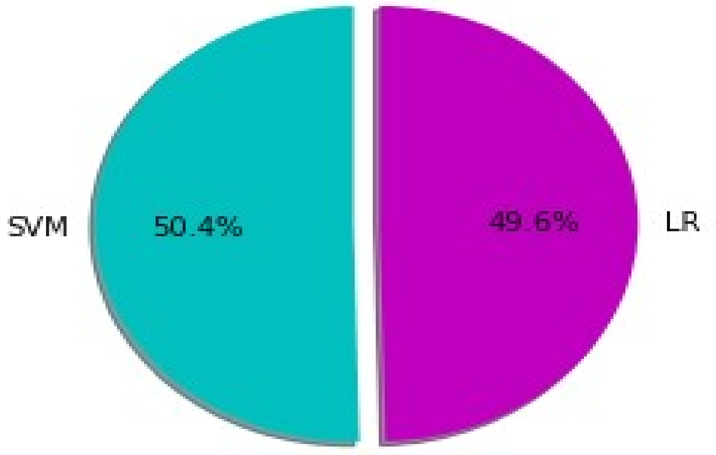

Figure 1: Data Mining Pie Plot of SVM and LR

Figure 1 depicts the pie chart for each area occupied by LR and SVM classifiers measured in percentage. The SVM occupied about 50.4% of the pie plot compared to SVM with 59.6% area within the chart. The result shows both models are occupying equal area in the pie plot.

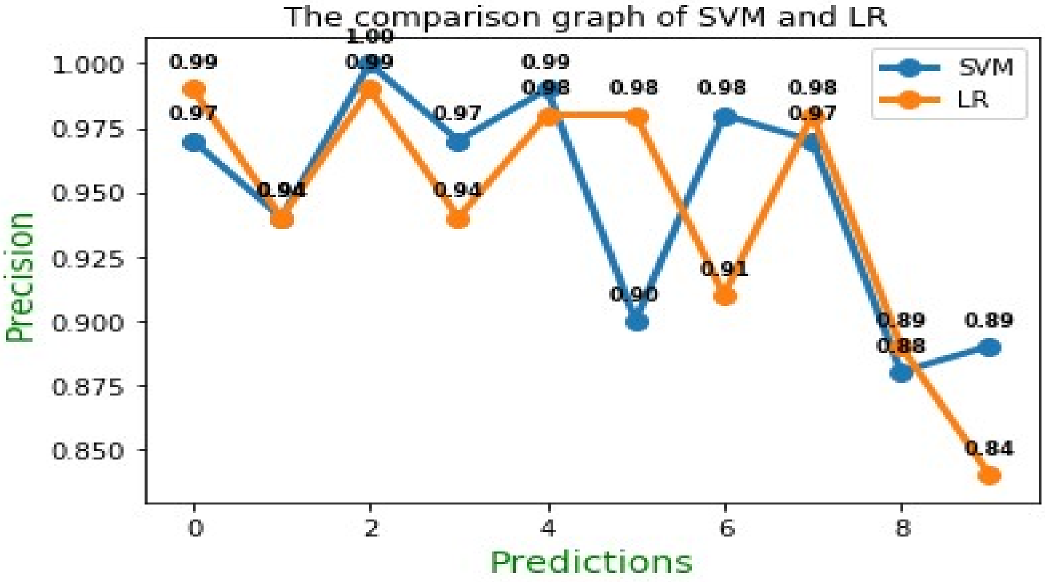

Figure 2: Variation in Prediction Accuracy.

Figure 2 shows the accuracy graph of SVM and LR in predicting digits from 0 to 9. At digit 0 the LR was higher, decreased and increased again in that order compared to SVM but produced the least at digit 9.

Table 1: The Training Time Variation Measured In Seconds

Methods	0	1	2	3	4	5	6	7	8	9
SVM	0.21	0.20	0.25	0.26	0.12	0.19	0.27	0.16	0.23	0.20
LR	7.21	6.44	7.37	7.59	4.81	7.00	6.61	7.37	7.45	7.53

Table I shows the training time variable of SVM and LR data mining classifiers with their variation in training time measured in seconds to recognize digits. The pixilated digit on images ranges from 0-to-9. The training time of SVM classifier is recorded to be faster and smaller than the logistic regression techniques but varies from one digit to another.

4.1 Performance evaluation

The accuracy, F1-score, precision and recall tools are employed to evaluate the performance of SVM and LR data mining classification models.

Classification accuracy is the ratio of the correctly classified data points to the total number of points in the dataset. This ranges from 0 to 100%

$$\text{Classification accuracy} = \frac{\text{Number of correct classifications}}{\text{Total number of classifications}} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

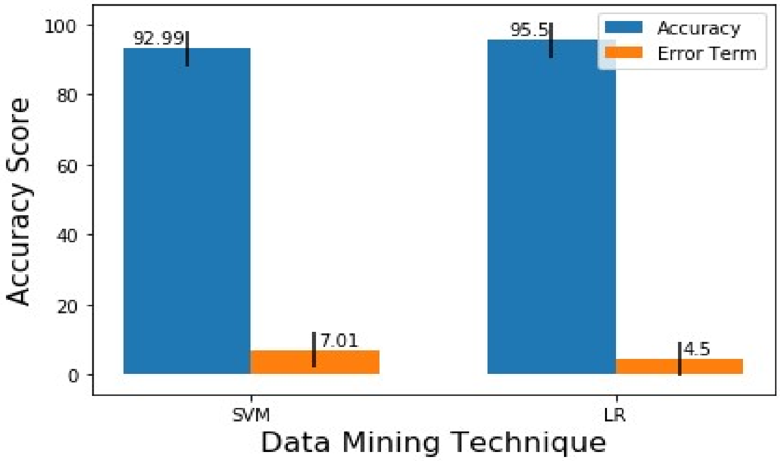

Figure 4: The accuracy and error terms of LR and SVM

Figure 3 shows the performance accuracy and error terms of SVM and LR measured in percentage after training and testing with same dataset. The LR gave 95.50% and 74.5% which was higher in recognition and low in error rate than the SVM that produced 92.99% with 7.01% error rate.

Precision: is a metrics used to measure the positive classifications represented as follows:

$$\text{Precision}=\frac{TP}{TP+FP} \qquad (2)$$

Recall is a metric used to measure the false negative classifications represented as:

$$\text{Recall} =\frac{TP}{TP+FN} \qquad (3)$$

F1-score: takes into consideration the true positive and false positive regardless of false negative and false positive classifications. The F1-score is sensitive to which class is positive and negative as given below in equation 4:

$$\text{F1-score}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2*TP}{2*TP+FP+FN} \qquad (4)$$

Table 2: Classification Report of SVM

Digit	Precision	Recall	F1-score	Support
0	0.97	0.99	0.95	88
1	0.94	0.90	0.92	91
2	1.00	0.99	0.99	86
3	0.97	0.86	0.91	91
4	0.99	0.95	0.97	92
5	0.90	0.97	0.93	91
6	0.98	0.99	0.98	91
7	0.97	0.96	0.96	89
8	0.88	0.92	0.90	88
9	0.87	0.93	0.90	92
accuracy			**0.94**	**899**
Macro avg	**0.95**	**0.94**	**0.94**	**899**
Weighted avg	**0.95**	**0.94**	**0.94**	**899**

Table 2: depicts the classification report of SVM containing the precession, recall and f1-score accuracy level. The precision accuracy score ranges from 0.88 to 1.00, recall score ranges from 0.99 to 0.68 and f1-score ranges from 0.90 to 0.98 inclusively. The recorded precision score shows to be higher than the recall and f1-score respectively.

Table 3: Classification Report of LR

Digit	Precision	Recall	F1-score	Support
0	0.99	0.95	0.97	88
1	0.94	0.90	0.92	91
2	0.99	0.98	0.98	86
3	0.94	0.84	0.88	91
4	0.98	0.91	0.94	92
5	0.88	0.95	0.91	91
6	0.91	0.99	0.95	91
7	0.98	0.96	0.97	89
8	0.89	0.90	0.89	88
9	0.84	0.93	0.89	92
accuracy			**0.93**	**899**
Macro avg	**0.93**	**0.93**	**0.93**	**899**
Weighted avg	**0.93**	**0.93**	**0.93**	**899**

Table 3 shows the classification report of logistic regression with precession, recall and f1-score accuracy from 0 to 1.00. The precision accuracy score ranges from 0.84 to 0.99, recall from 0.84 to 0.99 and f1-score to be 0.88 to 0.97 inclusively. The recorded precision score shows to be higher than the recall and f1-score respectively.

5. CONCLUSION

The existing methods are inefficient in identifying digits on pixilated images as required with Recognition System(RS). However, the metrics of accuracy of SVM was higher compared to LR. From the experimental results, we conclude that the SVM model performed better than the LR in terms of prediction accuracy and speed in recognizing digits on pixilated images.

REFERENCES

[1] Mamta, G, Muktsar, P., Deepika A., and Muktsar, P. (2013),A Novel Approach to Recognize the off-line Handwritten Numerals using MLP and SVM Classifiers, *International Journal of Computer Science and Engineering Technology (IJCSET), vol. 4, Issue. 7,pp952-958.*

[2] Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., and Keane, J. (2009). "Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction," *Inf. Syst. Front(ISF). vol. 11, Issue. 4, pp449–460.*

[3] Deng, L(2012). "The MNIST Database of Handwritten Digit Images for Machine Learning Research," *IEEE Signal Processing Magazine, vol. 29, Issue. 6, pp141-142*

[4] Bharati, M. and Ramageri, K.(2004). "Data Mining Techniques and Applications," *Indian Journal of Computer Science and Engineering(IJCSE), vol. 1, Issue. 4, pp301-305*

[5]. Shashank, M., Malathi, D. and Senthilkumar, K.(2018). "Digit Recognition using Deep Learning",, International Journal of Pure and Applied Mathematics, vol. 118, Issue. 22, pp2018, 95-301

[6]. Aditi, M. J., and kinjal T.(2018). "A Survey on Digit Recognition using Deep Learning," *International Journal of Novel Research and Development(IJNRD)*, vol. *3, Issue. 4, pp112-118.*

[7] Salvadoret, E., Maria J. C. B., Jorge G. M. and Francisco Z. M.(2014). “Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, Issue. 4, pp45-78*

[8]. Lauer, F., Suen, C., and Bloch, G. (2007). "A Trainable Feature Extractor for Handwritten Digit Recognition," *Pattern Recognition vol. 40, Issue. 6, pp1816–1824.*

[9] Sarma, P. Sarmah, S., Bhuyan, M. P., Hore, K. and Das, P. P.(2018) "Automatic Spoken Digit Recognition Using Artificial Neural Network," *International Journal of Science and Technology(IJST), vol. 8, Issue. 12, pp1400-1404.*

[10] Perwej, Y. and Chaturvedi, A. (2011). "Neural Networks for Handwritten English Alphabet Recognition," *Journal of Natural Sciences and Engineering*(JNSE), vol. 2, pp67-89.

[11]. Ziweritin, S., Ugboaja, U. C. A., and Osu, C. M.(2020). "Random Forest Model for Predicting Grayscale Digits on Images," *International Journal of Scientific Research in Computer Science and Engineering(IJSRCSE), vol. 8, Issue. 6, pp1-7.*

[12] Srivastava, S., Kalani, S., Hani, U., and Chakraborty, S.(2017)."Recognition of Handwritten Digits Using Machine Learning Techniques," *International Journal of Engineering Research and Technology(IJERT), vol. 6, Issue. 5, pp711-714.*

[13]. Ziweririn, S., Ukegbu, C. C., and Ezeorah, E. U.(2020), "Building Data Mining Classification Model for Pixilated Digit Recognition System," *SSRG International Journal of Computer Science and Engineering(IJCSE), vol. 7, Issue. 10, pp6-12.*

[14] Thangamariappan, P., and Pamila, J. C. M. J.(2020). "Handwritten Recognition by using Machine Learning Approach," *International Journal of Engineering Applied Sciences and Technology(IJEAST), vol. 4, Issue. 11, pp564-567.*

[15] Dixit, R., Kushwah, R., and Pashine, S. (2020). "Handwritten Digit Recognition using Machine Learning and Deep Learning," *International Journal of Computer Applications(IJCA), vol. 176, Issue. 42, pp27-33*.

[16] Bellazzi, R., and Zupan, B.(2008). "Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines," *International Journal of Medical. Information(IJMI), vol. 77, pp81–97.*

[17]. Ayush, P. and Chauhan, S. S.(2016) “A Literature Survey on Handwritten Character Recognition”, International Journal of comp. Science and Information Technologies, vol. 7, Issue. 1, pp1-5.

[18] Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J.(2010). "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," *Neural Computation(NC), vol. 22, Issue. 12, pp3207-3220.*

[19] Purohit, A., and Chauhan, S. S.(2016). "A Literature Survey on Handwritten Character Recognition," *International Journal of Computer Science and Information Technology(IJCSIT), vol. 7, Issue. 1, pp1-5.*

[20] Byun, H., and Lee, S. W.(2003). "A Survey on Pattern Recognition Applications of Support Vector Machines," *International Journal of Pattern Recognition and Artificial Intelligence(IJPRAI), vol. 17. Issue. 03, pp459-486.*