Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 8 July, 2021: 5987 – 6000

Research Article

Recognition of Math Expressions & Symbols using Machine Learning

Sagar Shinde[1*], Akil Mulagirisamy[2], Daulappa Bhalke[3], Lalitkumar Wadhwa[4]

[1]Postdoctoral Fellow, Department of E & C Engineering, Lincoln University College, Malaysia.

[2]Associate Professor, Department of E & C Engineering, Lincoln University College, Malaysia.

[3]Professor, Department of E & TC Engineering, AISSMS College Of Engineering, Pune, Maharashtra, India

[4]Principal, Nutan Maharashtra Institute of Engineering & Technology, Talegaon, Pune, Maharashtra, India.

Corresponding author: sagar.shinde5736@gmail.com

Abstract

The symbols and expressions used in math are very important and used in daily routines. It is necessary to have them in electronic form to access and produce them easily. The goal of this research is to provide a bridge between the knowledge of the organizer and the user inputs. By making math symbols and expressions visible to the non-native speakers, this research could introduce them to math notation. The method is to be recognized mathematical formula as well as symbols (printed or handwritten) based on feed forward back propagation neural network, support vector machine and K- nearest neighbor classifier. The noise free clean & clear input image can be obtained in preprocessing. The elements associated with math equations can be isolated with the utilization of segmentation. Finally feed forward back propagation neural network, support vector machine and K- nearest neighbor classifier with extracting static and complex features is used to recognize the expressions and symbols on the handwritten and printed image too. The receiving operating characteristics (ROC), confusion matrix & overall recognition system determines the accuracy and efficiency of the proposed system. The recognition of handwritten mathematical equations, symbols, digits provides a lot of real time applications as well as non-real time applications and most important application of math equation recognition system is math talk system for visually impaired people.

Keywords: math symbol, math expression, statistical, complex, ROC, SVM, K-NN, BPNN.

1. Introduction

Since the time of humans, writing has been an integral part of their lives. It is used to archive and preserve knowledge. The rise of technologies such as electronic and computing has made the automation of processing tasks, especially reading and research, increasingly difficult. This is of concern to researchers in fields such as form recognition. The concept of automatic recognition of writing dates back to the

1960s. Its complexity stems from the need to interpret printed text to a machine that can read it. The recognition of writing is a process utilized to translate printed text into a text that is understandable by a machine. It does so by sending a text to a machine that can read it. The recognition of writing is usually related to the mass processing of documents. This library provides a wide range of repetitive tasks that are commonly handled by large databases. Some of these include the processing of administrative files, reading of postal mail, processing of forms, checking bank balances, and searching for information in archives. It therefore covers large repetitive applications with large databases such as automatic processing of administrative files, automatic sorting of postal mail, reading of amounts and bank checks, processing of postal addresses, Processing of forms, keyboardless interfaces, analysis of written gestures, reading of legacy documents, indexing of library archives and searching for information in databases. The computerized reading of writing has seen measurable achievement from last couple of years. This is due, on the one hand, to the many works carried out leading to a variety of different approaches and, on the other hand, to the performance of computers and current acquisition systems coupled with modern statistical methods for example hidden markov models, support vector machines and neural networks. The existence of standard international databases that enable researchers to perform reports on the performance of their various approaches in handwriting and printing allowed them to easily compare their results. On the one hand, mathematical notation is an example of a very rich 2D language. On the other hand, the use of mathematical notations is indispensable in the scientific documentation. a system without constraints is much more natural to use but much more complex to achieve. The recognition of 2D languages is a real challenge that requires the marriage of skills from several fields. The proposed system can be used for scanning and recognizing the mathematics answer sheets in university, blind math applications etc.

2. Related Works

The recognition of character, symbol and math equations is an integral part of pattern recognition. It is utilized for recognizing optical character, symbol and expressions recognition. With the help of an image, an equation can be quickly interpreted to a computer. This technique can be utilized by different individuals and organizations to interpret math paper reports. **Mehmet et al. (2011)** represented a framework for a probabilistic system for recognizing math expressions. The current system can recognize short expressions in real time. This paper shows how it does so and what its context sensitive grammar does. "**Alvaro et al. (2013)** described a utilization of hybrid features which has mixed the online and offline data. This paper presents a recurrent neural network classifier that is based on a set of features that can detect different kinds of representations. It was compared to a standard neural network classifier. **Chaturvedi et al. (2014)** represented a framework for a pattern recognition system based on a neural network & an Izhikevich neuron model. This paper compares the feed-forward and the Izhikevich neural networks for handwritten pattern recognition. "**Le et al. (2014)** explained the implementation for identification of online manually written math expressions (MEs) and improvement of structure analysis. This paper proposes a method to learn and improve structural relations through two SVM models that are trained without making any heuristic decisions. The goal of this study was to demonstrate how an improved back propagation can speed up the training process. It allowed the neural network to train with fewer numbers of epochs. The framework can learn various patterns of the same character if it learns them under a similar label. **Ratnamala S.Patil, Shilpa [39],** Document image analysis is a challenging task that involves recognizing math symbols in a document and ordering the record as non-maths or math-

related documents. This paper proposes a novel method for document classification that takes into account the statistical features of a document.

Through the qualitative and quantitative analysis, it has been concluded that, in order to properly recognize complex handwritten math symbols & equations, the researchers need to conduct studies that are focused on the various issues that arise when over writing and touching symbols. The literature survey said that for high recognition rate, multilayer perceptron neural network with back propagation algorithm, support vector machine and nearest neighbor classifiers are mostly used. Some unique and weighted features have to extract to get overall good performance, efficiency, throughput and high recognition rate. The features like width to height ratio, centroid, zoning, boundary box, labeling and counting total number of elements in the equations, statistical features as well as complex features such as entropy, skew, kurtosis, variance, mean, standard deviation, number of elements in the equation, labeling etc. It is necessary to consider sensitivity, specificity, accuracy, precision in the recognition of handwritten mathematical equations and symbols to get good experimental results with simulation. The receiving operating characteristics and confusion matrix will be useful for measuring the performance of recognition in terms of efficiency and throughput.

3. Architecture & Methodology

System architecture involves the flow from preprocessing through recognition. The number of math symbols viz. sigma, ohm, plus, minus, zero, comma etc and equations viz. convolution summation, quadratic equation, law of indices, straight line etc have been considered for the testing purpose through simulation. The various math symbols and equations captured and image which is acquired as an input is gone through binary and noise reduction techniques. On noise free and clear image, segmentation has applied by performing some morphological operations and extracting weighted features such as labeling, counting total number of elements, statistical as well as complex features. Finally support vector machine and K-NN classifiers has been applied to recognized the input image. The dataset has been created through various people with the age in between 18 to 60 years and ranging from students, faculties to banking clerks. It has been observed that there is a remarkable variation due to style of handwriting, stroke based writing, overwriting and touching or breaking in symbols as well as in equations. The complete flow of recognition with all details has been mentioned in below fig.1.

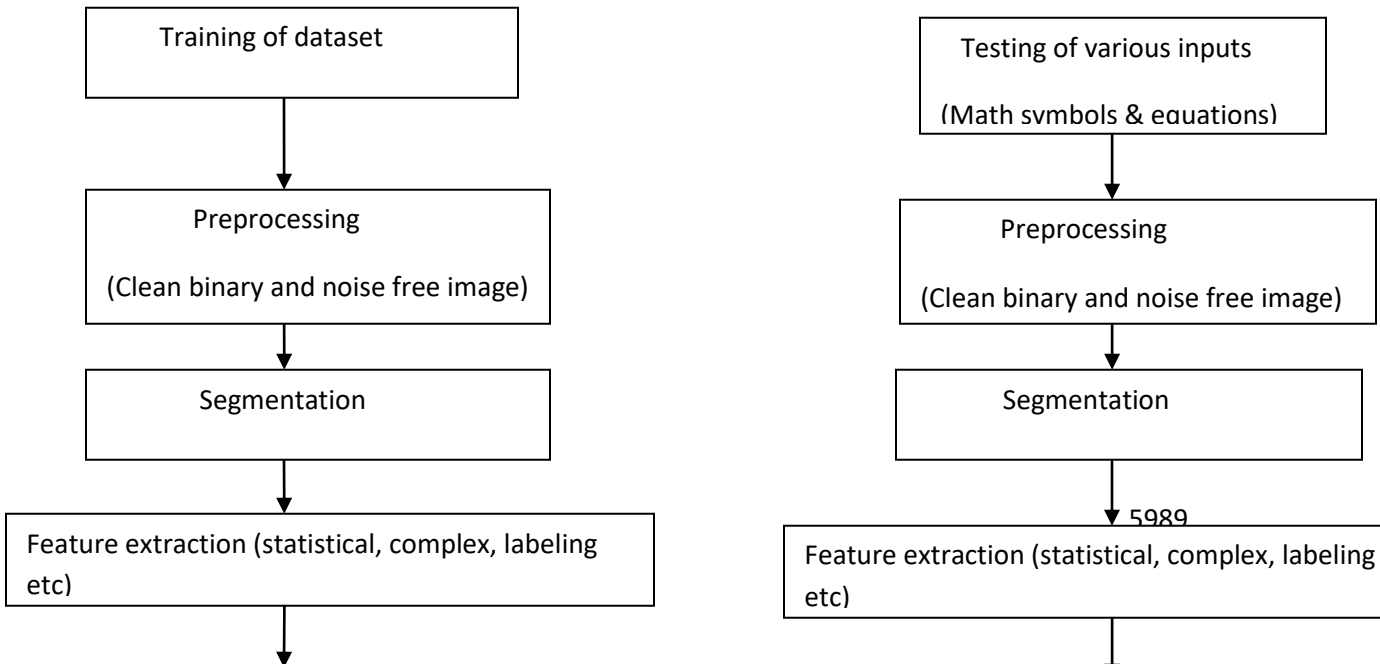


Fig. 1. Architecture & methodology of proposed system

4. Simulation Results & Discussion

The various math symbols and equations are the acquired images one by one as input and gone through system flow as shown in fig. 1. First of all, math symbol or equation to be recognized is preprocessed to get noise free clean & clear image. The preprocessor operation is carried out to convert an image into binary. The operation is done by taking the input image and converting it into binary. The successive formulary is utilized for the operation of preprocessing.

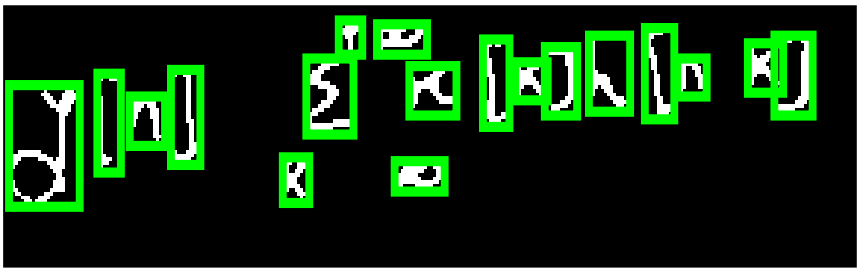
$$B_i(x,y) = \begin{Bmatrix} 1\text{ , if } P_i(x,y) > 1 \\ 0, \text{if } P_i(x,y), 1 \end{Bmatrix} \quad (1)$$

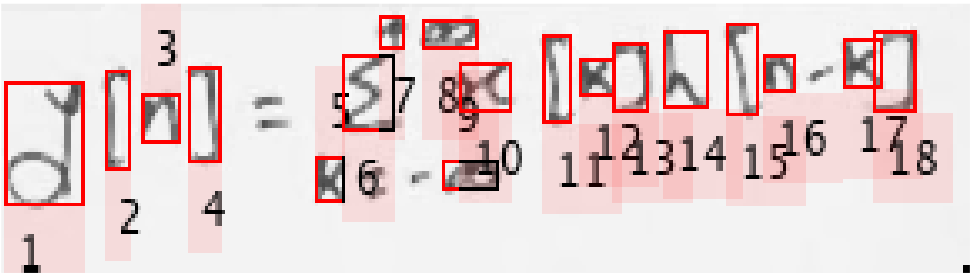
Where P_i = pre-processed and B_i = Binary Image.

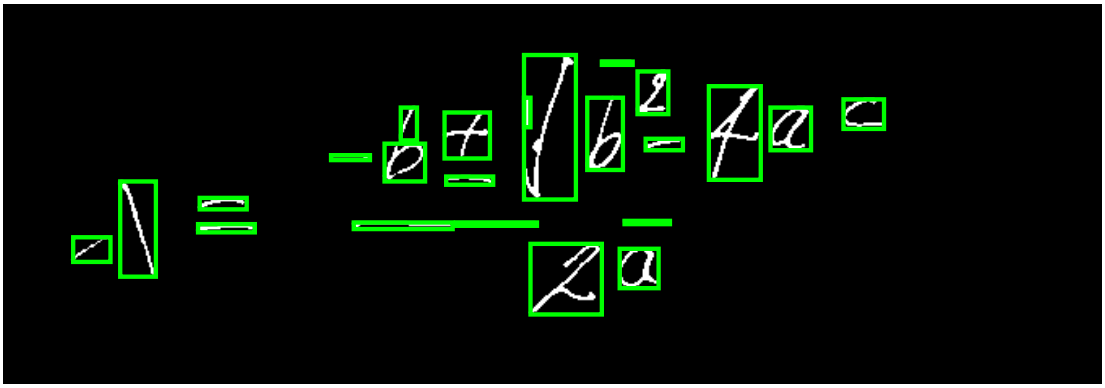
The row and column sum of the element's location to give the location of the component in the image.

4.1. Segmentation

The process of separating an object from the rest of the equation is known as segmentation. The total number of elements or objects has been counted. It is shown in fig. 2. Basically it has used for detecting an object from a background and breaks the image up into segments. Morphological operations are used to identify and classify objects in an image. Two basic morphological operations dilution and erosion are used to identify objects based on their shape. It can also detect the object of interest that is closest to the object.







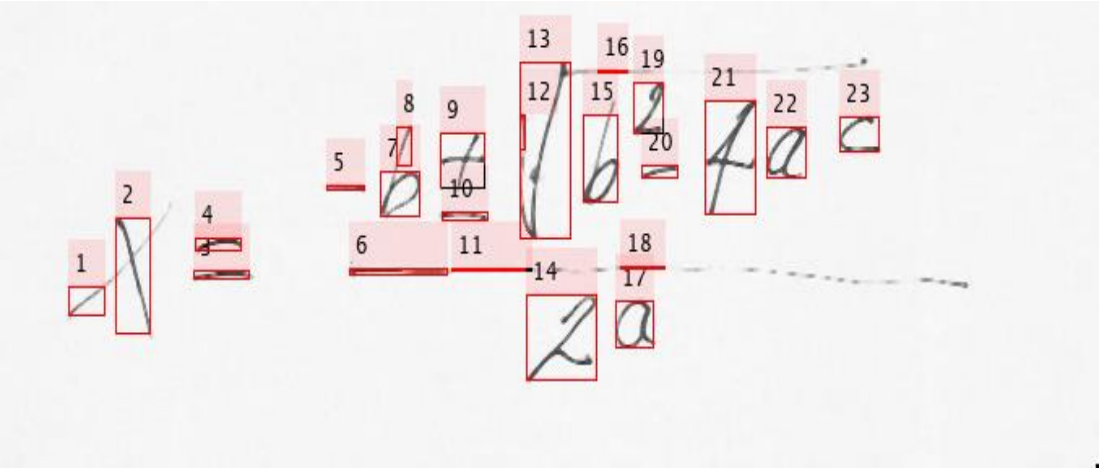


Fig. 2. Segmentation (Separation & Labeling)

4.2. Feature extraction

In order to learn, features are needed to be extracted from the data. This step helps in reducing the amount of redundant data and accelerates the learning process. The data extraction also helps in reducing the complexity of the model. Feature extraction is the task of pin pointing the features that will be utilized for investigation. It helps minimize the amount of data that's needed to build a model. The statistical & complex features have been extracted from the input symbols or equations to be recognized. Instead of trying to understand the structure of the texture, they represent it indirectly by describing the distributions and relationships among the grey levels of an Image. Actually the identification and effectiveness is based on the extracted features. The statistical and complex features viz. entropy, mean, variance, standard deviation, skewness, kurtosis, correlation and contrast have been extracted. The results of feature extraction has mentioned in below table 1.

Table I. Feature Extraction from various math equations & symbols

Sl. No.	Type of equation or symbol	Entropy	Mean	Variance	Standard deviation	Skewness	Kurtosis	Correlation	Contrast
1	Convolution Sum	5.3152	0.91823	0.00020847	0.071231	3.0502	14.8096	0.1017	692.375
2	Quadratic Equation	9.5483	0.95399	3.9957e-05	0.046919	-5.0122	43.4196	0.038443	686.1132

3	Plus/Addition	0.0089488	0.9859	0.0011784	0.10211	0	0	0	682.0409
4	Zero	0.018138	0.99521	0.00012126	0.061276	0	0	0	676.2367
5	Exclamatory	0.0048414	0.99491	0.00014844	0.064796	0	0	0	679.3357

4.3 Classifier

There are various classifiers are available in machine learning and most popular classifiers in the case of pattern recognition are support vector machine & K-nearest neighbor.

The support vector machine is a constitutive part of pattern perception. Its capability to classify future data makes it an ideal tool for pattern recognition. The various algorithms and models under SVM have been described for pattern recognition that uses SVM. They proved their efficiency when compared to other methods. SVM is a method that combines the separation of hyper planes and data points with the aim of reducing the upper bound of generalized error. It is widely used in pattern recognition. The goal is to create a hyper plane that maximizes the distance between two class boundaries, with all of the input data going to one side of the hyper plane and divide the data into two classes. The points of one class are on one side of the plane, while the points of the other class are on the other side. SVM classifier is a type of discriminative algorithm used for complex equations. It is good for both classifications and regression challenges. A hyper plane is a type of classification that can achieve a maximum distance between the classes. This ensures that it can detect patterns with high accuracy. This algorithm uses a subset of the decision function's training points to improve its performance. It does so by acquiring good performance without having any prior knowledge of the underlying data. SVM is a data-oriented algorithm that provides maximized margin in data. Its main goal is to separate the data into separate segments.

$$x \in R^1 \Phi(x) R^H \quad (2)$$

Where $\Phi(x)$ is kernel function, utilize to get hyper plane. The hyper plane with SVM is shown in below fig. 3. It is not essential to train the data frequently in SVM. Data is partition in two category as +1 or -1 for 2-D. Also it needs the two hyper planes to separate data points in 3-D data.

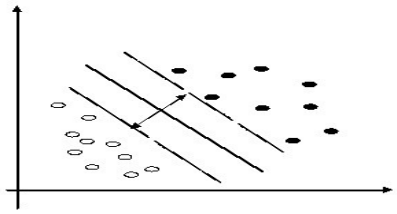


Fig.3. Hyper plane with SVM

The k-nearest neighbor is simplest machine learning algorithms for classification has been used as the second classifier in this implementation. Different feature set fed to the k-NN and the classification accuracy has been determined. The numbers of neighbors (k) for each feature set were varied and the accuracy of classification has been estimated. K-Nearest neighbor is a supervised learning algorithm which can be used for classification problems. It works by assuming that every data point falling in the same class is falling in the same direction. The nearest class to the point that is to be classified is computed using the Euclidean distance. It's a good classifier to implement but it can be very slow once the data gets big enough.

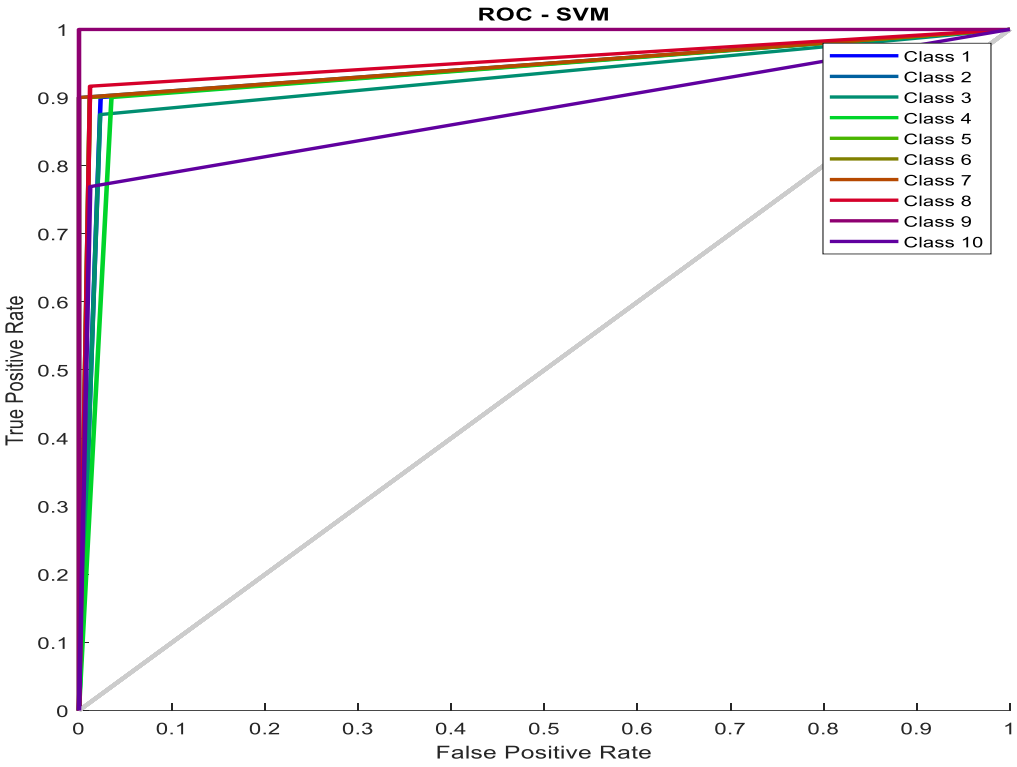


Fig.4. Receiver Operating Characteristics (SVM as a classifier)

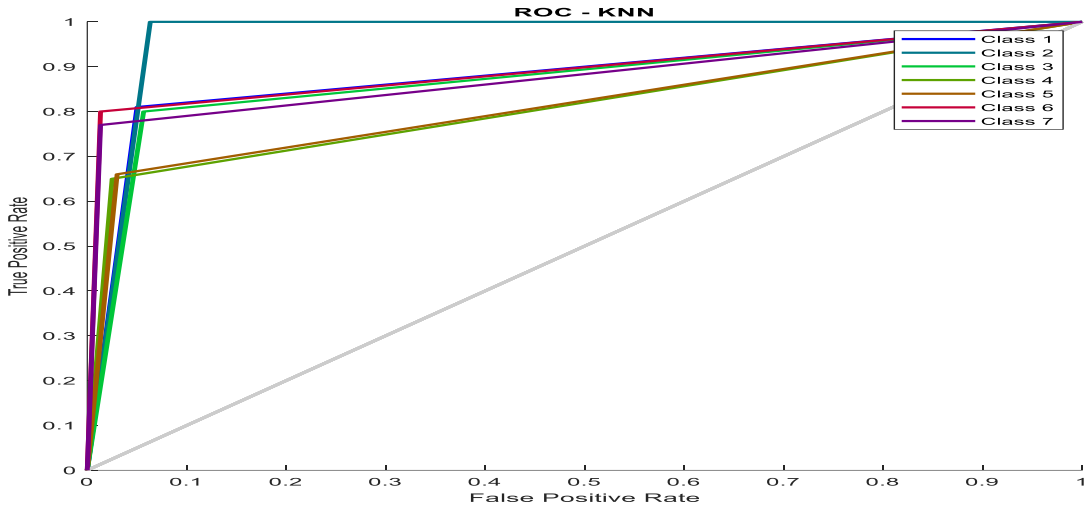
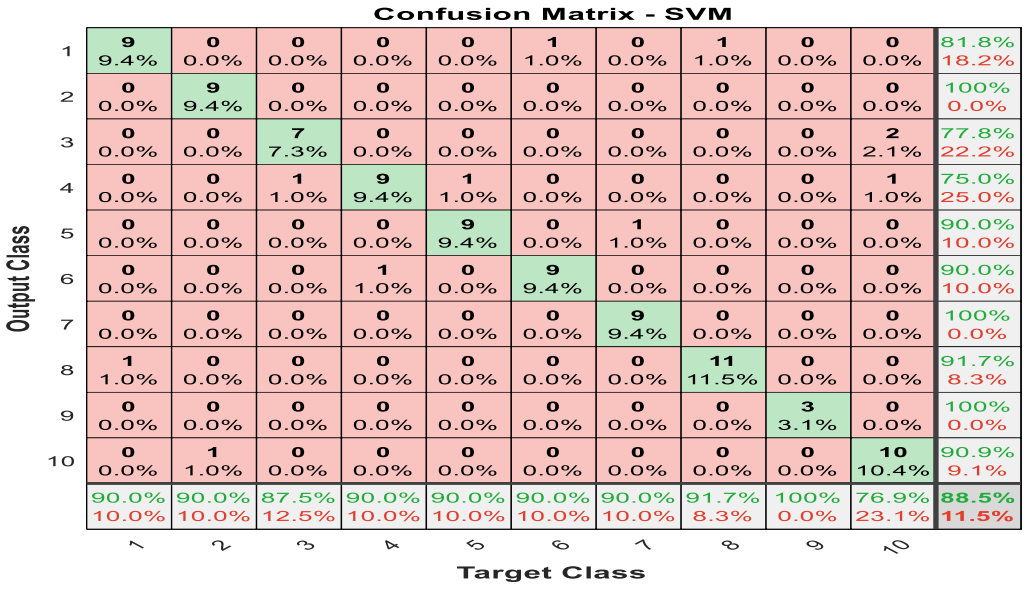


Fig.5. Receiver Operating Characteristics (K-NN as a classifier)

The fig. 4 and 5, describes the receiver operating characteristics (ROC) for each & every class of math expression with the utilization of classifiers such as support vector machine and K-NN have been represented with true positive rate (TPR) and false positive rate (FPR). Each class of equations showing dissimilar true positive rate and false positive rate. TP =True Positive - It is an outcome where the classifier truly forecast the positive class. FP =False Positive – It is an outcome where the classifier erroneous forecast the positive class.



Output Class \ Target Class	1	2	3	4	5	6	7	8	9	10	
1	9 9.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 1.0%	0 0.0%	1 1.0%	0 0.0%	0 0.0%	81.8% 18.2%
2	0 0.0%	9 9.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	7 7.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 2.1%	77.8% 22.2%
4	0 0.0%	0 0.0%	1 1.0%	9 9.4%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 1.0%	75.0% 25.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 9.4%	0 0.0%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	90.0% 10.0%
6	0 0.0%	0 0.0%	0 0.0%	1 1.0%	0 0.0%	9 9.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	90.0% 10.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 9.4%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
8	1 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	11 11.5%	0 0.0%	0 0.0%	91.7% 8.3%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 3.1%	0 0.0%	100% 0.0%
10	0 0.0%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	10 10.4%	90.9% 9.1%
	90.0% 10.0%	90.0% 10.0%	87.5% 12.5%	90.0% 10.0%	90.0% 10.0%	90.0% 10.0%	90.0% 10.0%	91.7% 8.3%	100% 0.0%	76.9% 23.1%	88.5% 11.5%

Fig. 6. Confusion Matrix with SVM.

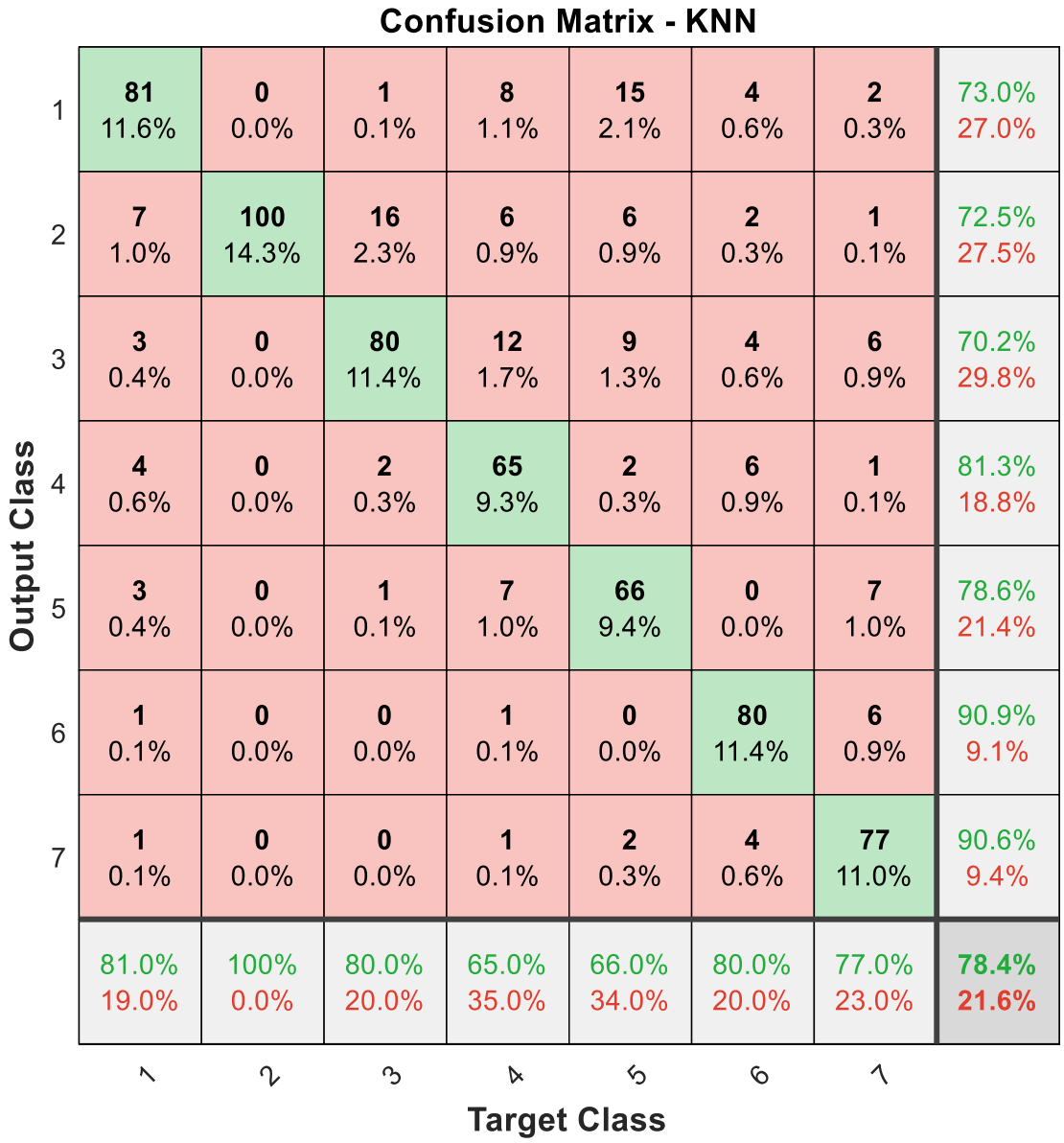


Fig.7. Confusion Matrix with K-NN.

The perception rate is basically relying on the classifier used and features have pulled out. The figure 6 and 7 have explained the confusion matrix with target class and output class with ten classes of equations. As shown in above figure 5 with SVM, it has been concluded that the precision of class nine is highest among the other classes of equations while class ten shows lowest precision among the other classes of equations and in figure 6 with K-NN, it has observed that the exactness of class 6 is more as compared to other classes but exactness of class 3 is too low. The overall accuracy obtained for proposed approach with SVM is 88.5 % and with K-NN is 78.4 %.

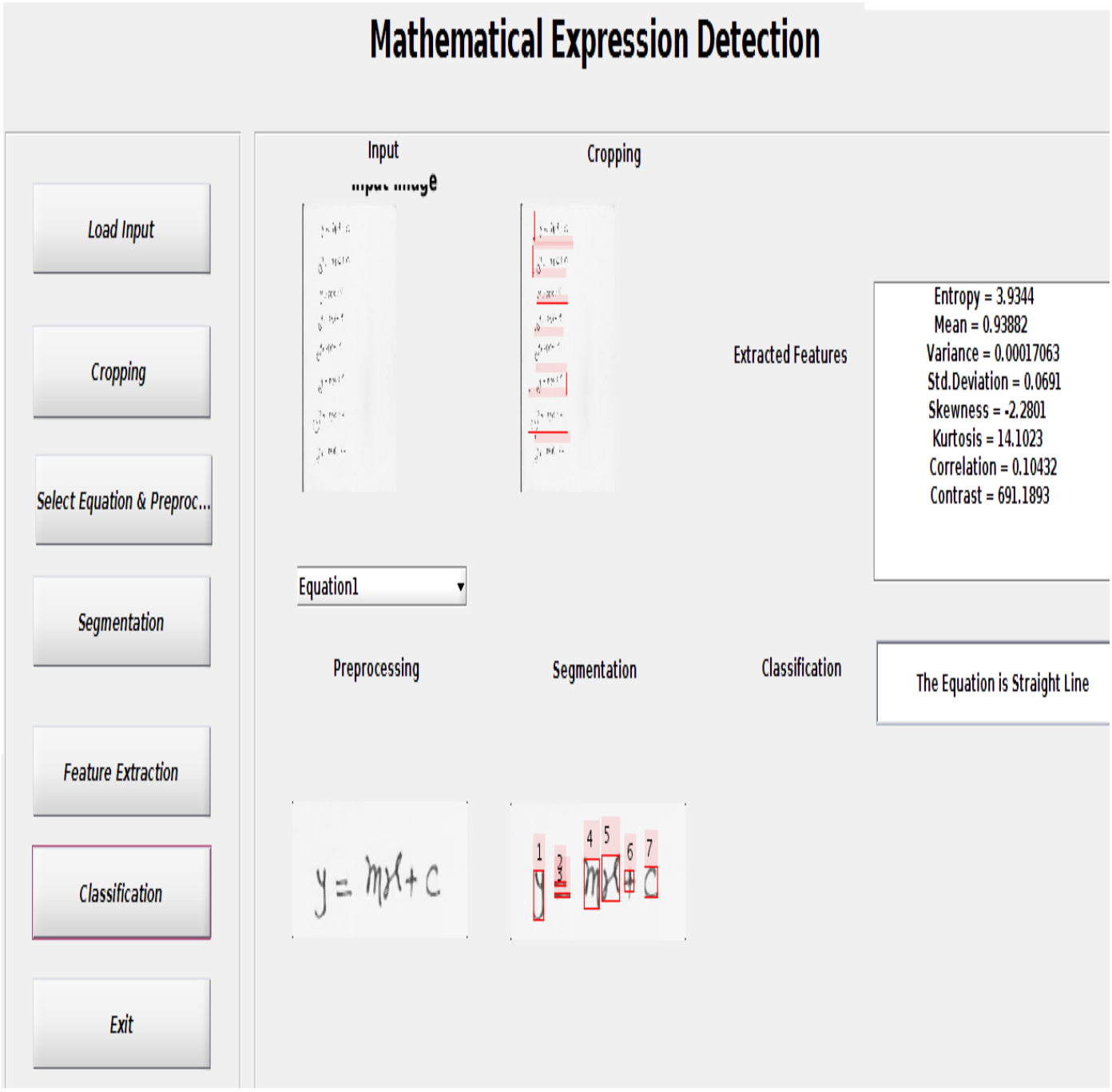


Fig.8. Graphical User Interface (Recognition of math expression)

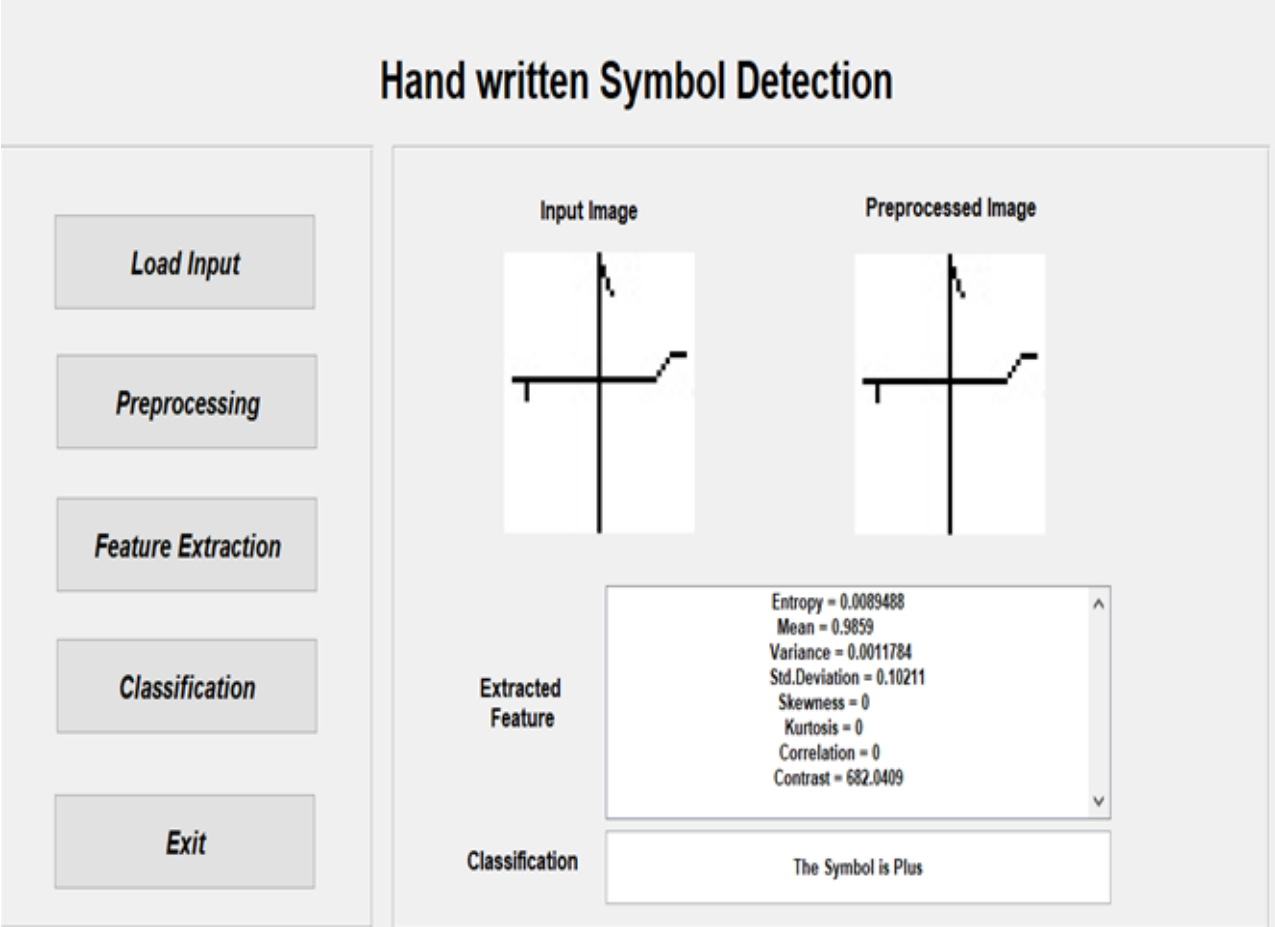


Fig.9. Graphical User Interface (Perception of math symbol)

The graphical user interface (GUI) is shown above in figure 8 and 9, which denotes the flow from loading of the input image through classification to get the identification of math equation & symbol. The bank of handwritten equations or symbols have acquired as a input and cropped it to select particular equation as per choice and the selected equation has undergone through preprocessing to get noise free clean and clear image and then passed through segmentation and finally statistical and complex features are extracted to get classification by using support vector machine & K-NN. The above figure shows the recognition of straight line equation and plus or addition symbol. Through this GUI, it have been noted that throughput and efficiency of the proposed system is remarkable.

5. Conclusions

It has observed that the complicated mathematical expressions as well as symbols have been recognized with remarkable outcomes in terms of efficiency & throughput by extracting the statistical and complex features with the utilization of support vector machine and K-NN as a classifier. The confusion matrix shows that the accuracy and efficiency using support vector machine has been more than that of K-NN. The overall accuracy by utilizing the proposed system is 88.5% and 78.4 % using SVM and K-NN respectively.

References

1. Çelik, Mehmet, and Berrin Yanikoğlu. "Handwritten mathematical formula recognition using a statistical approach." In *Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference on*, pp. 498-501. IEEE, 2011.
2. Álvaro, Francisco, Joan-Andreu Sánchez, and José-Miguel Benedí. "Classification of on-line mathematical symbols with hybrid features and recurrent neural networks." In *Document analysis and recognition (icdar), 2013 12th international conference on*, pp. 1012-1016. IEEE, 2013.
3. Chaturvedi, Soni, Rutika N. Titre, and Neha Sondhiya. "Review of handwritten pattern recognition of digits and special characters using feed forward neural network and Izhikevich neural model." In *Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014 International Conference on*, pp. 425-428. IEEE, 2014.
4. Le, Anh Duc, Truyen Van Phan, and Masaki Nakagawa. "A system for recognizing online handwritten mathematical expressions and improvement of structure analysis." In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pp. 51-55. IEEE, 2014.
5. Dai Nguyen, Hai, Anh Duc Le, and Masaki Nakagawa. "Deep neural networks for recognizing online handwritten mathematical symbols." In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pp. 121-125. IEEE, 2015.
6. Chajri, Yassine, Abdelkrim Maarir, and Belaid Bouikhalene. "A comparative study of handwritten mathematical symbols recognition." In *Computer Graphics, Imaging and Visualization (CGiV), 2016 13th International Conference on*, pp. 448-451. IEEE, 2016.
7. [7] Ashok Kumar, Pradeep Kumar Bhatia, " Offline Handwritten Character Recognition Using Improved Back Propagation Algorithm", International Journal of Advances in Engineering Sciences Vol.3(3), July 2013.
8. [8] Wenhao He, Yuxuan Luo, Fei Yin, Han Hu, Junyu Han, Errui Ding, & Cheng-Lin Liu. (2016), "Context-aware mathematical expression recognition: An end-to-end framework and a benchmark", 2016, IEEE 23rd International Conference on Pattern Recognition (ICPR), 978-1-5090-4847-2, doi:10.1109/icpr.2016.7900135.
9. [9] RatnamalaS.Patil, Shilpa, 2016, "A Novel Method For Handwritten
10. Mathematical Document Based On
11. Equation Symbols Recognition
12. Using K-Nn And Ann Classifiers" , International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 23 Issue 6 –OCTOBER 2016 (SPECIAL ISSUE).
13. [10] Sagar Shinde, R. B. Waghulade, D. S. Bormane, "A new neural network based algorithm for identifying handwritten mathematical equations" , IEEE International Conference on Trends in Electronics and Informatics (ICEI-11-12 May 2017), SCAD COET, Tirunelveli, India.
14. [11] Sagar Shinde, R. B. Waghulade, "An Improved Algorithm For Recognizing Mathematical Equations By Using Machine Learning Approach And Hybrid Feature Extraction Technique", IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE- 27-28 April 2017), M. Kumarasamy College of Engineering, Karur, Tamilnadu, India.