HANDWRITTING RECOGNIZER

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

COMPUTER SCIENCE

Submitted By

BISWARUP BHATTACHARJEE	*University Enrollment No*. 12019009023003
SUBHAJIT PATI	*University Enrollment No.* 12019009023006
RAHUL DEBNATH	*University Enrollment No*. 12019009023001
ARIJIT GOSWAMI	*University Enrollment No*. 12019009001175
SUBHADIP MAJI	*University Enrollment No.* 12019009001345
ANKITA SIKDER	*University Enrollment No.* 12019009023005

GROUP NO. 42

Under the guidance of

PROF. SUDIPTO KUMAR MONDOL

Department of Computer Science

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled **HANDWRITTING RECOGNIZER** submitted by **BISWARUP BHATTACHARJEE (University Roll No. 12019009023003), SUBHAJIT PATI (University Roll No. 12019009023006)**, **RAHUL DEBNATH (University Roll No. 12019009023001)**, **ARIJIT GOSWAMI (University Roll No. 12019009001175)**, **SUBHADIP MAJI(University Roll No. 12019009001345)** and **ANKITA SIKDER(University Roll No.12019009023005)** students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfilment of requirement for the degree of Bachelor of Computer Science, is a bonafide work carried out by them under the supervision and guidance of Prof. SUDIPTO KUMAR MONDOL during 3rd Semester of academic session of 2020 - 2021. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

Prof. Sudipto Kumar Mondol

Assistant Professor

Department of Computer Science

UEM, Kolkata

Prof. Sukalyan Goswami

Head of the Department

Department of Computer Science

UEM, Kolkata

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. **Sudipto Kumar Mondol** of the Department of Computer Science, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

We would also like to express our sincere gratitude to Prof. **Sukalyan Goswami**, HOD, Computer Science, UEM, Kolkata and all other departmental faculties for their ever-present assistant and encouragement.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

BISWARUP BHATTACHARJEE

SUBHAJIT PATI

RAHUL DEBNATH

ARIJIT GOSWAMI

SUBHADIP MAJI

ANKITA SIKDER

TABLE OF CONTENTS

ABSTRACT →

Handwritten character recognition is one of the practically important issues in pattern recognition applications. The applications of digit recognition include postal mail sorting, bank check processing, form data entry, etc. The heart of the problem lies within the ability to develop an efficient algorithm that can recognize handwritten digits and which is submitted by users by the way of a scanner, tablet, and other digital devices. This paper presents an approach to off-line handwritten digit recognition based on different machine learning techniques. The main objective of this paper is to ensure effective and reliable approaches for recognition of handwritten digits. Several machine learning algorithms namely, Multilayer Perceptron, Support Vector Machine, Naïve Bayes, Bayes Net, Random Forest, J48 and Random Tree have been used for the recognition of digits using WEKA. The result of this paper shows that the highest 90.37% accuracy has been obtained for the Multilayer Perceptron.

INTRODUCTION →

Intelligent image analysis is an appealing research area in Artificial Intelligence and also crucial for a variety of present open research difficulties. Handwritten digits recognition is a well-researched subarea within the field that is concerned with learning models to distinguish pre-segmented handwritten digits. It is one of the most important issues in data mining, machine learning, pattern recognition along with many other disciplines of artificial intelligence .The main application of machine learning methods over the last decade has determined efficacious in conforming decisive systems which are competing to human performance and which accomplish far improved than manually written classical artificial intelligence systems used in the beginnings of optical character recognition technology . However, not all features of those specific models have been previously inspected. A great attempt of research workers in machine learning and data mining has been contrived to achieve efficient approaches for approximation of recognition from data . In the twenty-first Century handwritten digit communication has its own standard and most of the times in daily life are being used as means of conversation and recording the information to be shared with individuals. One of the challenges in handwritten characters recognition wholly lies in the variation and distortion of handwritten character set because distinct communities may use diverse styles of handwriting, and control to draw the similar pattern of the characters of their recognized script. Identification of digit from where best discriminating features can be extracted is one of the major tasks in the area of digit recognition system. To locate such regions different kinds of region sampling techniques are used in pattern recognition .The challenge in handwritten character recognition is mainly caused by the large variation of individual writing styles . Hence, robust feature extraction is very important to improve the performance of a handwritten character recognition system. Nowadays handwritten digit recognition has obtained a lot of concentration in the area of pattern recognition system sowing its application in diverse fields. In the next few days, a character recognition system might serve as a cornerstone to initiate paperless surroundings by digitizing and processing existing paper documents. Handwritten digit datasets are vague in nature because there may not always be sharp and perfectly straight lines. The main goal in digit recognition is feature extraction is to remove the redundancy from the data and gain a more effective embodiment of the word image through a set of numerical attributes. It deals with extracting

most of the essential information from image raw data. In addition the curves are not necessarily smooth like the printed characters. Furthermore, characters dataset can be drawn in different sizes and the orientation which are always supposed to be written on a guideline in an upright or downright point. Accordingly, an efficient handwritten recognition system can be developed by considering these limitations. It is quite exhausting sometimes to identify handwritten characters as it can be seen that most of the human beings can't even recognize their own written scripts. Hence, there exists a constraint for a writer to write apparently for recognition of handwritten documents. Before revealing the method used in conducting this research, a software engineering module is first presented. Pattern recognition along with Image processing plays a compelling role in the area of handwritten character recognition. The study , describes numerous types of classification of feature extraction techniques like structural feature based methods, statistical feature based methods and global transformation techniques. Statistical approaches are established on planning of how data are selected. It utilizes the information of the statistical distribution of pixels in the image. The paper provided an SVM based offline handwritten digit recognition system. Authors claim that SVM outperforms in the experiment. Experiment is carried out on NIST SD19 standard dataset. The study provides the conversion of handwritten data into electronic data, nature of handwritten characters and the neural network approach to form a machine competent of recognizing handwritten characters. The study addresses a comprehensive criterion of handwritten digit recognition with various state of the art approaches, feature representations, and datasets. However, the relationship of training set size versus accuracy/error and the dataset-independence of the trained models are analyzed. The paper presents convolution neural networks into the handwritten digit recognition research and describes a system which can still be considered state of the art.

LITERATURE SURVEY →

An early notable attempt in the area of character recognition research is by Grimsdale in 1959. The origin of a great deal of research work in the early sixties was based on an approach known as analysis-by-synthesis method suggested by Eden in 1968. The great importance of Eden's work was that he formally proved that all handwritten characters are formed by a finite number of schematic features, a point that was implicitly included in previous works. This notion was later used in all methods in syntactic (structural) approaches of character recognition. K. Gaurav, Bhatia P. K. Etal, this paper deals with the various pre-processing techniques involved in the character recognition with different kinds of images ranging from a simple handwritten form based documents and documents containing colored and complex background and varied intensities. In this, different preprocessing techniques like skew detection and correction, image enhancement techniques of contrast stretching, binarization, noise removal techniques, normalization and segmentation, morphological processing techniques are discussed. It was concluded that using a single technique for preprocessing, we can't completely process the image. However, even after applying all the said techniques might not be possible to achieve full accuracy in a preprocessing system. Salvador España-Boquera et al , in this paper hybrid Hidden Markov Model (HMM) model is proposed for recognizing unconstrained offline handwritten texts. In this, the structural part of the optical model has been modelled with Markov chains, and a Multilayer Perceptron is used to estimate the emission probabilities. In this paper, different techniques are applied to remove slope and slant from handwritten text and to normalize the size of text images with supervised learning methods. The key features of this recognition system were to develop a system having high accuracy in preprocessing and recognition, which are both based on ANNs. In, a modified quadratic classifier based scheme to recognize the offline handwritten numerals of six popular Indian scripts is proposed. Multilayer perceptron has been used for recognizing

Handwritten English characters . The features are extracted from Boundary tracing and their Fourier Descriptors. The character is identified by analysing its shape and comparing its features that distinguish each character. Also an analysis has been carried out to determine the number of hidden layer nodes to achieve high performance of the back propagation network. A recognition accuracy of 94% has been reported for Handwritten English characters with less training time. In [9], diagonal feature extraction has been proposed for offline character recognition. It is based on the ANN model. Two approaches using 54 features and 69 features are chosen to build this Neural Network recognition system. To compare the recognition efficiency of the proposed diagonal method of feature extraction, the neural network recognition system is trained using the horizontal and vertical feature extraction methods. It is found that the diagonal method of feature extraction yields the recognition accuracy of 97.8 % for 54 features and 98.5% for 69 features. A. Brakensiek, J. Rottland, A. Kosmala, J. Rigollet al, in this paper a system for off-line cursive handwriting recognition is described which is based on Hidden Markov Models (HMM) using discrete and hybrid modelling techniques. Handwriting recognition experiments using a discrete and two different hybrid approaches, which consist of a discrete and semi-continuous structures, are compared. A segmentation free approach is considered to develop the system. It is found that the recognition rate performance can be improved of a hybrid modelling technique for HMMs, which depends on a neural vector quantizer (hybrid MMI), compared to discrete and hybrid HMMs, based on tired mixture structure (hybrid - TP), which may be caused by a relative small data set. R. Bajaj, L. Dey, S. Chaudharietal , employed three different kinds of features, namely, the density features, moment features and descriptive component features for classification of Devanagari Numerals. They proposed multi classifier connectionist architecture for increasing the recognition reliability and they obtained 89.6% accuracy for handwritten Devanagari numerals. Sandhya Arora in , used four feature extraction techniques namely, intersection, shadow feature, chain code histogram and straight line fitting features. Shadow features are computed globally for character image while intersection features, chain code histogram features and line fitting features are computed by dividing the character image into different segments. On experimentation with a dataset of 4900 samples the overall recognition rate observed was 92.80% for Devanagari characters. Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al Khawaldeh et al, this paper describes that Recognition of characters greatly depends upon the features used. Several features of the handwritten Arabic characters are selected and discussed. An off-line recognition system based on the selected features was built. The system was trained and tested with realistic samples of handwritten Arabic characters. Evaluation of the importance and accuracy of the selected features is made. The recognition based on the selected features give average accuracies of 88% and 70% for the numbers and letters, respectively. Further improvements are achieved by using feature weights based on insights gained from the accuracies of individual features.Sushree Sangita Patnaik and Anup Kumar Panda May 2011 et al, this paper proposes the implementation of particle swarm optimization (PSO) and bacterial foraging optimization (BFO) algorithms which are intended for optimal harmonic compensation by minimizing the undesirable losses occurring inside the APF itself. The efficiency and effectiveness of the implementation of two approaches are compared for two different conditions of supply. The total harmonic distortion (THD) in the source current which is a measure of APF performance is reduced drastically to nearly 1% by employing BFO. The results demonstrate that BFO outperforms the conventional and PSO based approaches by ensuring excellent functionality of APF and quick prevalence over harmonics in the source current even under unbalanced supply. In literature, T. Som have discussed fuzzy membership function based approaches for HCR. Character images are normalized to 20 X 10 pixels. Average image (fused image) is formed from 10 images of each character. Bounding box around character is determined by using vertical and horizontal projection of character. After cropping the image to the bounding box, it is resized to 10 X 10 pixels size. After that, things are performed and a thinned image is placed in one by one row of 100 X 100 canvas. Similarity score of test image is matched with fusion image and characters are classified. In [16], Renata F. P. Neves has proposed SVM based offline handwritten digit recognition. Authors claim that SVM outperforms the Multilayer perceptron classifier. Experiment is carried out on NIST SD19 standard dataset. Advantage of MLP is that it is able to segment non-linearly separable classes. However, MLP can easily fall into a region of local minimum, where the training will stop assuming it has achieved an optimal

point in the error surface. Another hindrance is defining the best network architecture to solve the problem, considering the number of layers and the number of perceptrons in each hidden layer. Because of these disadvantages, a digit recognizer using the MLP structure may not produce the desired low error rate. G. Pirlo and D. Impedovo in his work on , presented a new class of membership functions, which are called Fuzzy Membership functions (FMFs), for zoning-based classification. These FMFs can be easily adapted to the specific characteristics of a classification problem in order to maximize classification performance. In this research, a real coded genetic algorithm is presented to find, in a single optimization procedure, the optimal FMF, together with the optimal zoning described by Voronoi tessellation. The experimental results, which are carried out in the field of handwritten digit and character recognition, indicate that optimal FMF performs better than other membership functions based on abstract level, ranked-level, and measurement-level weighting models, which can be found in the literature. Yoshimasa Kimura presented a work on how to select features for Character Recognition Using Genetic Algorithms. The author proposes a novel method of feature selection for character recognition using genetic algorithms (GA). The proposed method selects only the genes for which the recognition rate of training samples exceeds than the predetermined threshold as a candidate of the parent gene and adopts a reduction ratio in the number of features used for recognition as the fitness value. Nafiz Arica at al.proposed a method which avoids most of the pre-processing operations, which causes loss of important information. One of the major contributions of the method is to development of a powerful segmentation algorithm. Utilization of the character boundaries, local maxima and minima, slant angle, upper and lower baselines, stroke height and width, and ascenders and descenders improve the search algorithm of the optimal segmentation path, applied on a gray-scale image. This approach decreases the over-segmentation. Another contribution is the use of Hidden Markov Models (HMM) training, not only for the estimation of model parameters, but also for the estimation of some global and feature space parameters. Also, HMM probabilities are used to measure the shape information and rank the candidate character. One dimensional representation of a two dimensional character image increases the power of the HMM shape recognizer. M. Hanmandlu, O.V. Ramana Murthy have presented in their study the recognition of handwritten Hindi and English numerals by representing them in the form of exponential membership functions which serve as a fuzzy model. The recognition is carried out by modifying the exponential membership functions fitted to the fuzzy sets. These fuzzy sets are derived from features consisting of normalized distances obtained using the Box approach. The membership function is modified by two structural parameters that are estimated by optimizing the entropy subject to the attainment of membership function to unity. The overall recognition rate is found to be 95% for Hindi numerals and 98.4% for English numerals. In, a method to construct a handwritten Tamil character by executing a sequence of strokes is proposed. A structure or shape-based representation of a stroke was used in which a stroke was represented as a string of shape features. Using this string representation, an unknown stroke was identified by comparing it with a database of strokes using a flexible string matching procedure. A full character was recognized by identifying all the component strokes.

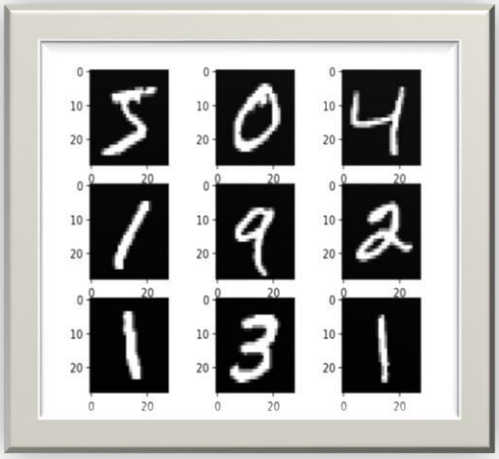
PROBLEM STATEMENT →

The goal of this project is to create a model that will be able to recognize and determine the handwritten digits from its image by using the concepts of Convolution Neural Network and Optical Character recognition with tesseract-OCR. Though the goal is to create a model which can recognize the digits, it can be extended to letters and an individual's handwriting. The major goal of the proposed system is understanding Convolutional Neural Network, and applying it to the handwritten recognition system. The another goal of this project is that make a simple user friendly graphical user interface to handle recognizing handwritten digits and OCR, create a developers section in graphical user interface where the all group members details.

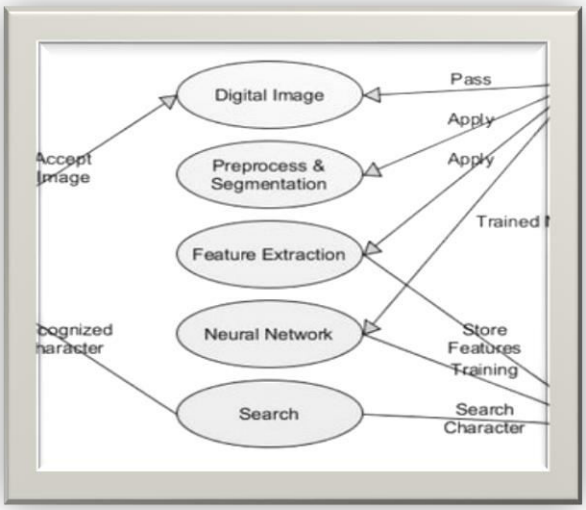
PROPOSED SOLUTION →

DATASET - MNIST database of handwritten digits is used as a dataset. It consists of a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image of 28*28 pixels (784 pixels).

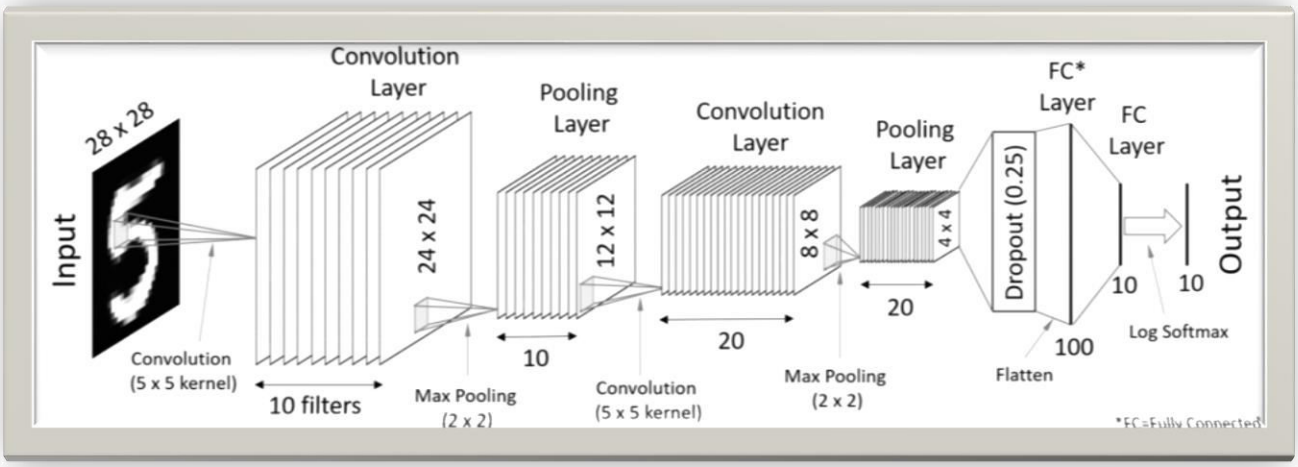
Modified National Institute of Standards and Technology (MNIST) is a large set of computer vision dataset which is extensively used for training and testing different systems. It was created from the two special datasets of the National Institute of Standards and Technology (NIST) which holds binary images of handwritten digits. The training set contains handwritten digits from 250 people, among them 50% training dataset was employees from the Census Bureau and the rest of it was from high school students. However, it is often attributed as the first datasets among other datasets to prove the effectiveness of the neural networks.



MNIST handwritten digit dataset



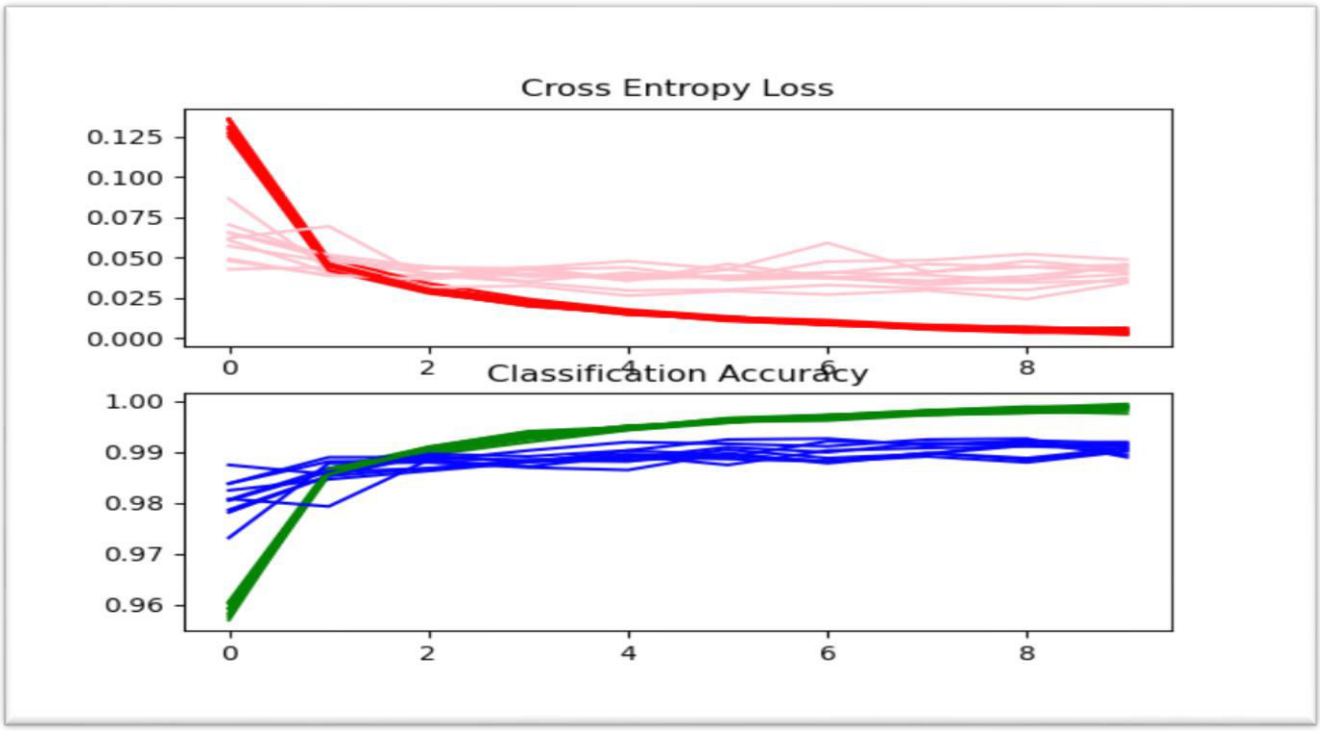
Steps of recognizing handwritten digits



CNN – ANALYSIS

MODEL VISUALIZATION:

We will add a double convolutional layer with 64 filters each, followed by another max pooling layer. A plot of the learning curves is created by matplotlib package in python, in this case showing that the models still have a good fit on the problem, with no clear signs of overfitting.

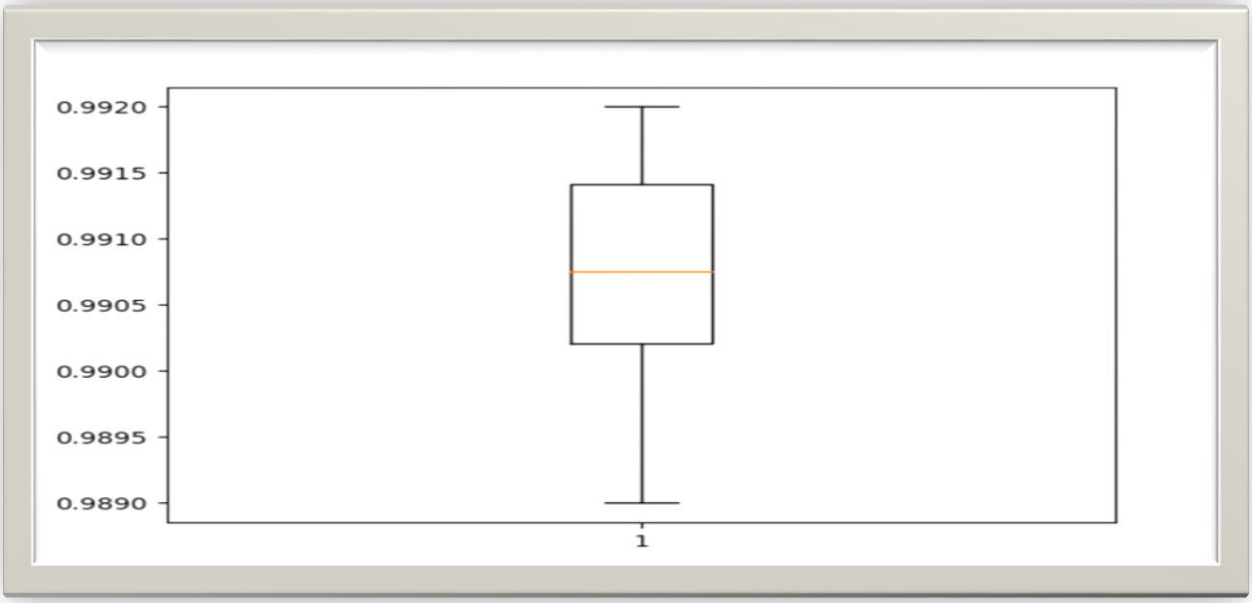


Consider running the example a few times and compare the average outcome, during model evaluation n folds is 10 so we get total 10 outcomes of our model accuracy:

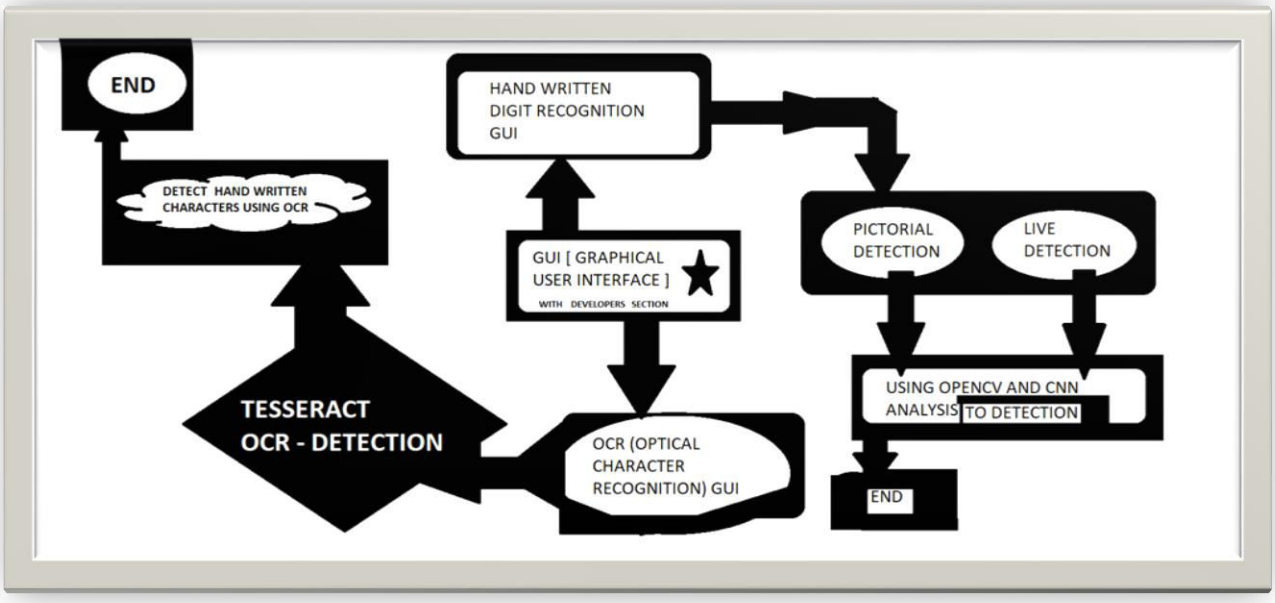
EPOCHS	BATCH SIZE	ACCURACY
10	32	98.950%
10	32	99.117%
10	32	98.900%
10	32	99.033%
10	32	99.017%
10	32	99.0.83%
10	32	99.127%
10	32	99.200%
10	32	99.150%
10	32	99.067%

The estimated performance of the model is presented, showing performance with a slight changing in the mean accuracy of the model: **99.30%.**

BOX PLOT FOR ACCURACY SCORES:



EXPERIMENTAL SETUP AND RESULT ANALYSIS →



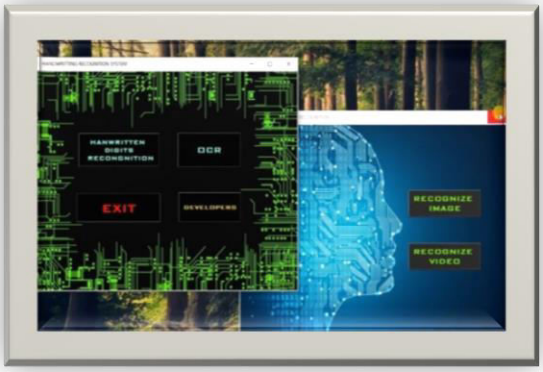
SIMPLE BLOCK DIAGRAM OF OUR PROJECT

Result Analysis - Our network has been trained with and tested with 70,000 datasets and we obtained an accuracy of > 99% which is good enough to test our classification implementation. We have given a learning rate

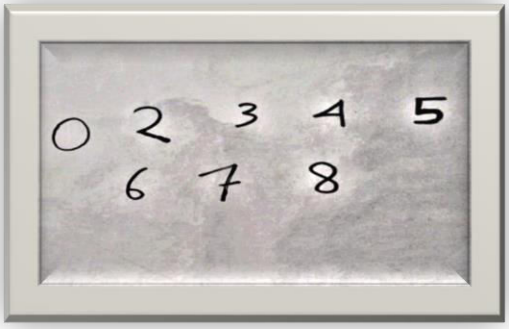
of 0.01 to our algorithm and obtained good classification results, every time after training we are taking random inputs from our testing dataset and calculating the efficiency each time it is executed. An interesting property of these networks is that the training error keeps decreasing over time but the test error goes through a minimum and starts increasing after a certain number of iterations , this is possibly because of the higher learning rate and by decreasing it we can get our results , if not reduced the learning rate the Stochastic gradient descent may get stuck in local minimum and finds it difficult to predict the optimized weights , which affects the prediction and accuracy of our network. The figure 5 shows how they change when the learning rate is high. In our discussion we can refer to other methods and their accuracy although all methods did well with all the classifiers, boosted LeNet 4 did best , achieving a score of 0.7% and the rest of them acquired better accuracy than other methods . So it is best to rely on LeNet architecture rather than other methods for classification. And tesseract-OCR is used to Optical character detect. Total recognizing based on machine learning CNN analysis. Through simple user friendly graphical interface our project can be executed.

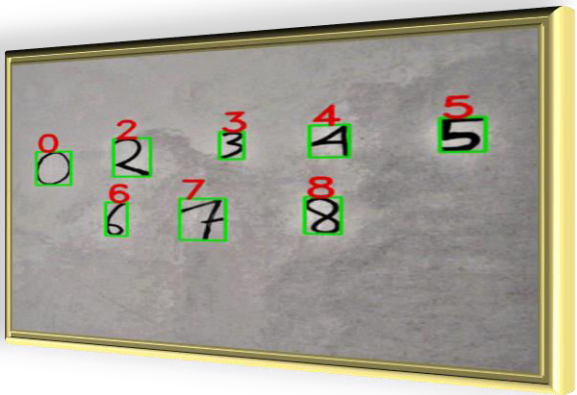
EXPERIMENTAL SCREENSHOTS:

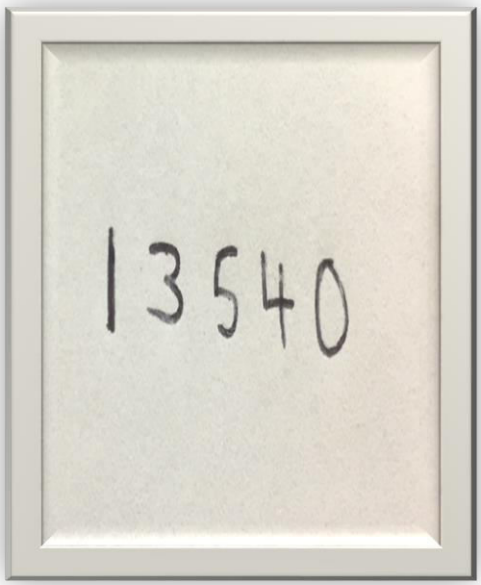
1st GUI-PART:

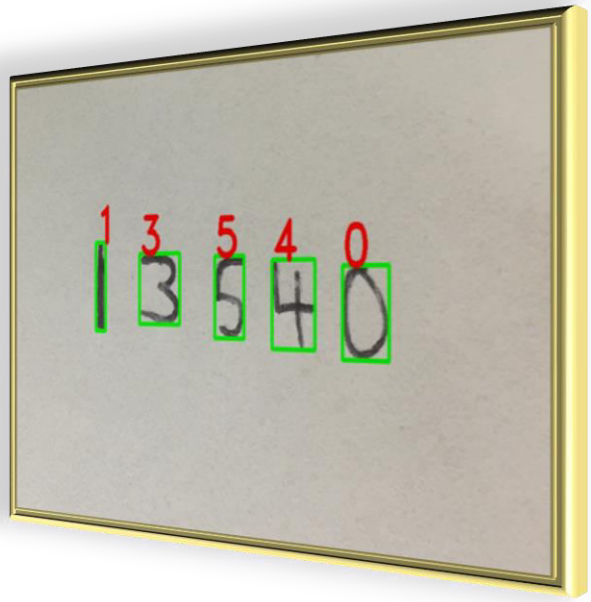


2nd RECOGNITION PART:

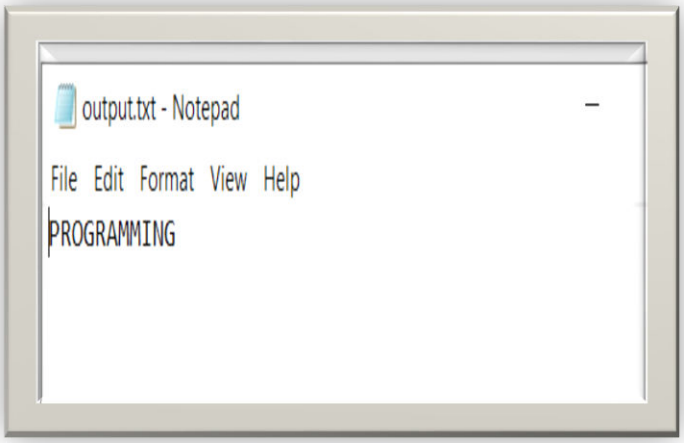








3rd OCR PART:



CONCLUSION AND FUTURE SCOPE →

Performance of a network depends on many factors like low memory requirements, low run time and better accuracy , although in this paper it is primarily focused on getting better accuracy rate for classification . Before Artificial neurons had better accuracy but now the branch of computer vision mainly depends on deep learning features like convolutional neural networks. Research is still going on in this field and researches have developed many forms of LeNet architecture like LeNet-1,LeNet-4, Boosted LeNet-4 and also combination of many methods like LeNet-4 with KNN's , but for a quite long time our LeNet architecture was considered as state of the art.Many other methods like Tangent Distance Classifier were developed using LeNet architecture. The main aim of this paper deals with one of the methods in which it can be implemented , there are several methods in which they can be done and using different frameworks like matlab,octave. The branch of computer vision in artificial intelligence primary motive is to develop a network which is better to every performance measure and provide

results for all kinds of datasets which can be trained and trained and recognized. Fixed size Convolutional Neural Networks has been applied to many applications like handwritten digit recognition , machine printed character recognition and on-line handwriting recognition, they can also be useful for signature verification .The more the training examples the more is the accuracy of the networks .Unsupervised machine learning was made easier using Convolutional Neural networks , some of the future works possible to implement by CNN's are compressing or obtaining same results from smaller networks by optimization tricks , more invariant feature learning such that the input images doesn't gets distorted. The major 3D vision networks is a scope for researchers to develop using LeNet architecture and more biologically concordant methods , a hope for the future is that Unsupervised CNN's .

BIBLIOGRAPHY →

1.http://yann.lecun.com/exdb/mnist

2.http://athena.ecs.csus.edu/~shahrb/ProjectProposalN1.pdf

3.https://pdfs.semanticscholar.org/6d9c/b8ff588f5a49a0d00b095d93d5f6d2f14771.pdf

4.http://cv-tricks.com/tensorflow-tutorial/training-convolutional-neural-network-for-imageclassification/

5. http://ijarcet.org/wp-content/uploads/IJARCET-VOL-6-ISSUE-7-990-997.pdf