Indian Institute of Information Technology Allahabad

Department of Information Technology

Mini Project 2021

Handwritten and Machine Printed Text Detection

Under supervision of Dr. Pavan Chakraborty & Dr. Sonali Agrawaal

As Presentation of C2 by

Enrollment No.	Name	GitHub ID.
IIT2019239	**Mrityunjaya Tiwari**	**Error404m**
IIT2019222	**Rauank Singh Rathore**	**Error404r**
IIB2019006	**Amanjeet Kumar**	**Amanjeetk11**
IIT2019202	**Jyoti Verma**	**Jyo123-verma**

1.Abstract:

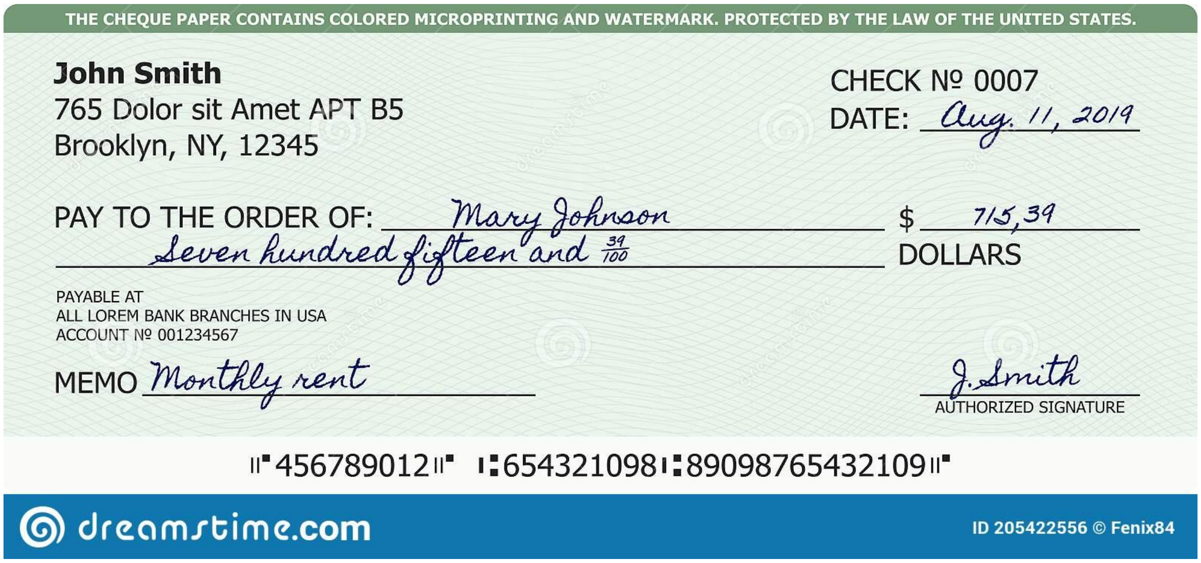
In this project , we are working to archive proper documents or books, starting from sets of short documents with annotations . It is necessary to separate the machine printed and handwritten text before applying different recognition methodologies to each. Because a document image having both types of texts (machine printed and handwritten text) may give rise to significant issues within a digitisation and recognition pipeline. In this project we will strive to figure out handwritten text from machine printed text using the Optical Character Recognition (OCR).

Keywords: Printed text · Handwritten text · OCR · Shape-based features · GUI

2.Introduction:

Nowadays, one can observe a rapidly growing number of digitization initiatives in libraries and archives, involving a variety of document types. The presence of printed and handwritten text in the same document image gives rise to significant issues since each modality requires different treatment to recognize the corresponding characters.

In many samples we come across, we can see the presence of horizontal and vertical lines and that tabular representation of information. To overcome the complexities like horizontal or vertical lines, tabular representation of data, existence of texts across both sides of a line, a simple but effective image-processing approach, which can capture a large majority of texts word-wise, irrespective of their orientations and nature, within a document is presented. It is a time-saving approach and can be widely used in real-life scenarios.



THE CHEQUE PAPER CONTAINS COLORED MICROPRINTING AND WATERMARK. PROTECTED BY THE LAW OF THE UNITED STATES.

John Smith
765 Dolor sit Amet APT B5
Brooklyn, NY, 12345

CHECK № 0007
DATE: Aug. 11, 2019

PAY TO THE ORDER OF: Mary Johnson $ 715,39
Seven hundred fifteen and 39/100 DOLLARS

PAYABLE AT
ALL LOREM BANK BRANCHES IN USA
ACCOUNT № 001234567

MEMO Monthly rent

J. Smith
AUTHORIZED SIGNATURE

⑈456789012⑈ ⑆654321098⑆89098765432109⑈

We can see the presence of ‘mixed text’ regions i.e. the juxtaposition of the printed text and the handwritten text. Therefore, keeping all of the conditions in mind, a method which can provide high accuracy is needed both in case of detection and classification of texts present in the documents. The proposed work has its own impacts. For example, such work can be used
(i) to localize as well as analyze information in a document resulting in faster information retrieval
(ii) to classify the printed text that can help in better keyword extraction mechanisms.

Extraction of a very less number (five) of simple features, upon which, using a machine learning based approach, we find those many features are enough to identify the following from the documents:

1) pure handwritten texts
2) non-texts which are localized by the previous approach
3) pure printed texts
4) mixed texts which arise due to the congested juxtaposition of printed and handwritten texts and exist less in number.

3.Literature Survey:

Though researchers have been working on document image processing, either handwritten or printed, since long, but to be specific, the domain of handwritten and printed text classification in the same document is almost unexplored. Here, we have briefed some of the past works found in the literature.

In the work by Banerjee, the classification between handwritten and machine-printed text is done on the basis of structural and statistical differences between these two kinds of texts, and a tainted text is put into any of the three categories: purely printed, purely handwritten or “mixed” (where parts or whole handwritten and printed texts are very close or overlap on each other).

Jindal et al propose the use of various characteristics of the texts to classify them as printed or handwritten with the claim of being advantageous compared to other classifiers in terms of efficiency, accuracy, simplicity and the low computation complexity. The various steps involved in classifying the texts as handwritten or printed are pre-processing, segmentation at character level, feature extraction from the characters and the classification of the same .
The work proposed by Malakar et al. classifies the printed and handwritten word images in a document image where both types of the texts are present. The word

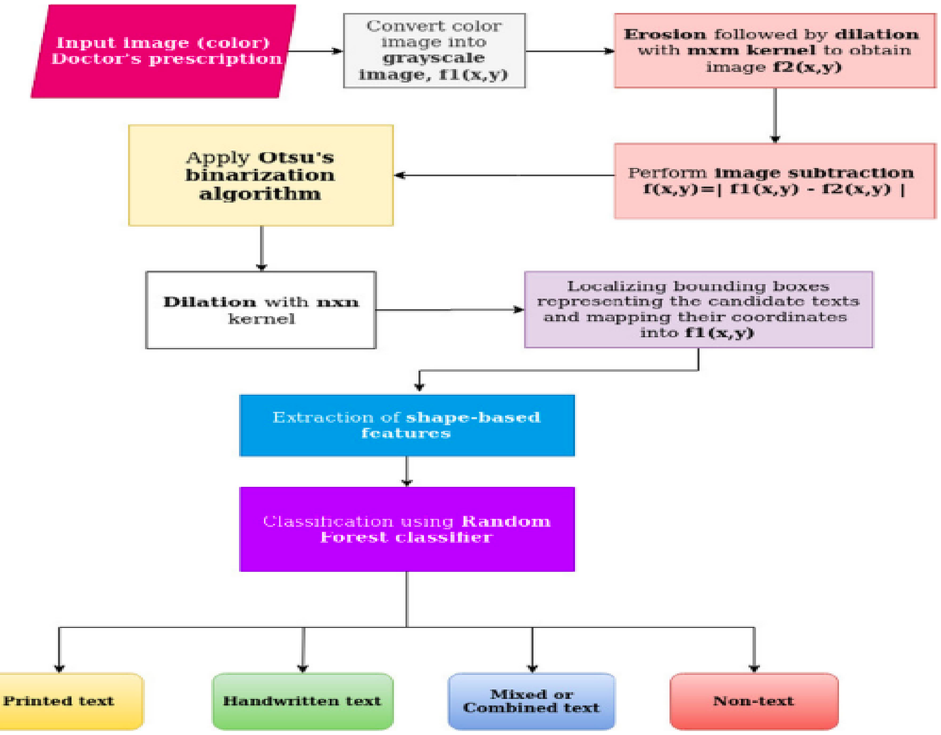
images of the two aforesaid types - handwritten and printed, are collected from various sources. Then a 6-element feature set for each word image is designed and the features are ranked. Depending upon the ranked features, a tree-like classifier has been designed and accordingly the unknown words are fed to assess the performance of the designed classifier.

4.The Proposed Approach:

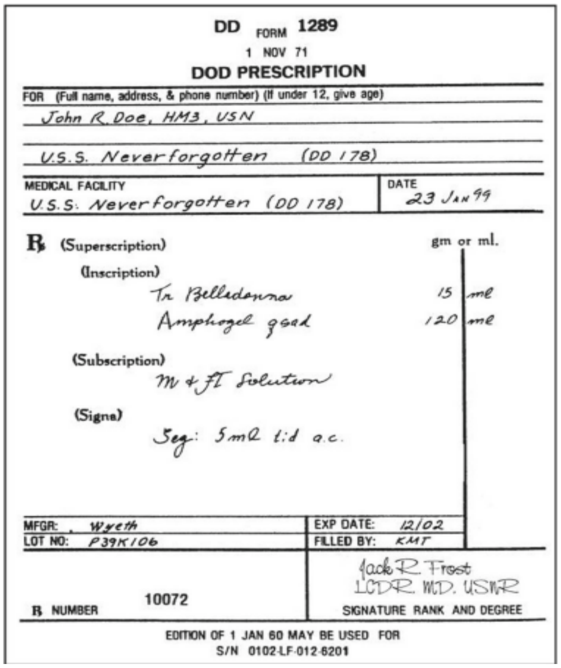
The proposed methodology consists of two stages. The first stage is to localize the possible text regions from the document images, and the second stage is to classify the localized portions of the image as handwritten, printed, non-text or in a few cases, mixed/combined text using the features extracted from the images of word or word-like segments. Most of the documents consist of horizontal and vertical lines, or tabular representation of data.That is why we remove the lines from the images followed by determining the position of the texts in the first stage. Flowchart describing the entire work .

- The first stage is to localize the possible text regions from the document images, and
- The second stage is to classify the localized portions of the image as handwritten, printed, non-text or in a few cases, mixed/combined text using the features extracted from the images of word or word-like segments.

Fig. Flowchart describing the key modules of the proposed method


Input image (color) Doctor's prescription
Convert color image into grayscale image, f1(x,y)
Erosion followed by dilation with mxm kernel to obtain image f2(x,y)
Perform image subtraction f(x,y)=| f1(x,y) - f2(x,y) |
Apply Otsu's binarization algorithm
Dilation with nxn kernel
Localizing bounding boxes representing the candidate texts and mapping their coordinates into f1(x,y)
Extraction of shape-based features
Classification using Random Forest classifier
Printed text
Handwritten text
Mixed or Combined text
Non-text

Step 1: The image is first converted into a single channel grayscale image f1(x,y). This given figure shows a sample document image converted into a grayscale.



DD FORM 1289
1 NOV 71
DOD PRESCRIPTION

FOR (Full name, address, & phone number) (If under 12, give age)
John R. Doe, HM3, USN
U.S.S. Neverforgotten (DD 178)

MEDICAL FACILITY
U.S.S. Neverforgotten (DD 178)

DATE
23 Jan 99

℞ (Superscription) gm or ml.

(Inscription)

Tr Belladonna 15 ml

Amphogel qsad 120 ml

(Subscription)

M & Ft Solution

(Signa)

Sig: 5ml t.i.d a.c.

MFGR: Wyeth
LOT NO: P39K106

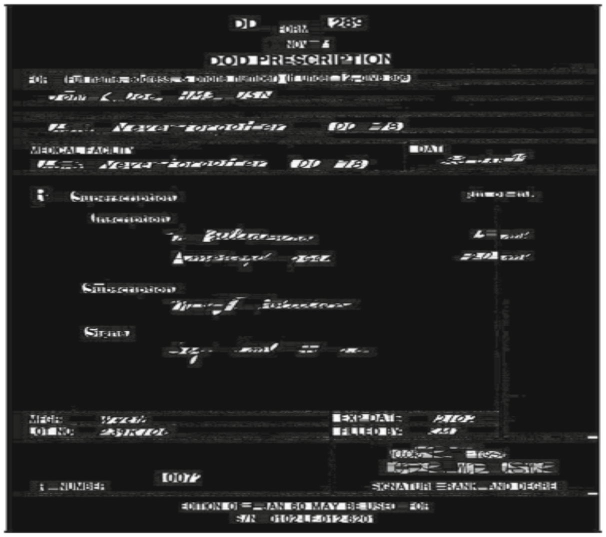
EXP DATE: 12/02
FILLED BY: KMT

Jack R. Frost
LCDR. MD. USNR
SIGNATURE RANK AND DEGREE

℞ NUMBER 10072

EDITION OF 1 JAN 60 MAY BE USED FOR
S/N 0102-LF-012-6201

Step 2: Upon, determining the difference of the corresponding pixel values of the grayscale image f1(x,y) and the resultant image f2(x,y), a new image is obtained as output which contains lower-valued pixels for the strokes part, and lower valued pixels for the background part because the background has higher pixel values for both of the two operand parts . As a result, the lines are removed from the text. Given figure shows the output of the image produced after the image subtraction procedure.



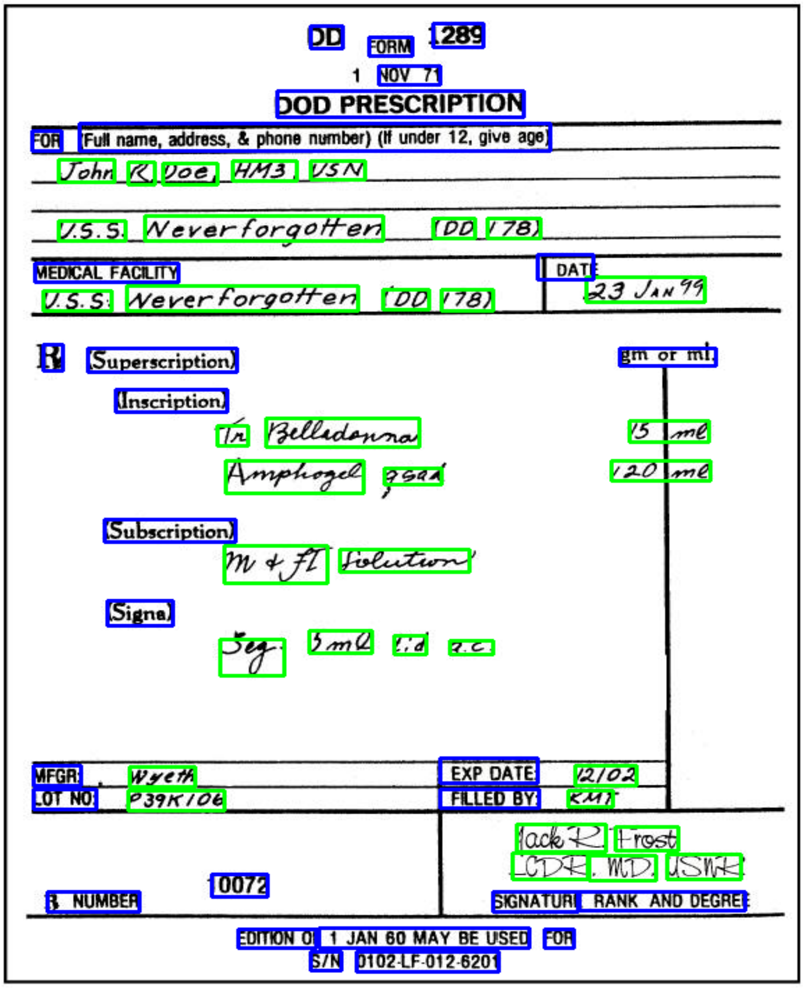
Step 3: Now Otsu’s binarization is performed on the output image, followed by dilation operation with a square structuring element c(x,y) having size 5×5. It is done in order to reshape the patches to get them in their actual size and also to remove the unnecessary dots and points scattered throughout the image . The resultant image after performing Otsu’s binarization is shown in this figure.

Step 4: Now, the bounded box of each of the patches is determined and these parts are cropped out from the original image, and five shape-based parameters are calculated from that cropped binarized image. These features are defined as follows.

- Height (hp)
- Width (wp)
- Aspect Ratio (ar)
- Height Text Ratio (HTR)
- Transition Count Ratio (TCR)

After the classification is performed, the final output of the sample document image is obtained as shown in Fig:

Final Output



DD FORM 1289
1 NOV 71
DOD PRESCRIPTION

FOR (Full name, address, & phone number) (If under 12, give age)
John R. Doe, HM3, USN
U.S.S. Neverforgotten (DD 178)

MEDICAL FACILITY
U.S.S. Neverforgotten (DD 178)

DATE
23 JAN 99

℞ (Superscription)

gm or ml

(Inscription)
Tr Belladonna 15 ml
Amphogel qsad 120 ml

(Subscription)
M & Ft Solution

(Signa)
Sig. 5 ml t.i.d a.c.

MFGR: Wyeth
LOT NO: P39K106
EXP DATE 12/02
FILLED BY: KMT

℞ NUMBER 10072

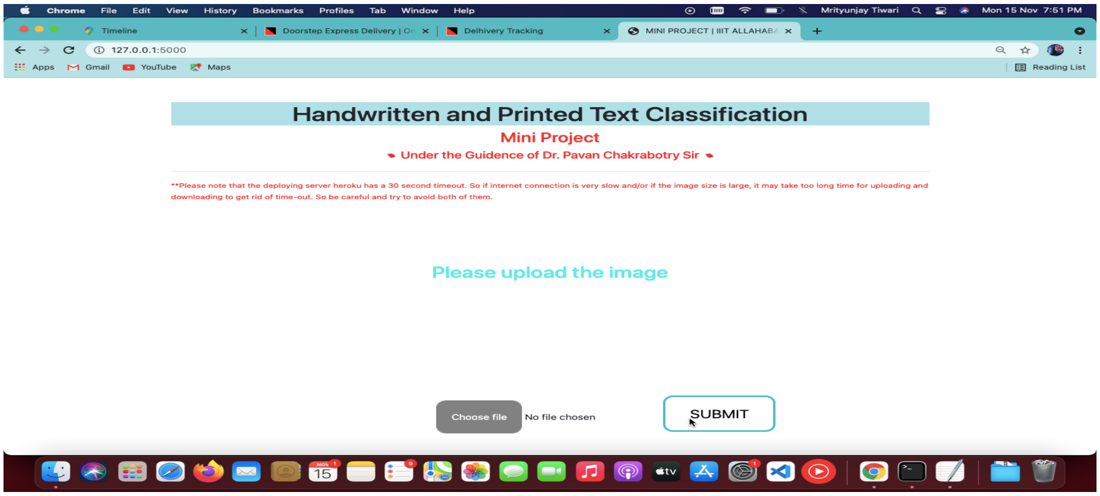
Jack R. Frost
LCDR, MD, USNR

SIGNATURE RANK AND DEGREE

EDITION OF 1 JAN 60 MAY BE USED FOR
S/N 0102-LF-012-6201

Fig. Final output image classifying the printed texts (shown in blue color),handwritten texts (shown in green color) and non-texts (shown in yellow color).

Final Step: After training the model, we have deployed our model on a flask server . This is the UI screen for deployed model:


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Mrityunjay Tiwari
Mon 15 Nov 7:51 PM
Timeline
Doorstep Express Delivery | On
Delhivery Tracking
MINI PROJECT | IIIT ALLAHABA
127.0.0.1:5000
Apps Gmail YouTube Maps
Reading List

Handwritten and Printed Text Classification

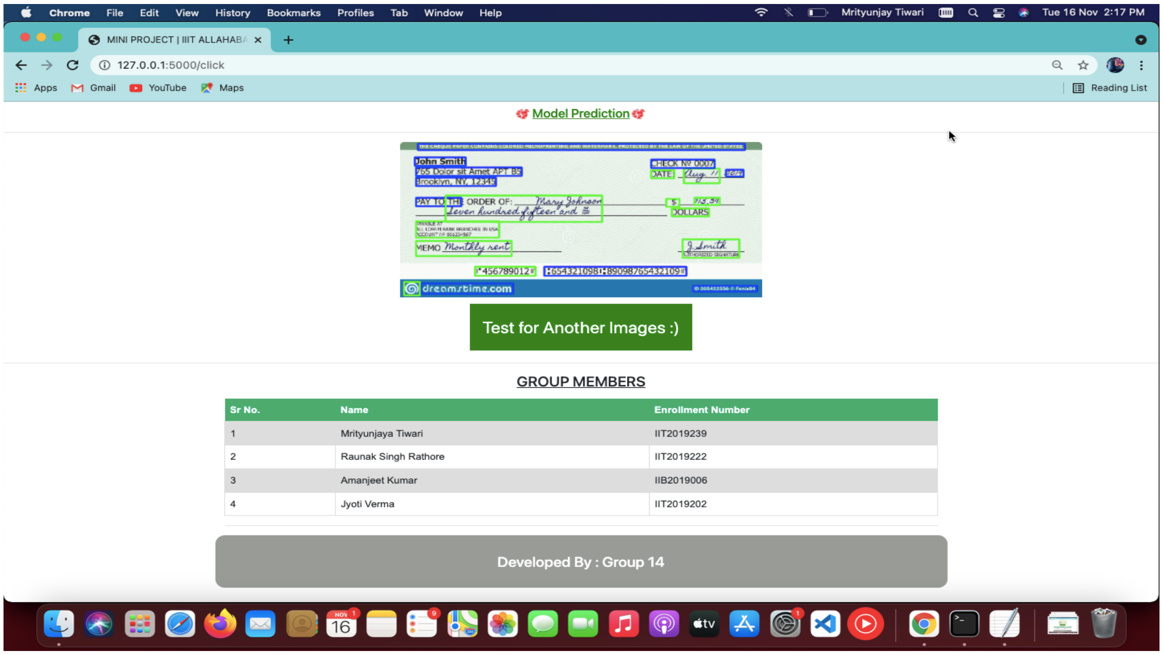
Mini Project

Under the Guidence of Dr. Pavan Chakrabotry Sir

**Please note that the deploying server heroku has a 30 second timeout. So if internet connection is very slow and/or if the image size is large, it may take too long time for uploading and downloading to get rid of time-out. So be careful and try to avoid both of them.

Please upload the image

Here images can be uploaded by clicking on CHOOSE FILE or simply drag and drop images . After uploading images click on the submit button for the output . Output screen is as shown below:


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Mrityunjay Tiwari
Tue 16 Nov 2:17 PM
MINI PROJECT | IIIT ALLAHABA
127.0.0.1:5000/click
Apps Gmail YouTube Maps
Reading List

Model Prediction

GROUP MEMBERS

Sr No.	Name	Enrollment Number
1	Mrityunjaya Tiwari	IIT2019239
2	Raunak Singh Rathore	IIT2019222
3	Amanjeet Kumar	IIB2019006
4	Jyoti Verma	IIT2019202

Developed By : Group 14

5. Results and discussion:

To the best of our knowledge, there exists no standard database to perform research on handwritten and machine printed text classification . Our dataset consists of 11340 images. We have split the entire dataset into train, validation and test sets by variable percentages maintaining mutual exclusion.

We have used the method of synthetic data generation to increase the size of our dataset. Finally, in the testing of the document a test image is found to contain

3,114 handwritten texts, 5,220 non-texts, 126 texts (which are in mixed or combined forms) and 2,880 texts which are printed as words or phrases. Features are extracted from a total of 11,340 images and the classification is done with the help of Random Forest classifier using 10-fold cross validation.

We have obtained 11308 instances to be correctly classified whereas 32 instances are wrongly classified. The proposed method obtains 0.0074 as mean absolute error and 0.0438 as root mean squared error. Relative absolute error is 2.2754%.

6. Performance analysis of our method:

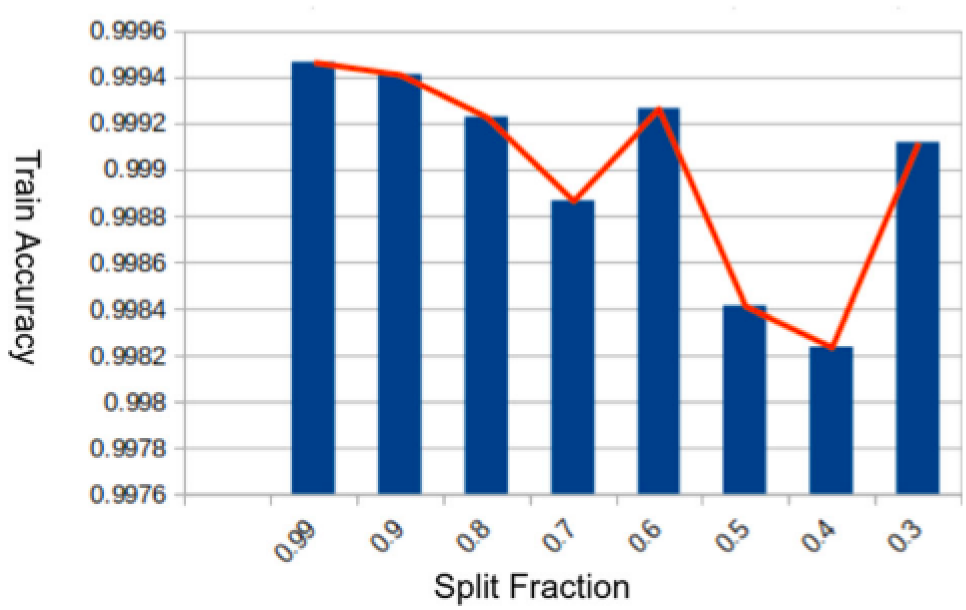
We have determined True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) as per the convention :

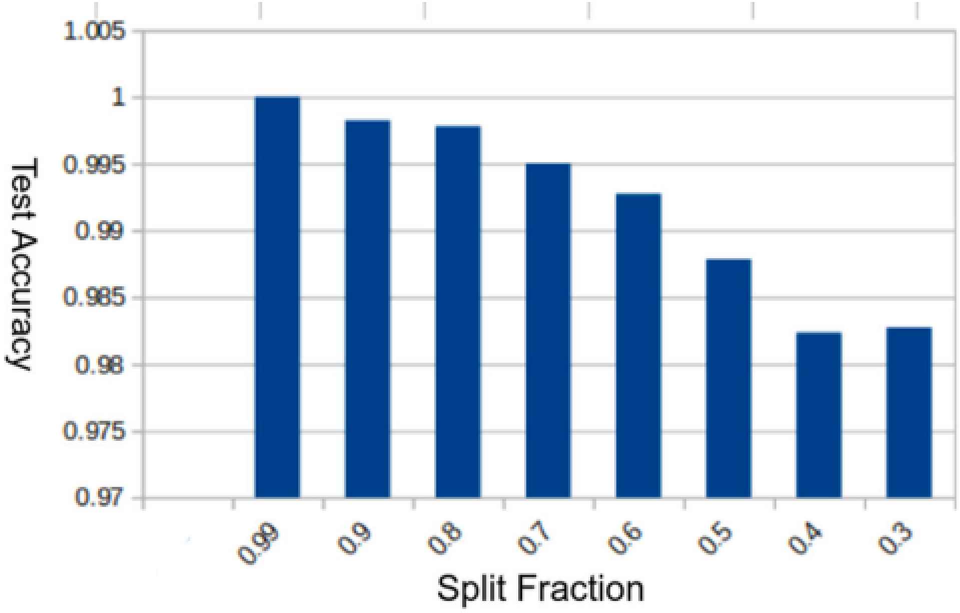
Stratified cross validation results using Random Forest classifier

Correctly Classified Instances	11308	99.72%
Incorrectly Classified Instances	32	0.28%
Kappa statistic	0.9956	
Mean absolute error	0.0074	
Root mean squared error	0.0438	
Relative absolute error	2.2754%	
Root relative squared error	10.872%	
Total Number of Instances	11340	
Running Time	3.02 second	

The variation of training and testing accuracies at different split fractions can be observed by the given charts .Testing accuracy increases definitely with the increment of the size of training dataset,whereas training accuracy, in general, increases, though a few exceptions are there.

The training and testing accuracy values at different split fractions are also shown in fig:





7. Work Done So Far:

Till now, we have deployed our model on our local system . It runs properly with good accuracy and in very less efficient time . Now, we are working to deploy our model in live servers like AWS, Heroku .

8. Future work:

Though, there are some aspects which could be looked into in future. For example, in the case of printed text, our method cannot produce good results for some complex fonts. Also the issue of touching text, where machine printed text gets joined with handwritten text, needs to be addressed. In few cases, documents illustrate some symptoms and mechanisms with some figures and diagrams , which have some significance but they are lost. Retrieval of such figures would also be tried in future. Another important issue is the preparation of realistic dataset instead of using synthetic data, which is also listed in our future plan.

9. Conclusion:

In this project, a method has been proposed to classify printed and handwritten texts in documents. As the proposed method has successfully classified the printed and handwritten texts in the documents and with a very low complexity, this can easily be embedded with a recognition module as an additional resource requirement. The scope of the present work can be used for direct application by the researchers and the netizens according to their requirements.

10. References:

- V.Pal and B.B.Chaudhuri, “Machine-printed and handwritten text lines identification”, Pattern Recognition Letters, 22, 2001, pp.431-441.
- U. Marti and H. Bunke, “A full English sentence database for off-line handwriting recognition”. Proc. 5th Int. Conference on Document Analysis and Recognition, ICDAR'99, 1999, pp. 705 – 708.
- E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, “Un Off-line Unconstrained Handwriting Recognition System”, International Journal of Document Analysis and Recognition, no 4, 2002, pp. 226-242.
- https://searchcontentmanagement.techtarget.com/definition/OCR-optical-character-recognition
- How to create salt and pepper noise in an image. https://www.projectrhea.org/rhea/index.php/How to Create Salt and Pepper Noise in an Image
- Lookup table (2020) https://en.wikipedia.org/wiki/Lookup table
- Moghaddam RF, Cheriet M (2012) Adotsu: An adaptive and parameterless generalization of otsu’s method for document image binarization. Pattern Recognit 45(6):2419–2431. https://doi.org/10.1016/j.patcog.2011.12.013 . Brain Decoding

- Chanda S, Franke K, Pal U (2010) Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments. In: Proceedings of the 2010 ACM symposium on applied computing, SAC ’10. Association for Computing Machinery, New York, NY, USA, pp 18–22. https://doi.org/10.1145/1774088.1774093
- Breiman L (2001) Random forests. Mach Learn 45(1):5–32