

An Automated Optical Character Recognition of Handwritten English Letters using Decision Trees &Random Forests.

Contents

[Cover Sheet 2](#_Toc121515029)

[Overview 3](#_Toc121515030)

[Similar applications 3](#_Toc121515031)

[literature review](#_Toc121515032) 5

[The Dataset](#_Toc121515033) 8

Applied Algorithms .........................................................................................................................8

Experiments & Results ..................................................................................................................9

Analysis & Discussion & Concolusion ........................................................................................11

[Diagrams 1](#_Toc121515034)3

**Artificial Intelligence**

# Cover Sheet

**Project Name: An Automated Optical Character Recognition of Handwritten English Letters using Decision Trees& Random Forests.**

**Team Information *(typed not handwritten, except for the attendance signature)*:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ID**  **[Ordered by ID]** | **Full Name**  **[In Arabic]** | **Attendance**  **[Handwritten Signature]** | **Final Grade** |
| **1** | 202000172 | اياد امين احمد احمد الفار |  |  |
| **2** | 202000574 | علي محمود حمزه محمد |  |  |
| **3** | 202001070 | يوسف ابراهيم احمد |  |  |
| **4** | 202000719 | محمد احمد احمد مجاهد |  |  |
| **5** | 202000655 | كارين اسامة العبد حنا |  |  |
| **6** | 202000612 | عمر نادر محمد نبيه |  |  |

**Project link:**

**shorturl.at/jqJLS**

# Overview

In this paper a detailed review of Handwritten Character Recognition is presented. Some features of human beings are unique to individuals like iris, fingerprint, DNA etc. Handwriting is one such feature which is different for each human being and it has been proven scientifically. In Handwritten Character Recognition (HCR) the task is to identify the characters written by humans and convert it into digital text. Handwritten Character Recognition is a field where plenty of research has been done but still there is scope in terms of improving accuracy and efficiency. Digitizing manually written text is very useful in today's world as it makes information readily available anywhere and anytime. Digitized text can be used for commercial purposes and it is more safe and environment friendly as compared to manual text.

Similar applications



**1. Microsoft OneNote :**

Converts legible handwriting into text. Download the OneNote mobile app for free through Apple’s App Store or Google Play or get the stand-alone desktop app for Windows PCs and Macs. However, you’ll get OneNote’s full functionality as part of the broader Microsoft 365 office productivity suite on Windows devices.

OneNote supports drag-and-drop text boxes and accepts stylus inputs. You can also use it to take a photograph of handwritten content before parsing it into a text document. As you’d expect, as part of the Microsoft 365 platform, OneNote relies on cloud-based storage and access.

Unlike some apps in this list, OneNote supports internal linking and version histories. It’s a surprisingly powerful tool once you get into advanced features like on-page content positioning and the ability to translate parts of a page rather than automatically translating the whole thing. Speaking of translation, OneNote supports 35 languages.

Features and functions:

1) Type information in your notebook or insert it from other apps and web pages.

2) Take handwritten notes or draw your ideas.

3) Use highlighting and tags for easy follow-up.

4) Share notebooks to collaborate with others.

5) Access your notebooks from any device.



**2. MyScript Nebo**

MyScript Nebo is designed specifically to use with tablets, including the Apple iPad , and newer Android devices with a stylus like Samsung’s S Pen.

Nebo uses its interactive ink tech to facilitate writing, drawing and formatting notes. It also converts text into shareable documents. Recent updates facilitate better performance with math objects and bring overall note-management improvements. A new library lets you view notebooks and collections in a side panel so you can rearrange documents via drag-and-drop functionality. An enhanced search engine covers your entire library.

Along with editing and formatting in an impressive 65 languages, you can sketch; annotate images; create editable equations; adjust type size and device orientation; organize notes in pages, notebooks and folders; and search, store and sync with Dropbox. You can export documents in text, Word, PDF and HTML formats. You can also copy, paste and edit diagrams to use in PowerPoint.

Features and functions:

1-variety of papers and pens

2-PDF annotation

3-digital scrap booking

4-Voice memos

***Abstract:***

***Optical Character Recognition is an application. This paper combines the functionality of Optical Character Recognition. The objective is to develop an application which performs image to editable character system. The OCR takes image as the input, gets text from that image and then converts it into speech. This system can be useful in various applications like banking, legal industry, other industries, and home and office automation. It mainly designed for people who are unable to read any type of text documents. In this paper, the character recognition method is presented by using OCR technology with random forests***

literature review

-OCR is the acronym for Optical Character Recognition , which is used for transform images that contain characters to editable characters and that is widely used in a lot of fields , text recognition is very important and it uses character recognition because it transfers text to list of characters and then it apply character recognition to them with a lot of more algorithms like fuzzy logic , words matrix and others.

-OCR is challenging as everyone has its own font and as there is a similarity between characters like ( ‘0’ , ‘o‘, ‘O’ ) and ( ‘i’ , ‘l’ , ‘I’ , ‘1’) and a lot of others.

The problem of single character recognition that it was over-fitting and lower accuracy because of similar characters.

-Random forest technique is one of the best models used for character recognition as it is accurate because of its algorithmic structure which is composed of many trees and takes the votes of them to choose the class ( character ).

- What everyone needs is to achieve best accuracy with the lowest samples that simulates most of characters with every shape.

- Most of the time the handwritten characters contains noise and it needs to be removed to not affect the model , luckily random forest is better that decision tree in that case , because random forest is less sensitive to outliers.

- OCR is very important as it is used in a lot of important fields like:

***a) Text Analysis & Detection:***

The Text Analysis is part of pre-processing. It analyzes the

input text and organizes into manageable list of words. Then it

transforms them into full text. Text detection localizes the text

areas from printed documents.

***b) Legal industry:***

it is used a lot in legal industry as most of evidence needs to be saved in papers after edits , OCR makes it easier as it makes images of characters to be editable

***c) Historic documents saver:***

It provides phonetic alphabets. The grapheme to phoneme

conversion is done. It is actually a conversion of

orthographical symbols into phonological symbols.

Resources:

1) Optical character recognition

[https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6a4b4f04d5ce3c3592832eb40c23cc8fc5a9131e](https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6a4b4f04d5ce3c3592832eb40c23cc8fc5a9131e" \o "OCR)

2) Random forests and decision trees

<https://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf>

3) Optical character recognition techniques

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=375e6f94bb9039f3df4fa2a625f11ac59db3629f

4) Isolated arabic characters OCR

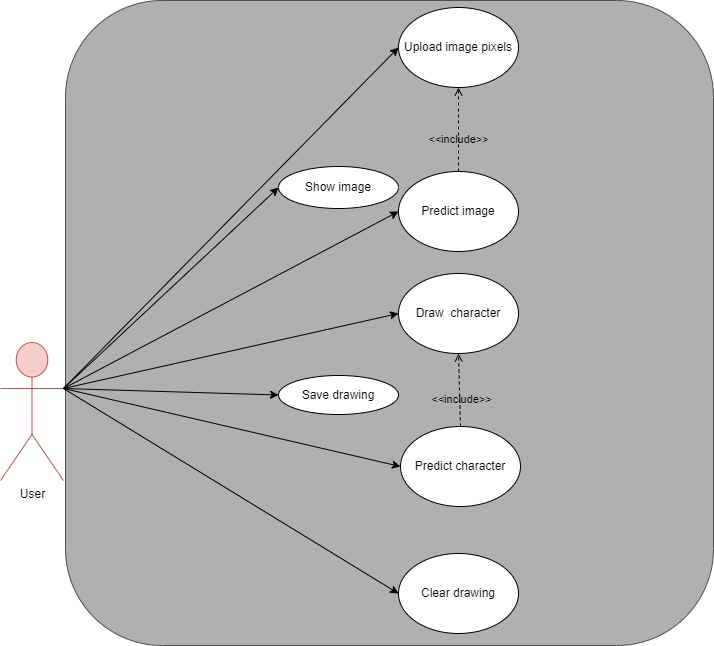
https://link.springer.com/content/pdf/10.1007/978-3-319-13461-1\_2.pdf?pdf=inline%20link

5) Springers EMNIST

https://link.springer.com/chapter/10.1007/978-3-030-43192-1\_73

The Dataset & Proposed solution

Proposed solution:



The Dataset employed:

resource:

[EMNIST (Extended MNIST) | Kaggle](https://www.kaggle.com/datasets/crawford/emnist)

<https://www.kaggle.com/datasets/crawford/emnist?select=emnist-balanced-train.csv>

[EMNIST (Extended MNIST) | Kaggle](https://www.kaggle.com/datasets/crawford/emnist?select=emnist-balanced-test.csv)

[EMNIST (Extended MNIST) | Kaggle](https://www.kaggle.com/datasets/crawford/emnist?select=emnist-balanced-mapping.txt)

Applied Algorithms

***Decision Tree Classifier:***

It is one tree contained from maximum number of important futures.

***Random Forest Classifier:***

It is a forest contains multiple random decision trees ,

Trees build in random forests as follows:

- Let Training objects as D , Number of features as F

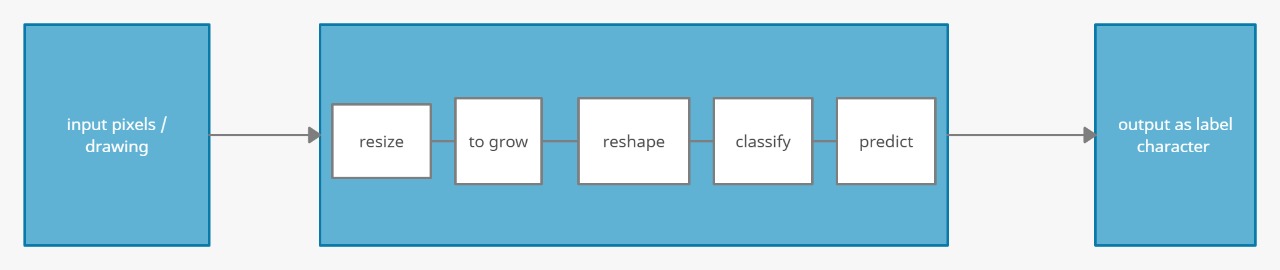
- Training of each tree is by choosing D from training objects , and f (where f << F )

from features randomly

- Each tree build largest set of questions

- Every tree classify the test data

- The forest choose the majority of votes as the predicted result



Experiments & Results

Experiments:

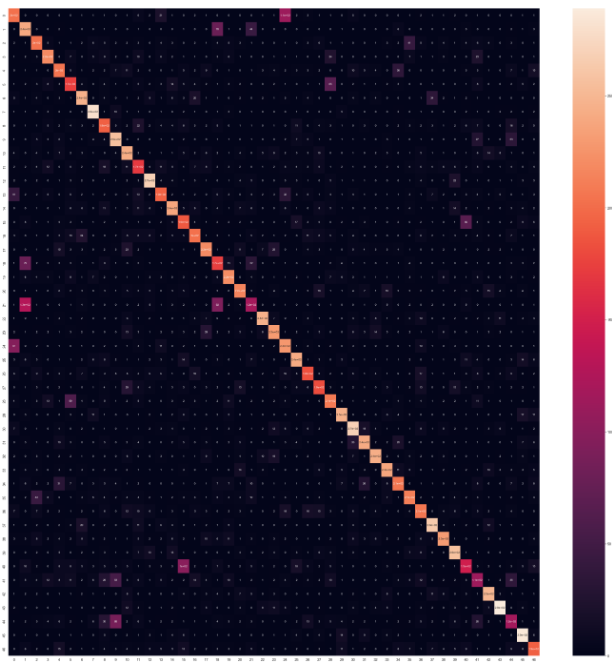
We have tried Random forest and Decision tree , since Random forest was significantly better than Decision tree , we tuned the parameters with grid search and cross validating until we achieved accuracy about 80.5% from 76.7% which is a lot better.

We found that the accuracy after deployed is very good but the drawback of the accuracy is due to characters that are similar in pixels.

Results:

**- Decision Tree**

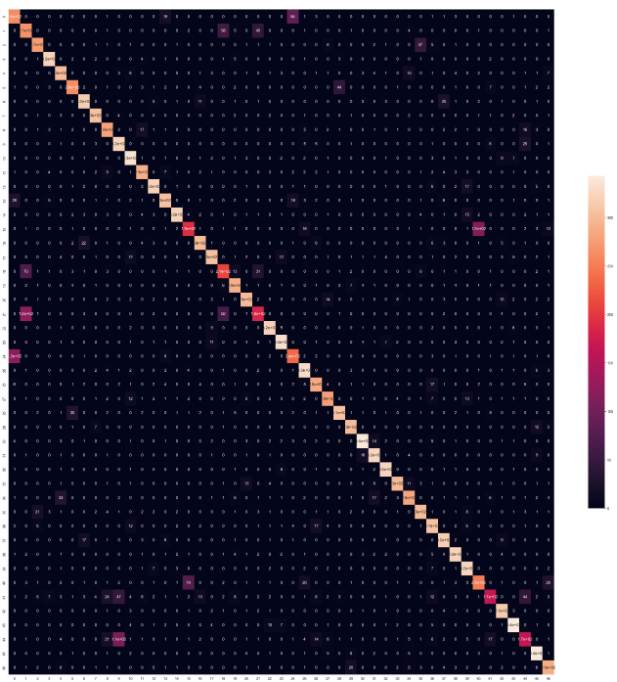
The First Algorithm applied is Decision Tree , by using the many different values of the decision tree model parameters , we got 60% test accuracy of the model.



**- Random Forest**

The Second Algorithm applied is Random Forest , by using the default values of the random forest model parameters , we got ¬ 76% accuracy of the model , and after many experiments on tuning the values with grid search and cross validation and k folds , the best result we achieved was test accuracy to 80.8%

which is a lot better since the accuracy is low because similar pixels characters as shown in confusion matrix.



Analysis, Discussion, and Future Work

***- Analysis of the results, what are the insights?***

Overall Random forest algorithm was better than Decision tree and that’s was obvious in this data-set because of the shape of it , and because Decision trees is more sensitive to outliers

***- What are the advantages / disadvantages?***

|  |  |
| --- | --- |
| **Decision Tree** | **Random Forest** |
| * Very over fitted | * A bit over fitted to data |
| * Bad overall accuracy | * Reasonable accuracy |
| * Not very bad after deployment | * Very good after deployment |
| * Fitted very fast as it is one tree | * A lot of time, resources to fit |
| * Playing with depth only under-over fit | * Playing with the numbers, depth found a sweet spot |
| * Overall, not a good model for data in that shape | * Overall, a good model |

***- Why did the algorithm behave in such a way? What might be the future modifications you’d like to try when solving this problem?***

The algorithms behaved that way because it takes images as pixels , since a lot of characters have a lot of matched pixels , it becomes difficult for the algorithm to predict that characters which decreased the accuracy.

Some future modifications that will increase accuracy is that using a word (bag of characters that have meaning) and predict the pixels of them with CNN , that will make model predict similar characters easily because that there’s a word like ( ‘more , More’ , but no ‘mOre’ or ‘m0re’ ) so it may be will a better solution.

***Development platform:***

*Tools: Anaconda , Jupyter Notebook , Jupyter lab , diagrams .*

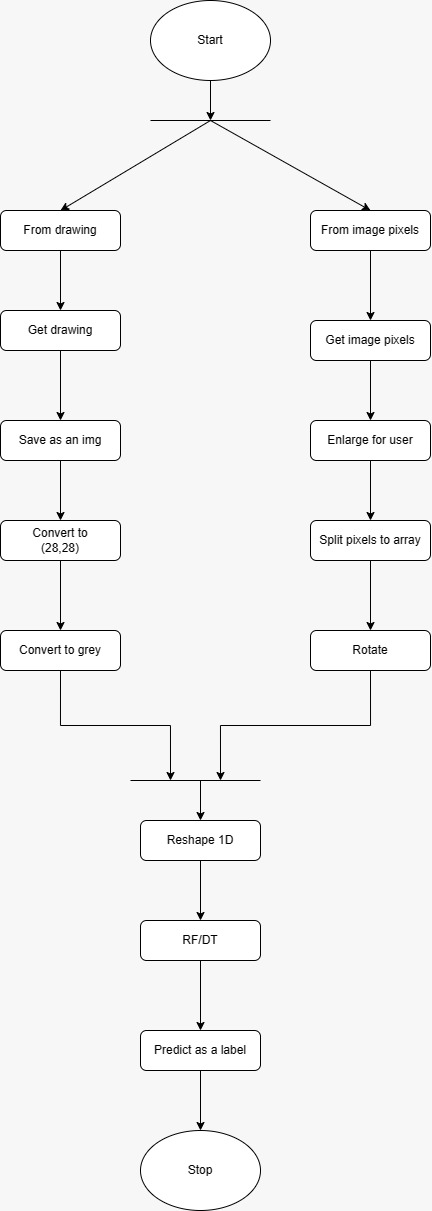
*Programming Languages : Python.*

*Python Libraries : pandas, numpy , matplotlib ,seaborn , cv2 ,*

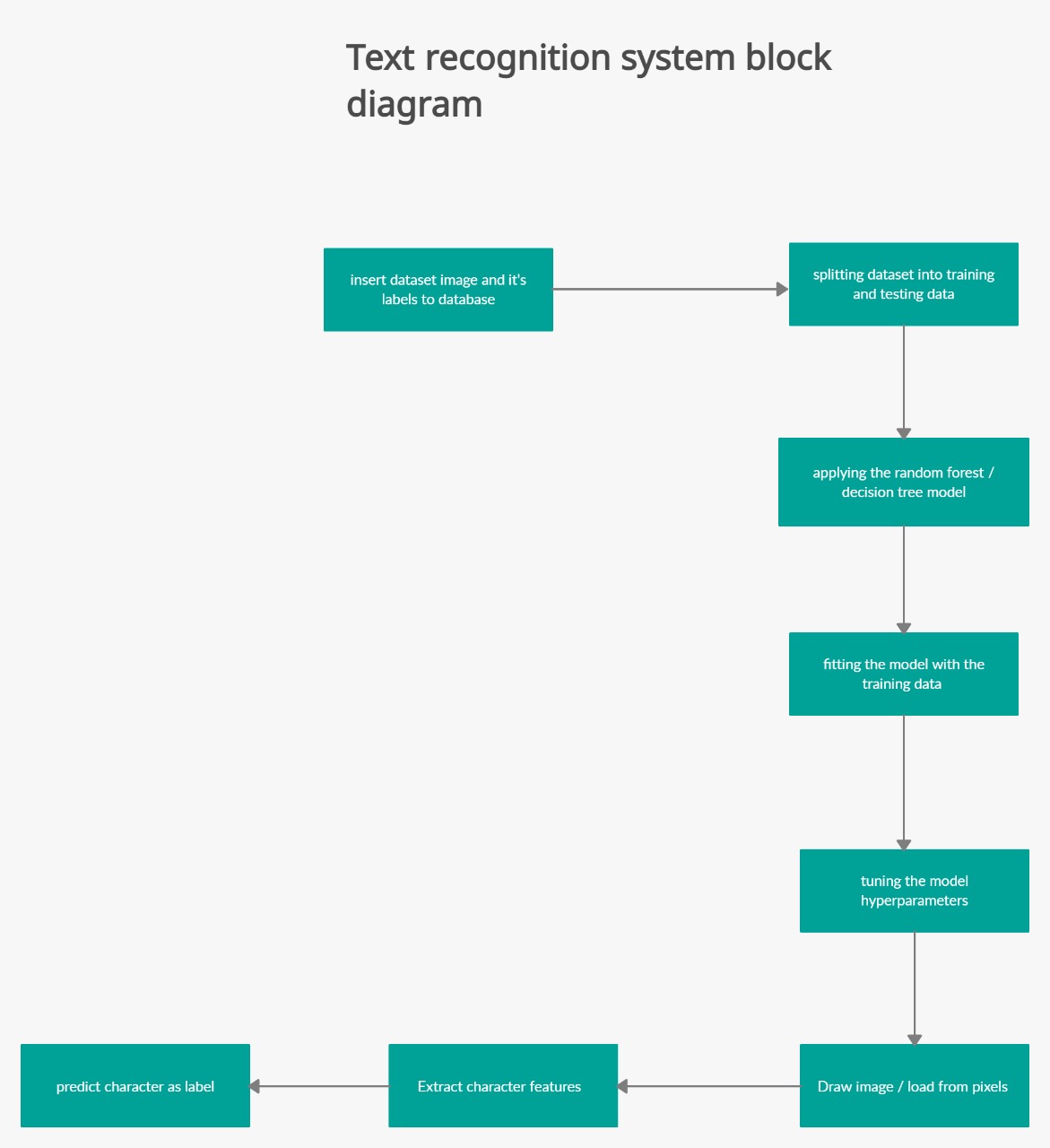
*sci-kit learn ,pickle , tkinter , pillow.*

Diagrams

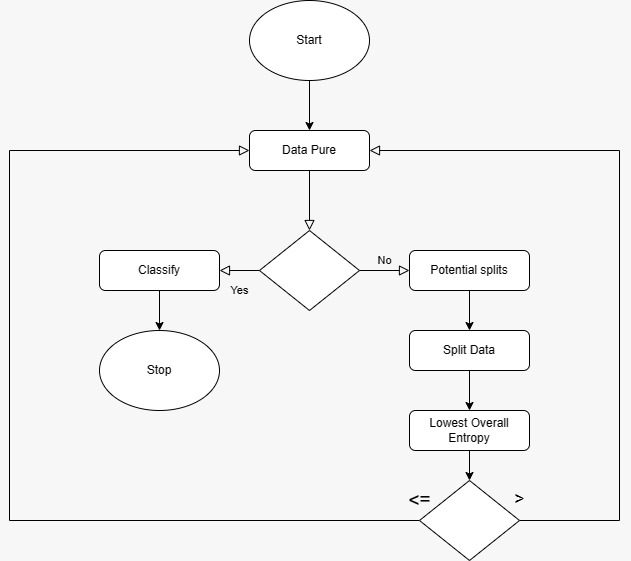
Activity diagram of the system:



Block diagram of the system:



Decision Tree diagram:



Random forest diagram:

