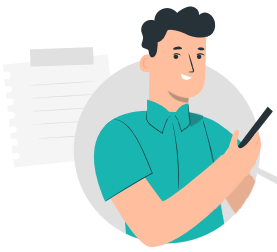
#teamFrontiers



17CE031

Naimish Ghevariya

17CE035

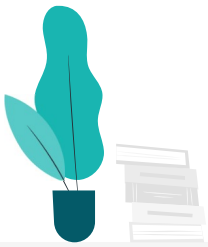
Riya Intwala

17CE044

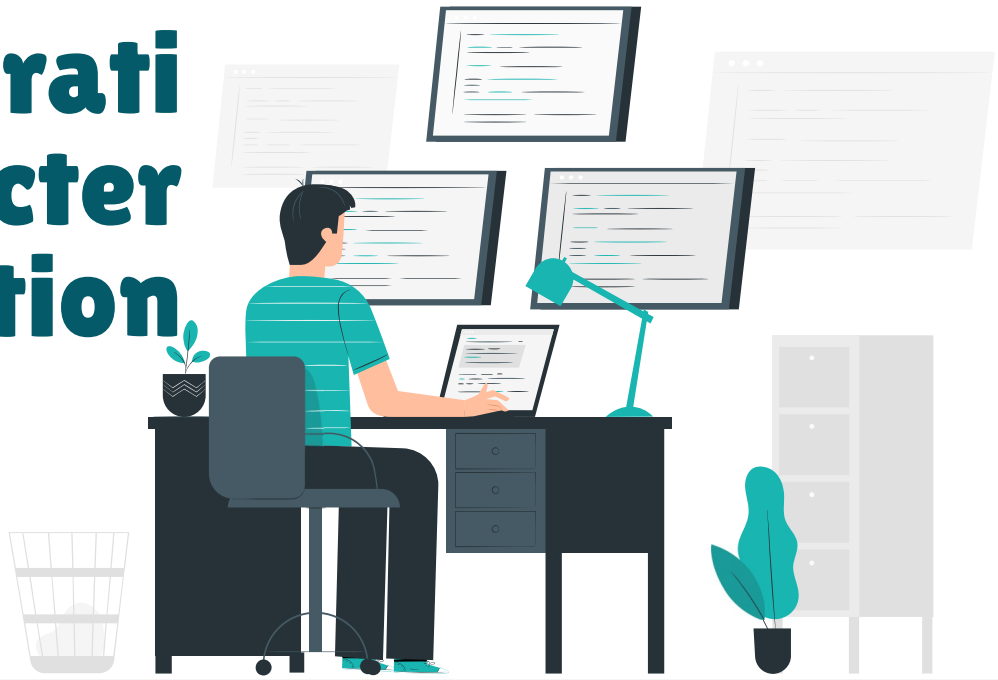
Jay Khatri

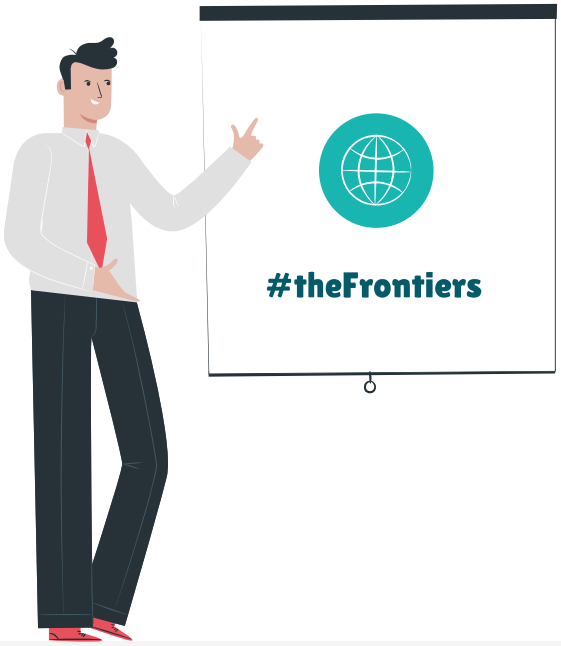
17CE045

Isha Khimsurya



Gujarati Character Recognition





OBJECTIVE

Aimed to develop an environment that can detect Gujarati character from images which can be handwritten or printed.

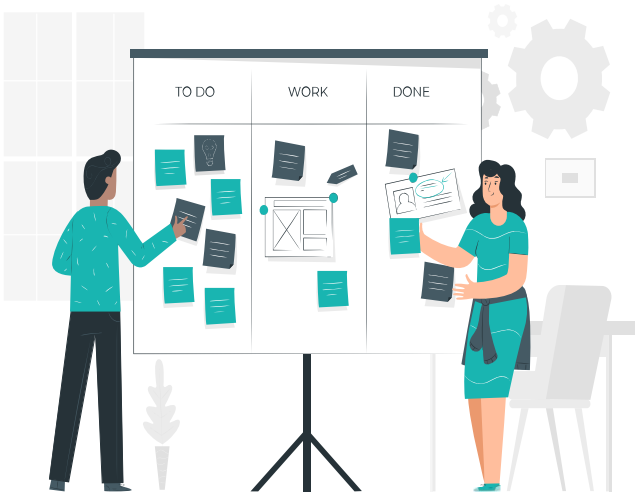
MODEL TRAINING

- developed using different machine learning and deep learning techniques

- gathering data of handwritten Gujarati characters Including…

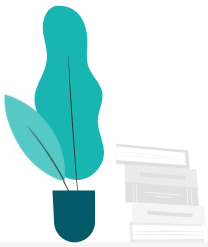
12 Vowels

35 Consonants



PROJECT PURPOSE

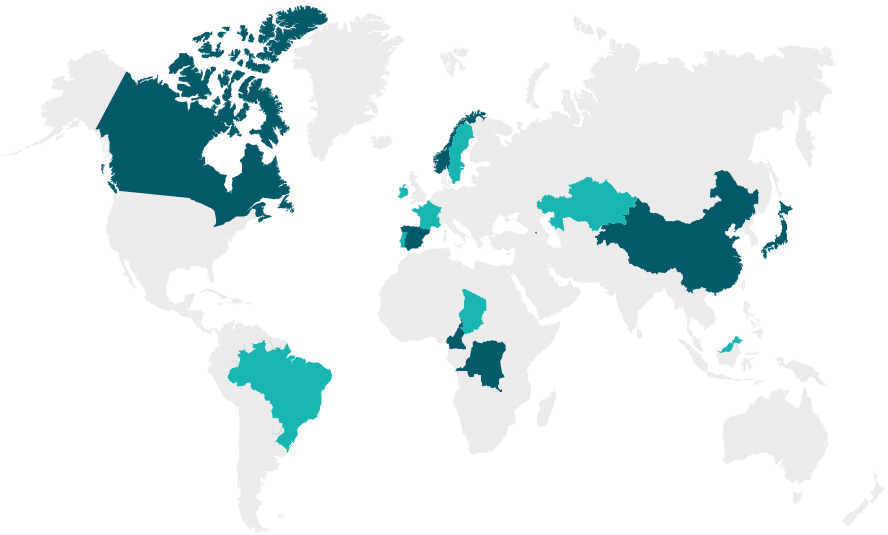
To help on a regional level by helping in digitizing the Gujarati documents easily and push the regional entities towards digital and modern time.



The Gujarati character recognition is the first step towards it.

Scope

huge opportunities to grow



India is going to be the **leading country** in terms of data and digital life.

India has many languages and Gujarati is one of the important parts of it.

55 M

Speakers in India

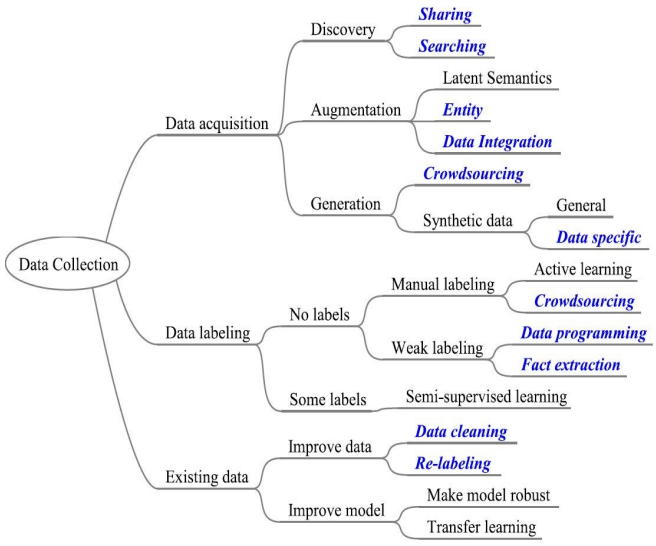
65 M

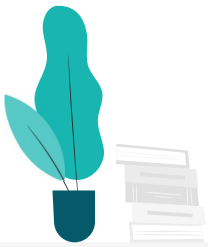
Speakers around the world

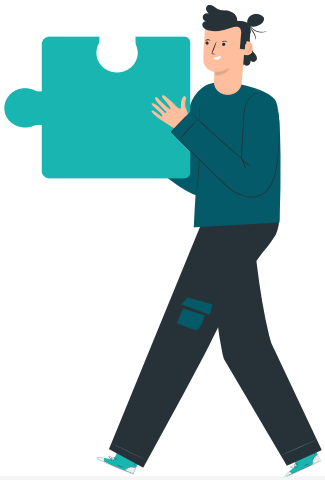
DATA COLLECTION

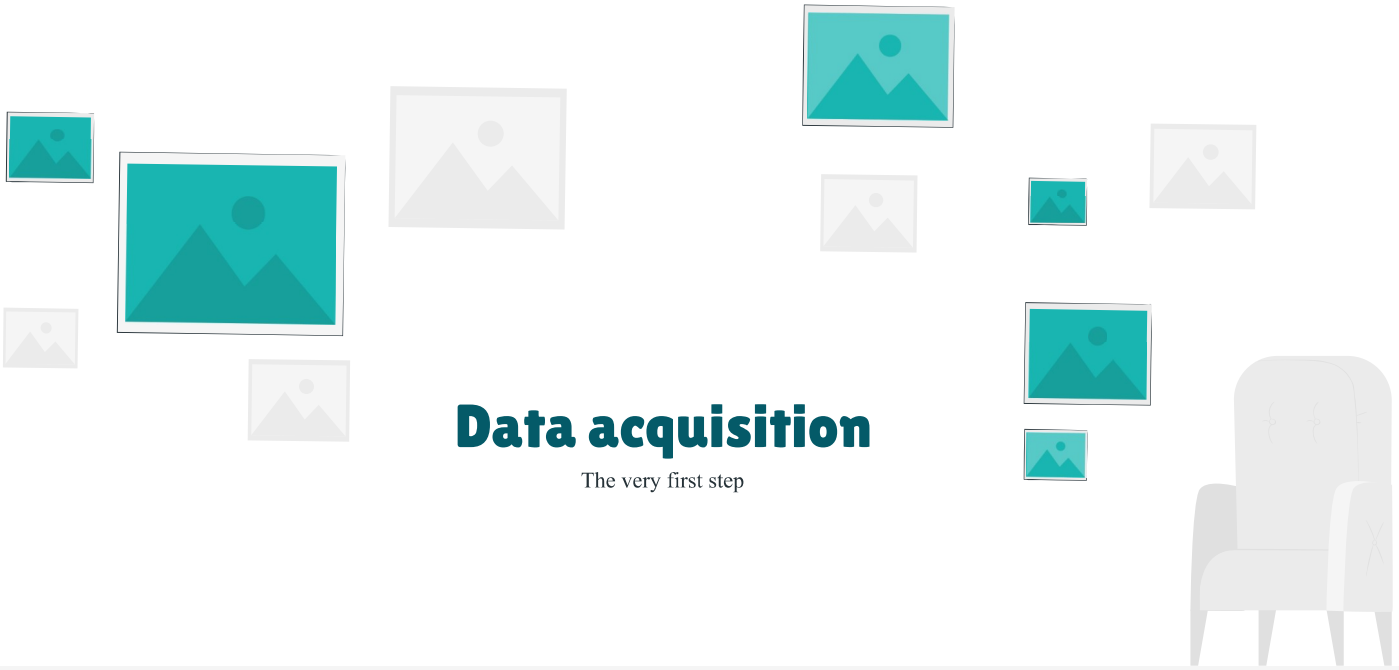
Standard research landscape of data collection

for machine learning







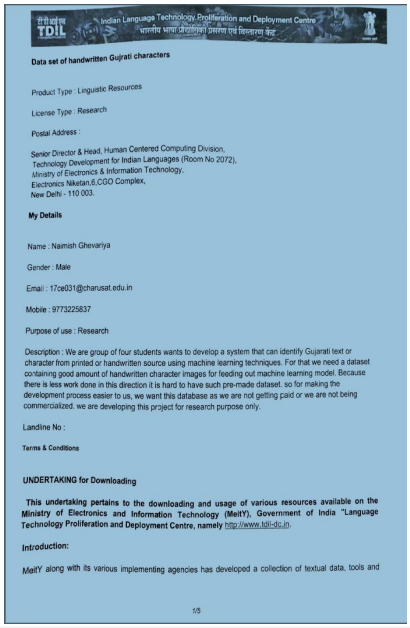


Data acquisition

The very first step

By sharing or searching

We gave an application requesting for Gujarati handwritten characters dataset.
Which has small dataset including almost 50 images per character.



Indian Language Technology Proliferation and Deployment Centre

Data set of handwritten Gujrati characters

Product Type : Linguistic Resources

License Type : Research

Postal Address :

Senior Director & Head, Human Centered Computing Division,
Technology Development for Indian Languages (Room No 2072),
Ministry of Electronics & Information Technology,
Electronics Niketan,6,CGO Complex,
New Delhi - 110 003.

My Details

Name : Naimish Ghevariya

Gender : Male

Email : 17ce031@charusat.edu.in

Mobile : 9773225837

Purpose of use : Research

Description : We are group of four students wants to develop a system that can identify Gujarati text or character from printed or handwritten source using machine learning techniques. For that we need a dataset co-taining good amount of handwritten character images for feeding out machine learning model. Because there is less work done in this direction it is hard to have such pre-made dataset. so for making the development process easier to us, we want this database as we are not getting paid or we are not being commercialized. we are developing this project for research purpose only.

Landline No :

Terms & Conditions

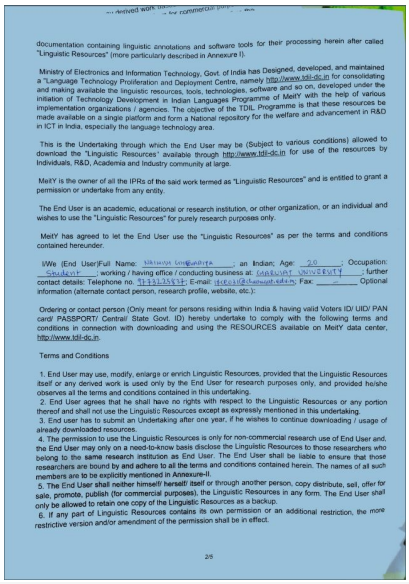
UNDERTAKING for Downloading

This undertaking pertains to the downloading and usage of various resources available on the Ministry of Electronics and Information Technology (MeitY), Government of India "Language Technology Proliferation and Deployment Centre, namely http://www.tdil-dc.in.

Introduction:

MeitY along with its various implementing agencies has developed a collection of textual data, tools and

1/5



documentation containing linguistic annotations and software tools for their processing herein after called "Linguistic Resources" (more particularly described in Annexure I).

Ministry of Electronics and Information Technology, Govt. of India has Designed, developed, and maintained a "Language Technology Proliferation and Deployment Centre, namely http://www.tdil-dc.in for consolidating and making available the linguistic resources, tools, technologies, software and so on, developed under the initiation of Technology Development in Indian Languages Programme of MeitY with the help of various implementation organizations / agencies. The objective of the TDIL Programme is that these resources be made available on a single platform and form a National repository for the welfare and advancement in R&D in ICT in India, especially the language technology area.

This is the Undertaking through which the End User may be (Subject to various conditions) allowed to download the "Linguistic Resources" available through http://www.tdil-dc.in for use of the resources by Individuals, R&D, Academia and Industry community at large.

MeitY is the owner of all the IPRs of the said work termed as "Linguistic Resources" and is entitled to grant a permission or undertake from any entity.

The End User is an academic, educational or research institution, or other organization, or an individual and wishes to use the "Linguistic Resources" for purely research purposes only.

MeitY has agreed to let the End User use the "Linguistic Resources" as per the terms and conditions contained hereunder.

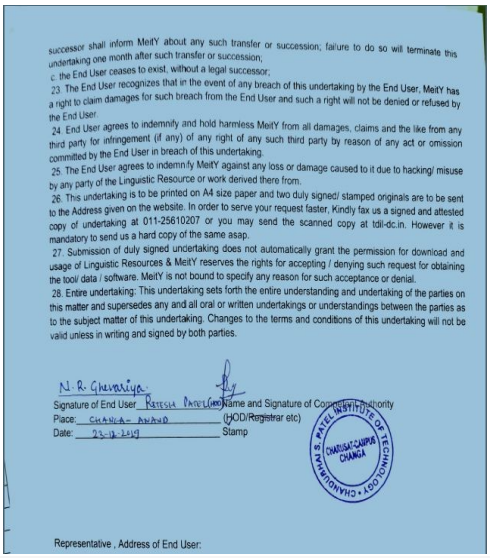
I/We (End User)Full Name: Naimish Ghevariya an Indian; Age: 20 ; Occupation: Student working / having office / conducting business at: CHARUSAT UNIVERSITY further contact details: Telephone no. 9773225837; E-mail: 17ce031@charusat.edu.in; Fax: — Optional information (alternate contact person, research profile, website, etc.):

Ordering or contact person (Only meant for persons residing within India & having valid Voters ID/ UID/ PAN card/ PASSPORT/ Central/ State Govt. ID) hereby undertake to comply with the following terms and conditions in connection with downloading and using the RESOURCES available on MeitY data center, http://www.tdil-dc.in.

Terms and Conditions

1. End User may use, modify, enlarge or enrich Linguistic Resources, provided that the Linguistic Resources itself or any derived work is used only by the End User for research purposes only, and provided he/she observes all the terms and conditions contained in this undertaking.
2. End User agrees that he shall have no rights with respect to the Linguistic Resources or any portion thereof and shall not use the Linguistic Resources except as expressly mentioned in this undertaking.
3. End user has to submit an Undertaking after one year, if he wishes to continue downloading / usage of already downloaded resources.
4. The permission to use the Linguistic Resources is only for non-commercial research use of End User and, the End User may only on a need-to-know basis disclose the Linguistic Resources to those researchers who belong to the same research institution as End User. The End User shall be liable to ensure that those researchers are bound by and adhere to all the terms and conditions contained herein. The names of all such members are to be explicitly mentioned in Annexure-II.
5. The End User shall neither himself/ herself itself or through another person, copy distribute, sell, offer for sale, promote, publish (for commercial purposes), the Linguistic Resources in any form. The End User shall only be allowed to retain one copy of the Linguistic Resources as a backup.
6. If any part of Linguistic Resources contains its own permission or an additional restriction, the more restrictive version and/or amendment of the permission shall be in effect.

2/5



successor shall inform MeitY about any such transfer or succession; failure to do so will terminate this undertaking one month after such transfer or succession;
c. the End User ceases to exist, without a legal successor;
23. The End User recognizes that in the event of any breach of this undertaking by the End User, MeitY has a right to claim damages for such breach from the End User and such a right will not be denied or refused by the End User.
24. End User agrees to indemnify and hold harmless MeitY from all damages, claims and the like from any third party for infringement (if any) of any right of any such third party by reason of any act or omission committed by the End User in breach of this undertaking.
25. The End User agrees to indemnify MeitY against any loss or damage caused to it due to hacking/ misuse by any party of the Linguistic Resource or work derived there from.
26. This undertaking is to be printed on A4 size paper and two duly signed/ stamped originals are to be sent to the Address given on the website. In order to serve your request faster, Kindly fax us a signed and attested copy of undertaking at 011-25610207 or you may send the scanned copy at tdil-dc.in. However it is mandatory to send us a hard copy of the same asap.
27. Submission of duly signed undertaking does not automatically grant the permission for download and usage of Linguistic Resources & MeitY reserves the rights for accepting / denying such request for obtaining the tool/ data / software. MeitY is not bound to specify any reason for such acceptance or denial.
28. Entire undertaking: This undertaking sets forth the entire understanding and undertaking of the parties on this matter and supersedes any and all oral or written undertakings or understandings between the parties as to the subject matter of this undertaking. Changes to the terms and conditions of this undertaking will not be valid unless in writing and signed by both parties.

N. R. Ghevariya

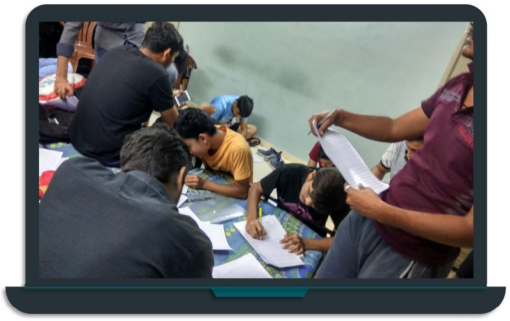
Signature of End User Ritesh Patel(Guide) Name and Signature of Competent Authority
Place: CHANGA- ANAND (HOD/Registrar etc)
Date: 25-12-2019 Stamp

CHARUSAT CAMPUS CHANGA

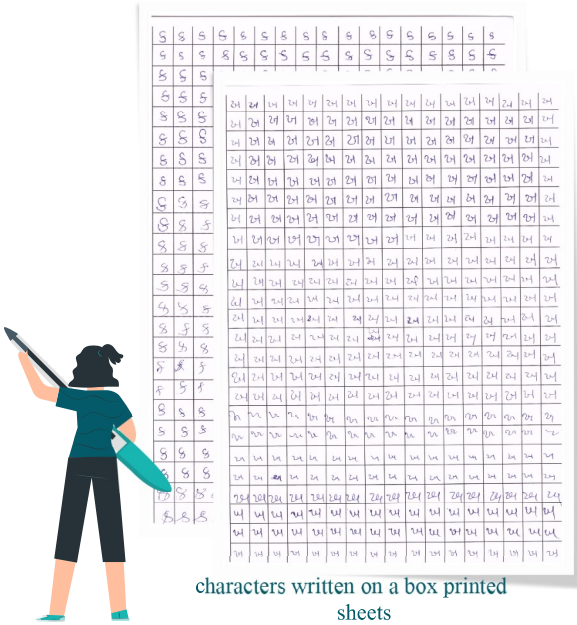
Representative , Address of End User:

No response !!

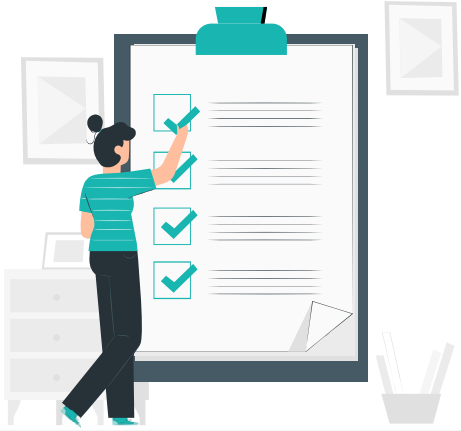
By crowdsourcing...



age group of 15 to 50-years people,
From different professions and different
linguistic backgrounds.



characters written on a box printed
sheets



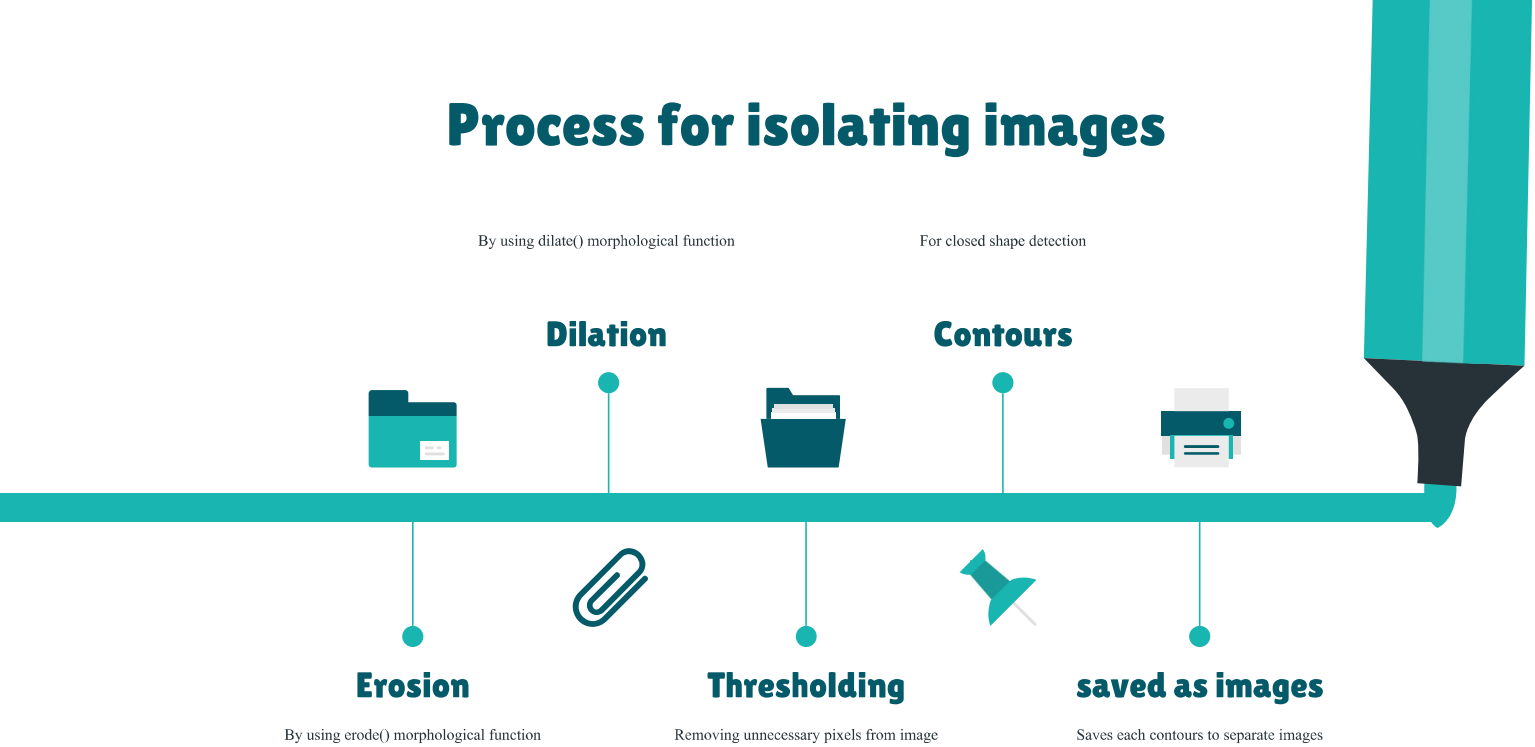
Challenge

to isolate those each characters from the single handwritten sheets into separate images

For that, we used

- edge detection
- closed box extractions.

Process for isolating images



By using dilate() morphological function

Dilation

For closed shape detection

Contours

Erosion

By using erode() morphological function

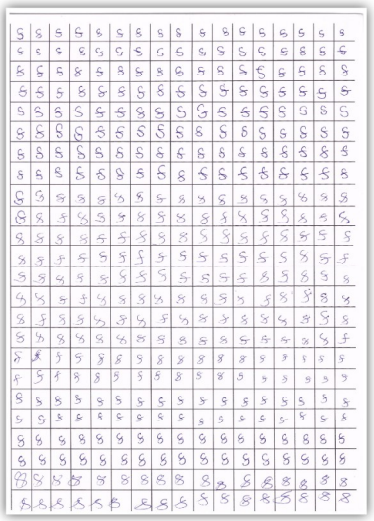
Thresholding

Removing unnecessary pixels from image

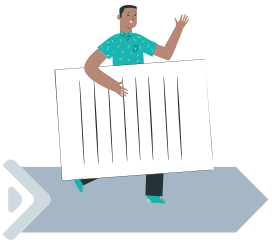
saved as images

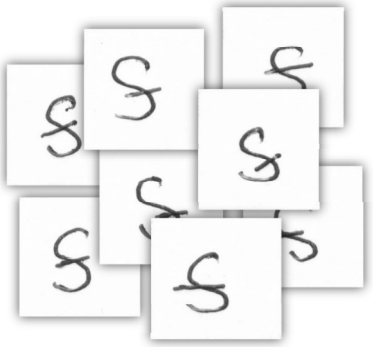
Saves each contours to separate images

Outcome

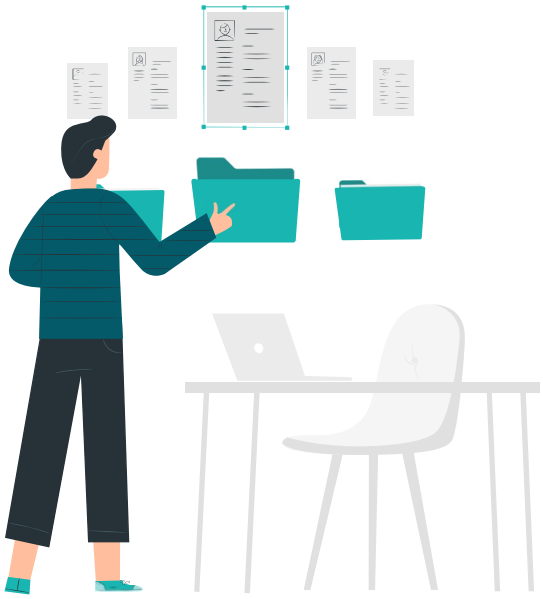


Non-isolated Gujarati characters





Isolated Gujarati characters



As a result we collected…

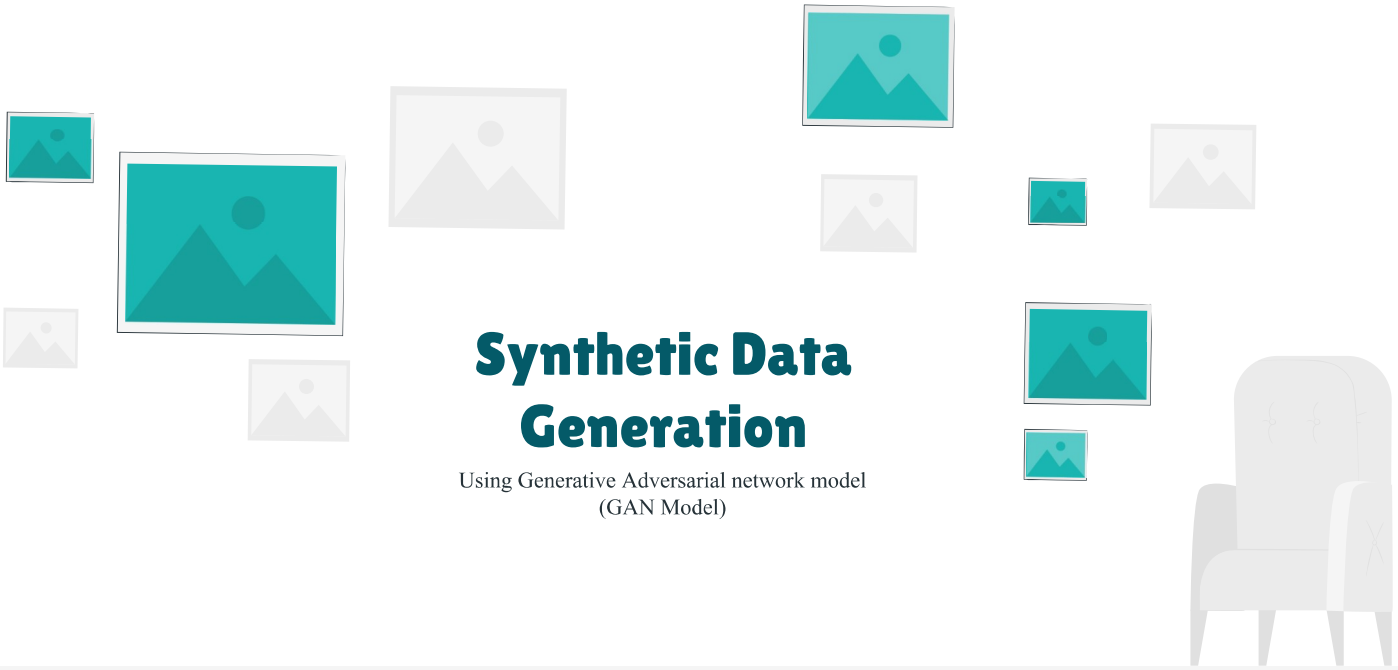
33,000 +

Total isolated characters

800 +

Image per character

The largest dataset available anywhere till now



Synthetic Data Generation

Using Generative Adversarial network model
(GAN Model)

Generative adversarial networks (GANs)

algorithmic architectures that uses two neural networks.

1. The generator

2. The discriminator

networks, putting one against the other (thus the “adversarial”) to generate new and synthetic instances of data that can pass for real data.

It is the same as that a thief or generator tries to produce fake data and police or discriminator try to catch the thief. The success will be there if the thief successfully produces fake images similar to the real one and never caught by the police

The generator model

It generates the fake images from noise created in a real image similar to the real image that we provide by using a convolutional neural network

- reshaped noise into 30*30 pixels
- added dense layer for the neuron to connect completely with all neurons
- passed 3 layers of convolutional layers and 3 layers of activation
- among them 2 them using activation function ReLU for linear increment and 1 of tanh for non-linear increment.
- have added 2 layers of batch normalization for stabilization

Output : an image of 120*120 pixel

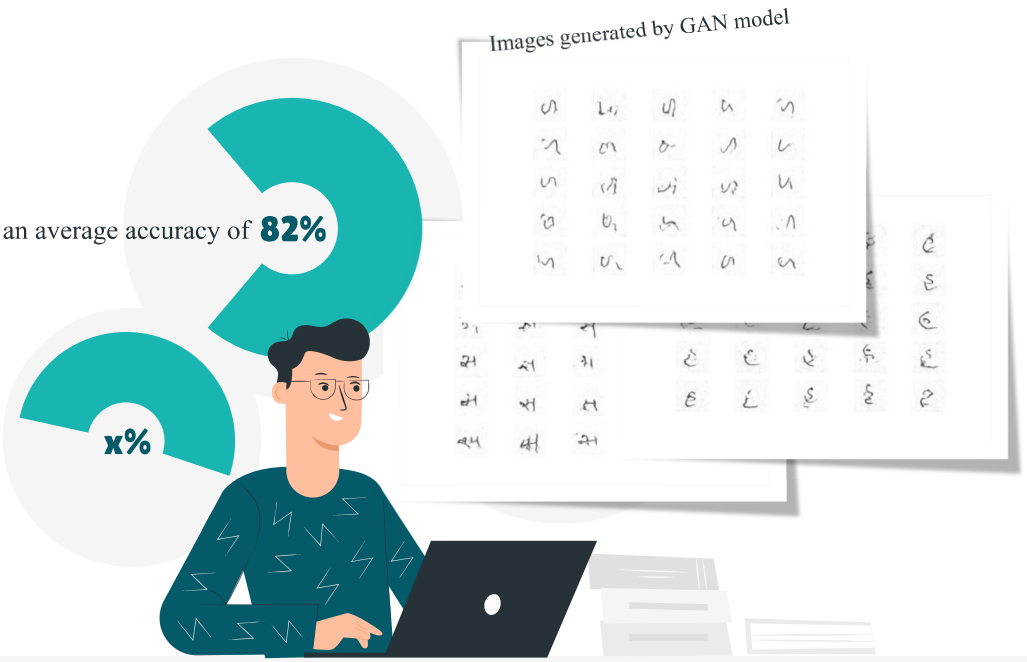
The discriminator model

Discriminator algorithm is used to train the Discriminator model that differentiates between real and fake Images. And it also used Convolutional Neural Network (CNN).

- used 4 layers of CNN
- In each layer, we have increased the filters in multiples of 2 such as 16,32 and 64 with stride value 2.
- For fixing dying ReLU, we have used leaky ReLU activation function. To remove neuron sets we also have used a dropout function that removes the random number of neuron sets.
- a flatten function to transform a 2d matrix into a vector.
- used epoch = 10000 and divided it into 400 batches. In one main image, there is a total of 25 images (5*5 matrix).

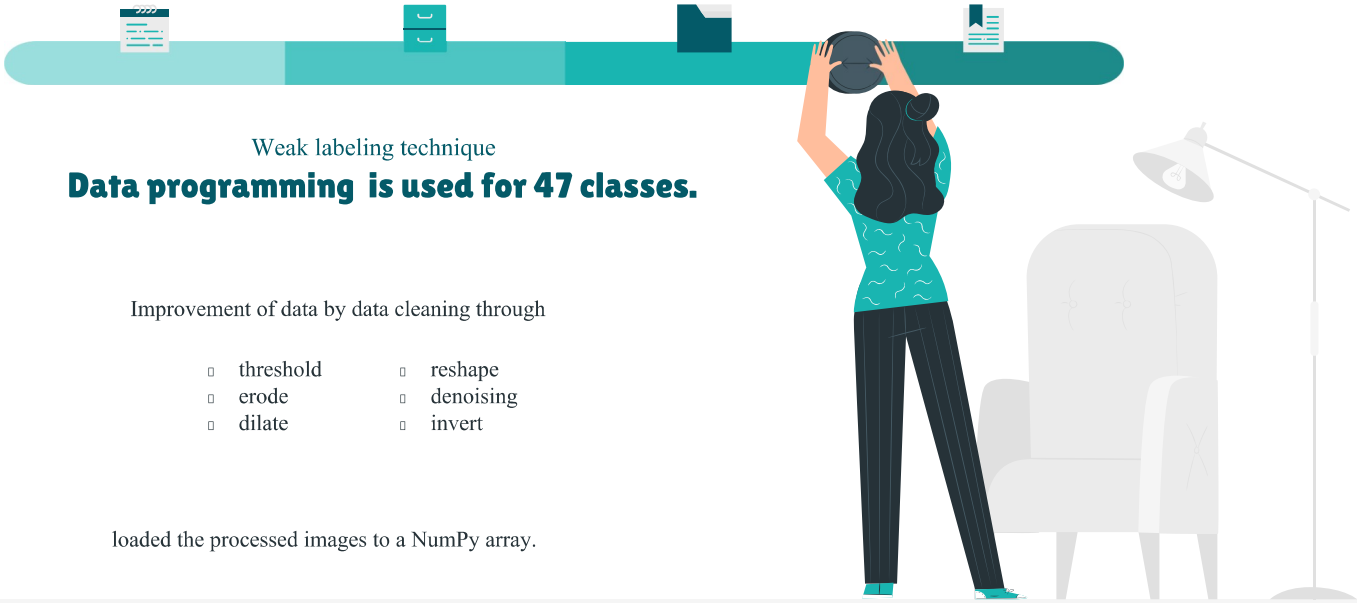
Output : an image of 120*120 pixel

We have achieved an average accuracy of **82%** per image.



But still **research is in progress** for this model.

Data labeling and preprocessing



Weak labeling technique

Data programming is used for 47 classes.

Improvement of data by data cleaning through

- threshold
- erode
- dilate
- reshape
- denoising
- invert

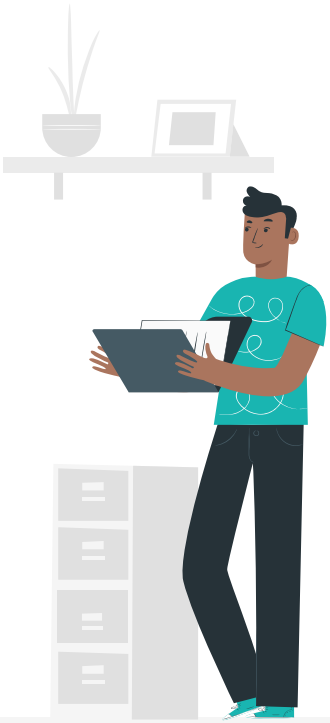
loaded the processed images to a NumPy array.

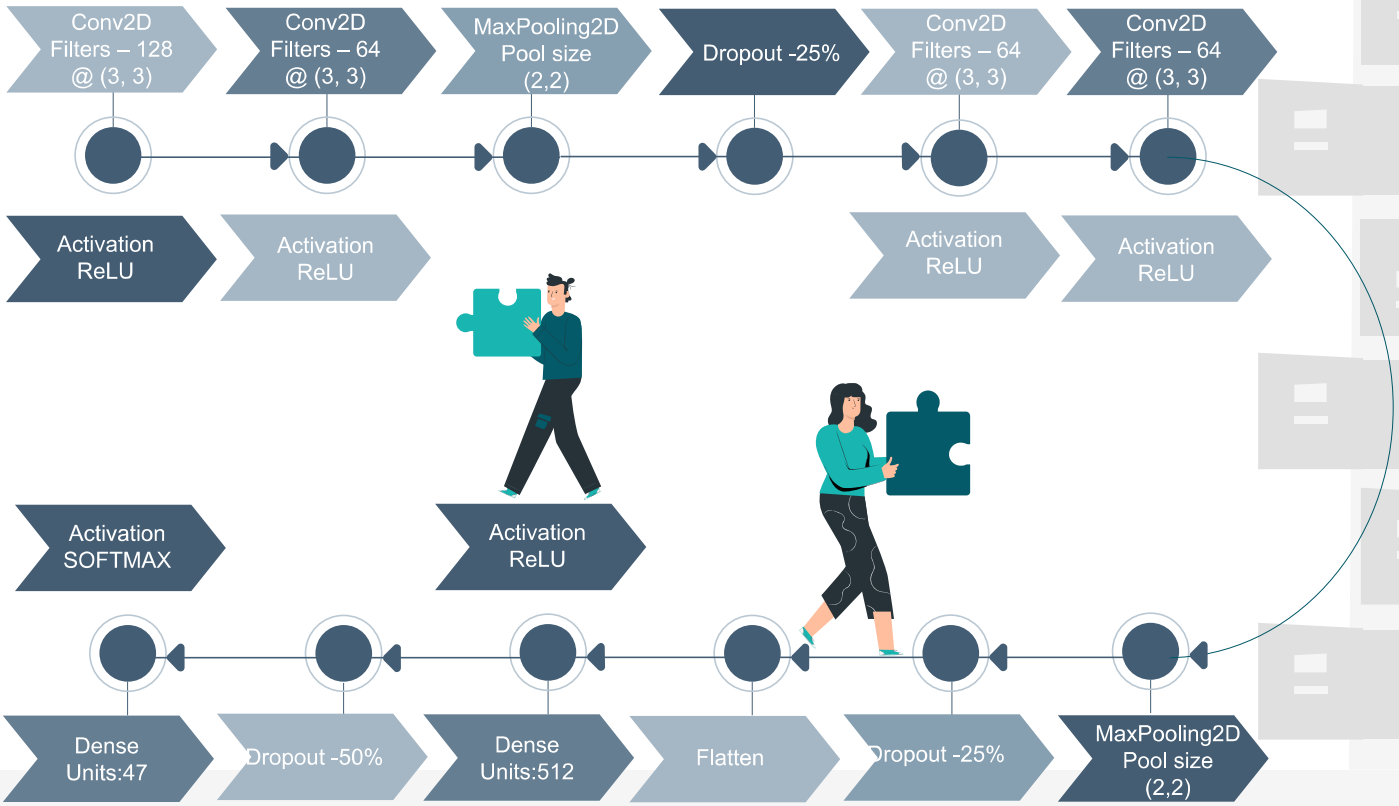
Training and prediction

Environment : Google Colab GPU as runtime

Training data : Testing data 85:15

Model architecture : Sequential model





MODEL DESIGN

Optimizer used for Keras is “Adam”

Accuracy

over the training data

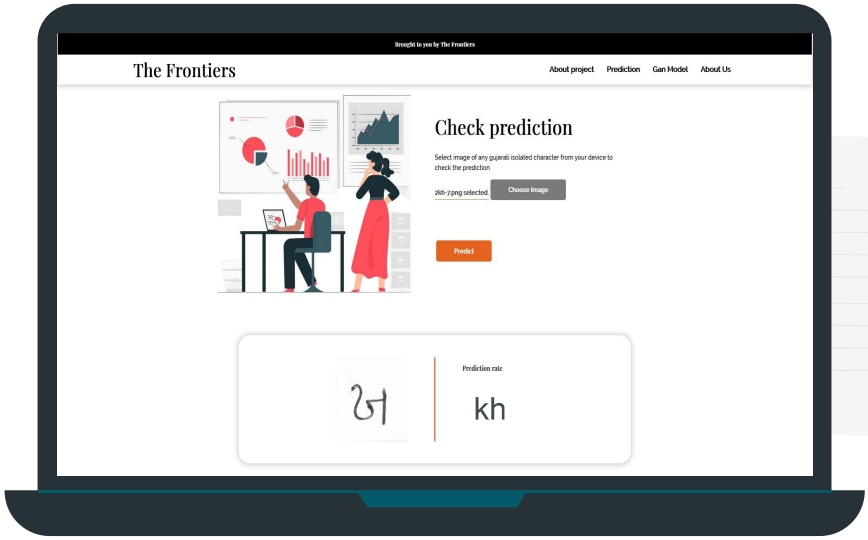
90.82%

86.03%

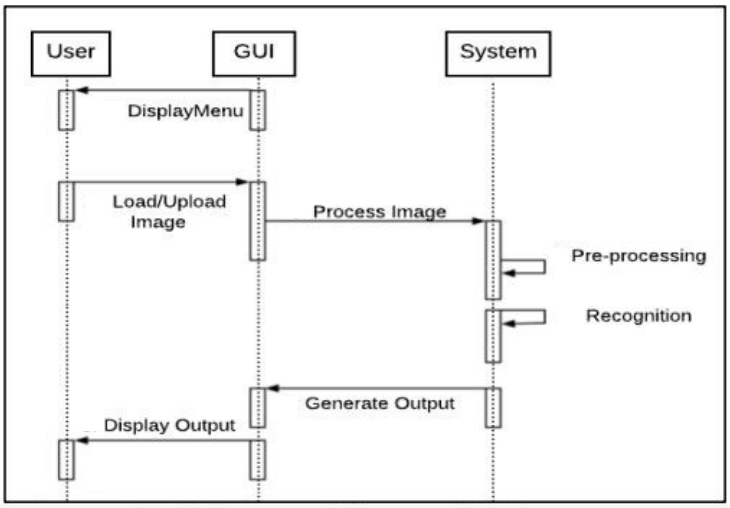
Over the test data

Prediction and GUI

For, Prediction part we are taking input as an image. Further taken input would be pre-processed. After pre-processing, feeding the pre-processed image into the trained model for prediction by classification.



The sequence diagram for GUI



References

- Handwritten text recognition: with deep learning and android *published by: blue eyes intelligence engineering & sciences publication retrieval number: C11730283S19/19©BEIESP*
- A survey on data collection for machine learning, A big data - AI integration perspective by yuji roh, geon heo, steven euijong whang, arxiv:1811.03402v2 [cs.LG] 12 aug 2019
- A Survey of Gujarati Handwritten Character Recognition Techniques, International Journal for Research in Applied Science & Engineering Technology (IJRASET) *ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887*
- https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py
- https://nextjournal.com/gkoehler/digit-recognition-with-keras
- https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-from-scratch-for-mnist-handwritten-digit-classification/
- https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa
- https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/
- https://pathmind.com/wiki/generative-adversarial-network-gan
- https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5

THANKS

For watching.