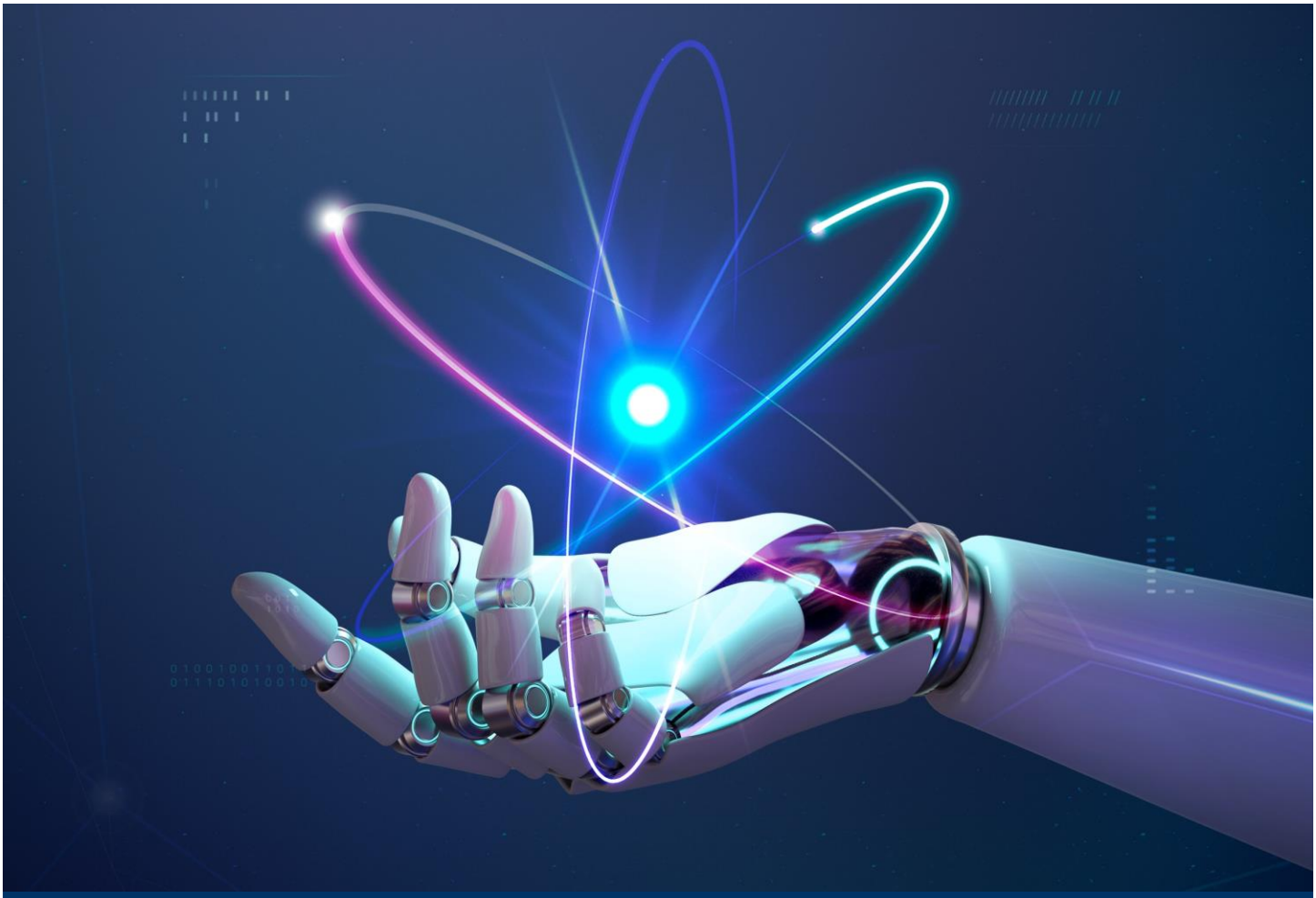


Handwriting Based Gender Classification

5/30/22

Neural Networks and Pattern Recognition

Team 5

Team Members:

Name	Section	BN
Raghad Khaled	**1**	**31**
Menna Allah Ahmed	**2**	**30**
Nada El-Sayed	**2**	**33**
Mohamed Khaled	**2**	**13**

Overview

1. Automatic analysis of handwriting for gender classification

Database

in this paper they are use the Qatar University Writer Identification (QUWI) database and a custom-developed Multiscript Handwritten Database (MSHD).
QUWI: this is English and Arabic data, it about 1,900 samples and 475 writers.
MSHD: this is French and Arabic data, it about 1,044 samples and 78 writers.

Features

They divided the features into three categories:

- Orientation and curvature: in this set they use multiple extractor Chain code-based features, and Polygon-based features.
- Fractal features: they calculate fractal dimension using box counting
- Texture-based features: Local binary patterns, AR coefficients (we didn't use this)

Classifiers

They use SVM and ANN (3-layers network) for classification.

Results

Data set	Features	Mode			
		Text-dependent		Text-Independent	
		SVM (%)	ANN (%)	SVM (%)	ANN (%)
QUWI-English	Slant and Curv.	68.00	70.00	70.00	66.00
	Texture	63.00	62.00	62.00	61.00
	Fractal	65.00	65.00	65.00	64.00
QUWI-Arabic	Slant and Curv.	69.00	71.00	63.00	62.00
	Texture	65.00	63.00	63.00	63.00
	Fractal	66.00	66.00	62.00	65.00
MSHD-French	Slant and Curv.	68.25	67.06	67.46	66.27
	Texture	66.67	66.27	66.27	65.48
	Fractal	64.68	66.27	63.09	65.87
MSHD-Arabic	Slant and Curv.	73.41	72.62	68.65	69.44
	Texture	74.20	72.22	72.22	71.43
	Fractal	65.08	65.87	64.28	65.08

2. A method for automatic classification of gender based on text-independent handwriting

Database

in this paper they are use ICDAR 2013- Gender prediction
competition
ICDAR: this is English and Arabic data, it about 1,900 samples and 475 writers.

Features

They use features extractor

- Region properties (we didn't use this).
- Perimeter (we didn't use this).
- Circle counting (we didn't use this).
- Pixel counting (we didn't use this).
- Skew angle.

Classifiers

They use SVM, Logistic regression and K nearest neighbor, then they use

Majority voting classifier to classify the input

Results

Classifiers	Classification Results
SVM	61.5%
Logistic Regression	63.3%
KNN	56.5%
Majority Voting	**65.71%**

3. "Improving Handwriting based Gender Classification using Ensemble Classifiers"

The proposed method addressed in this paper is:

- Feature Extraction: using set of features like (segmentation-based Fractal texture analysis), (Local binary pattern LBP), (histogram of oriented gradients HoG), (Gray Level co-occurrence matrix GLCM).

  This table shows the dimension of the extracted features:

Table 2: Summary of Features

SNo.	Feature	Dimension
1.	SFTA	24
2.	LBP	243
3.	HoG	81
4.	GLCM Statistics	20

- Classification:
  This step is carried out using multiple learners like (Decision Tree), SVM, KNN, ANN, RF
  Those classifiers are applied to each feature on its own, then applied to the combinations of all features used (LBP+HOG+GLCM+SFTA).

  Then **bagging** method is used to select the class of the test point. The class is chosen by most of the classifiers is voted as the decision.

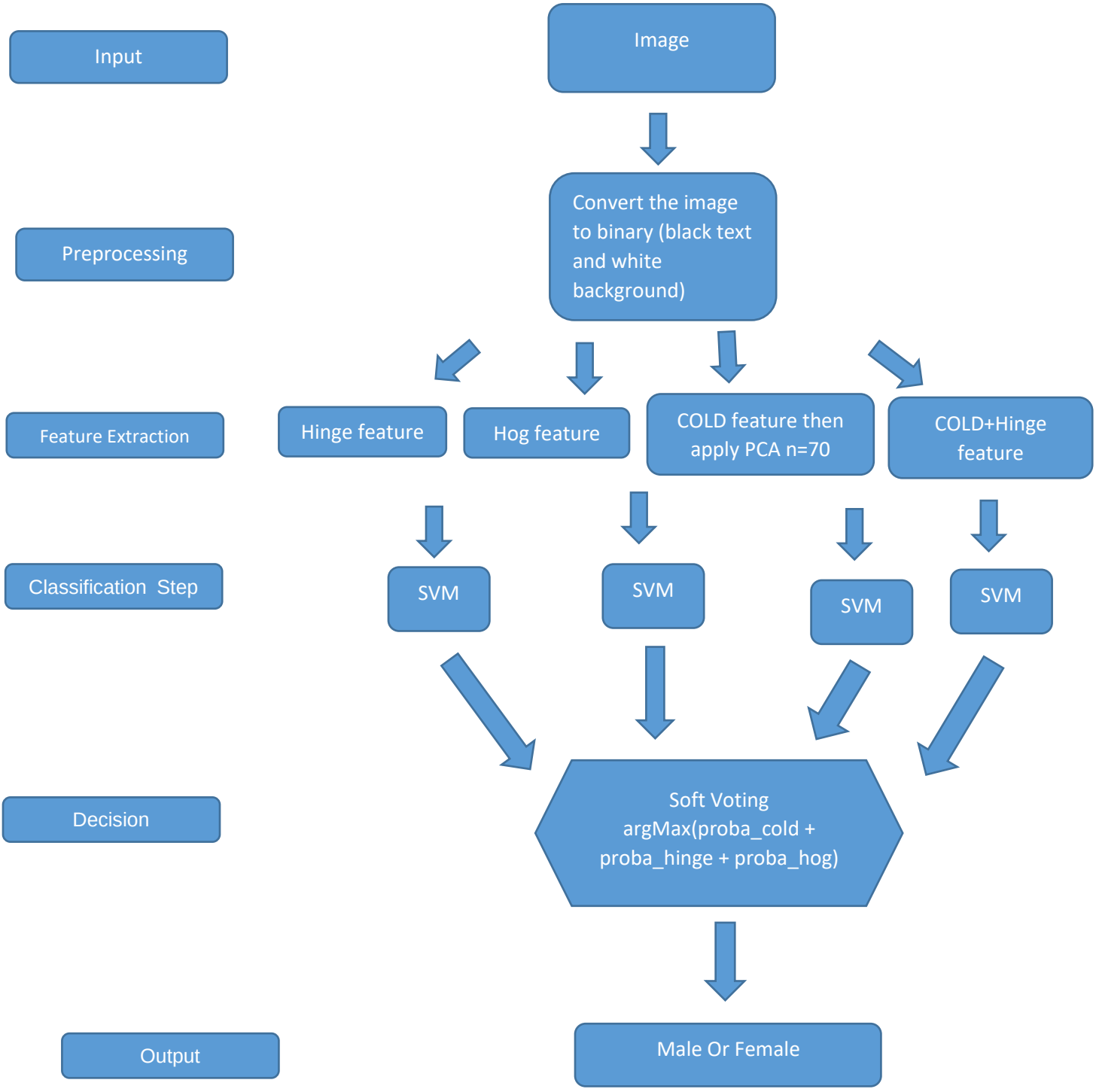
  The paper reached out that there are unstable learning algorithms like DT and ANN which small changes in the training data will give different results, while stable learners like KNN will show no or little changes.
  So, it reached out that the stacking method for the classification will lead to highest accuracy and reduce error rate of one or more classifiers.

  The results:

Table 5: Performance (in terms of classification rates) of ensemble classifiers for Scenario-A

Meta Learner	Base Learner		LBP	HoG	GLCM	SFTA	All
Bagging	ANN		74	63	75	50	74
	SVM		69	54	77	55	68
	DT		71	62	67	46	57
	KNN		67	62	59	55	55
	RF		61	55	70	60	68
Voting	KNN, SVM, ANN		69	63	77	56	69
	KNN, SVM, DT		70	63	72	55	69
	SVM, DT, ANN		74	58	78	59	75
	KNN, DT, ANN		**79**	**66**	73	**60**	73
	KNN, SVM, ANN, DT		68	63	72	56	69
Stacking	Level 0	Level 1					
	SVM + KNN	DT	65	57	71	49	69
	ANN+SVM	DT	74	63	78	49	**76**
	SVM+ANN	KNN	67	63	**79**	56	58
	SVM+DT	KNN	66	61	**79**	56	58
	SVM+KNN+DT	KNN	66	65	73	58	58
	SVM+ANN+KNN	DT	74	63	71	49	**76**

Project Pipeline



Preprocessing

After playing a lot with the data and exploring different features, we figured out that we need the images in binary form in order to make the model less error-prone because of the lighting of the image.

steps:

1. resize the image with a suitable ratio because the ram was full and the kernel crashed when the images were full-sized.
2. gaussian blur the image to decrease the noise by spreading it
3. Use adaptive thresholding in order to make the thresholding independent of the lighting.
4. Median the image to remove salt and pepper noise.
5. Erode the image to make the black writing larger.

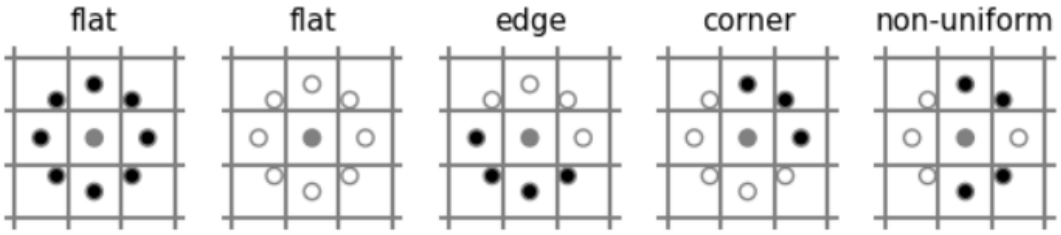
Feature Extraction/Selection Module

Used Two of Texture-based features as Texture analysis of handwriting considers each writing as a visually distinctive texture.

1. Local binary patterns

The approach is to Consider a set of neighborhood pixels V = (V0; V1; . . .; V8), the adjacent pixels are compared to the central pixel V0 to generate a binary pattern. The binary assignment is performed as follows. For i =1; 2; . . .;8 if Vi<V0 we assign the value 0 to the neighboring pixel i, otherwise, it is assigned the value 1. The resulting pattern is considered as a binary number, and multiplying each bit by the respective weight and summing the values together the LBP code for the central pixel is computed. The extensions to this original method what was suggested to use which include neighborhood of different sizes and the concept of uniform (less than 2 transitions between 0s and 1s ex: 00000000 ,01000000) and non-uniform (more than 2 transitions ex:01101100) binary patterns. There is a separate bin for each uniform pattern and all the non-uniform patterns are counted in a single bin.

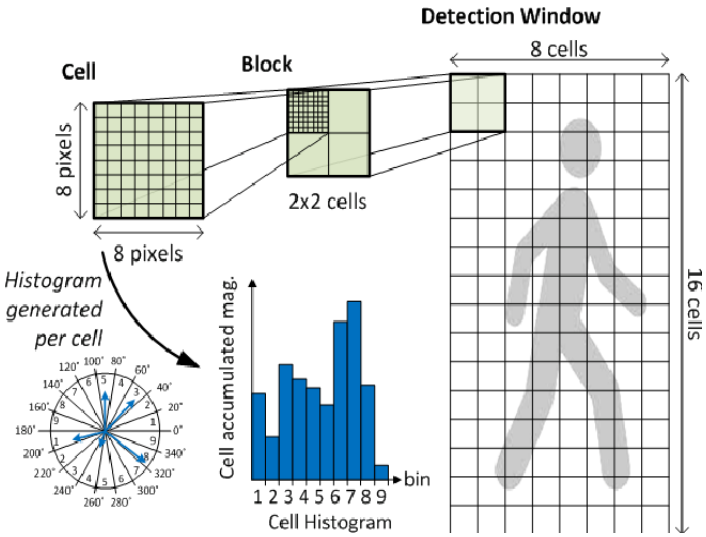
We Used the implementation that was suggested by papers [1] and [2] to compute the (16, 2) LBP as 16 is the neighboring pixels at a distance of 2 pixels from the central pixel. For 16 neighboring points, there is a total of (16*(16-2)-2) =242 uniform patterns. The LBP histogram, therefore, has 243 bins, 242 for the uniform patterns and 1 for all the non-uniform patterns



2. Histogram of Oriented Gradients

Histogram of Oriented Gradient (HoG) feature descriptor, vastly used for object detection The
key idea of HoG is to characterize an object by the distribution of edge (gradient) directions. Paper [2] shows that the recent studies shown that HoG features have effective performance for analysis of handwritten texts.

For implementation, the handwritten image is resized by (64,128) divided into cells with 8*8 pixels the image become 8*16 cells. For each pixel within a cell, the gradient vector (magnitude and direction) is computed. A histogram of orientations is then computed for each cell and the descriptor for the complete image is calculated by concatenating the histograms of all the cells which will have 7 *15 *36 =3780 dimensions as the block is 2*2 cell.



3. COLD "cloud of line distribution"

This feature relies on extract the curvature information of the handwriting from the contours. As discussed in papers [3,4].
Our algorithm goes into the following steps:

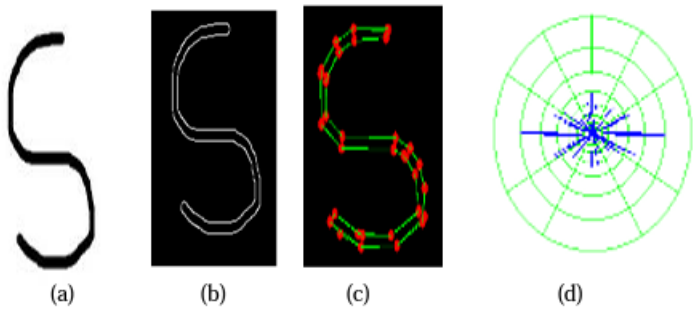
- Preprocessing step to binarize the image using adaptive thresholding method.
- Get the contours of the binary image using findCountours built-in in OpenCV.
- Then find the dominant points from the contours, connect every pair from the obtained points.
- Calculate in the polar domain: for every pair

$$\begin{cases} \Theta = \arctan\left(\dfrac{yi + k - yi}{xi + k - xi}\right) \\ \rho = \sqrt{(yi + k - yi)^2 + (xi + k - xi)^2} \end{cases}$$

(X_i, Yi) and $(X_{i+k,}\ Y_{i+k})$ are the coordinates of the dominant points P_i and $P_{i+k,}$ and each line has $\boldsymbol{\Theta}$ and $\boldsymbol{\rho}$ in the polar coordinates space.

- Hence until now COLD deals with the point-to-point way which founded that this way will be very sensitive to the noise.

- log-polar space has been used to overcome this problem to make it sensitive to the regions in the center not those further way from center, paper [2] gives an experimented-value for the used parameters in building the log-polar space
  "N_ANGLE_BINS"→ the number of the angular intervals = 12.
  "N_PHO_BINS"→ the number of distance intervals = 7.
  "R_INNER" → the distance between two consecutive rings in the log space = 5.
- The final feature vector is the normalized histogram.



4. Chain Code (unused feature)

This method was covered in the paper, and its idea is to create a code that expresses the sequence of pixels in the borders. But it has a bad result because it needs to link the paragraph completely.
We tried cropping the text into smaller versions and applying the feature algorithm on them but the only effective way was to handle each character by itself which is very difficult and computationally expensive so we ruled out this feature.

5. Hinge Feature (used feature)

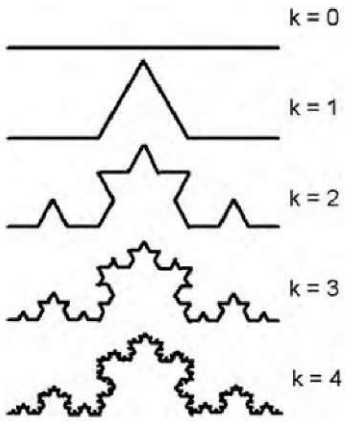
The hinge feature is a contour-based feature that is designed to get the curvature of the ink of the handwriting. The hinge feature calculates the pdf of the two phis of the curvature inside the window it calculates thus it can represent the curvature of the handwriting in a compressed manner. This feature was proved to be highly effective in writer identification problems which made the researchers use it for gender classification and also proved its effectiveness. Making a model based only on the hinge feature gave an accuracy of around 80%.

6. GLCM (unused feature)

The gray level co-occurrence matrix is a statistical method of examining texture that considers the spatial relationship of pixels. This feature was not helpful as it depended heavily on the lighting of the picture of the handwriting so the model based on it was really in-accurate.

7. Fractal Dimension feature:

Fractal dimension is a mathematical concept; that should describe the real dimension of the object, for example in this figure when K=4 the shape is a line (1D) but it has an area and the shapes these have area must be a 2D shapes then this shape is a 1D shape with some properties of the 2D shapes these properties we can describe with a float number could fractal dimension.
Paper [1] shows that the recent studies shown that fractal features have effective performance for analysis of handwritten texts.



We use box counting to approximately calculate the fractal dimension: to calculate the fractal dimension we should imagine this fractal lying on a spaced grid and count how many boxes are required to cover the set, then repeat this again with smaller scale for the grid, then we fit a line between the scale and number of boxes in this scale then with the following equation we can calculate the dimension.

$$-D = \frac{\log N}{\log 1/\varepsilon}$$

since N is the number of boxes and ε is the scale of the grid.
Finally, we didn't consider this feature because it wasn't promising it's accuracy was less than 50%.

8. Skew detection feature:

Skew detection calculates the angle of line with the horizontal. Paper [5] mention this feature that can effectively describe handwritten text then we can classify gender with it. Using Hough Transform we could calculate the angle for each line then the sum of these angles represent the feature.
Finally, we didn't consider this feature because it wasn't promising it's accuracy was less than 55%.

Model Selection/Training Module

1. The Random Forest Classifier

Paper [2] suggested to use RF as one of the Classifiers. RF consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. In implementation number of trees is 500 and use "gini" as function to measure the quality of a split.

2. Shallow Neural Network (NN)

It is a type of the Neural networks with one or two hidden layers, that maps input data sets to a set of appropriate outputs. Each layer is fully connected to the

following layer and the nodes of the layer is neurons with nonlinear activation function.

Parameters used in the Shallow-NN classifier:

`solver='lbfgs'` → parameter used for the weights given to the nodes.

`hidden_layer_sizes= (2,)` → specify one hidden layer with 2 epochs "neuron"

`random_state=1)` → Determines random number generation for weights and bias initialization.

`activation='relu')` → returns f(x) = max (0, x)

for the cold feature → we used two hidden layers.

3. KNN

K Nearest Neighbors algorithm one of the most famous supervised learning algorithms that uses to classification and regression. To classify new point in KNN we vote k nearest neighbor points classes, then point classify as most common class in voting. The nearest points are the points with minimum euclidean distance.

4. SVM

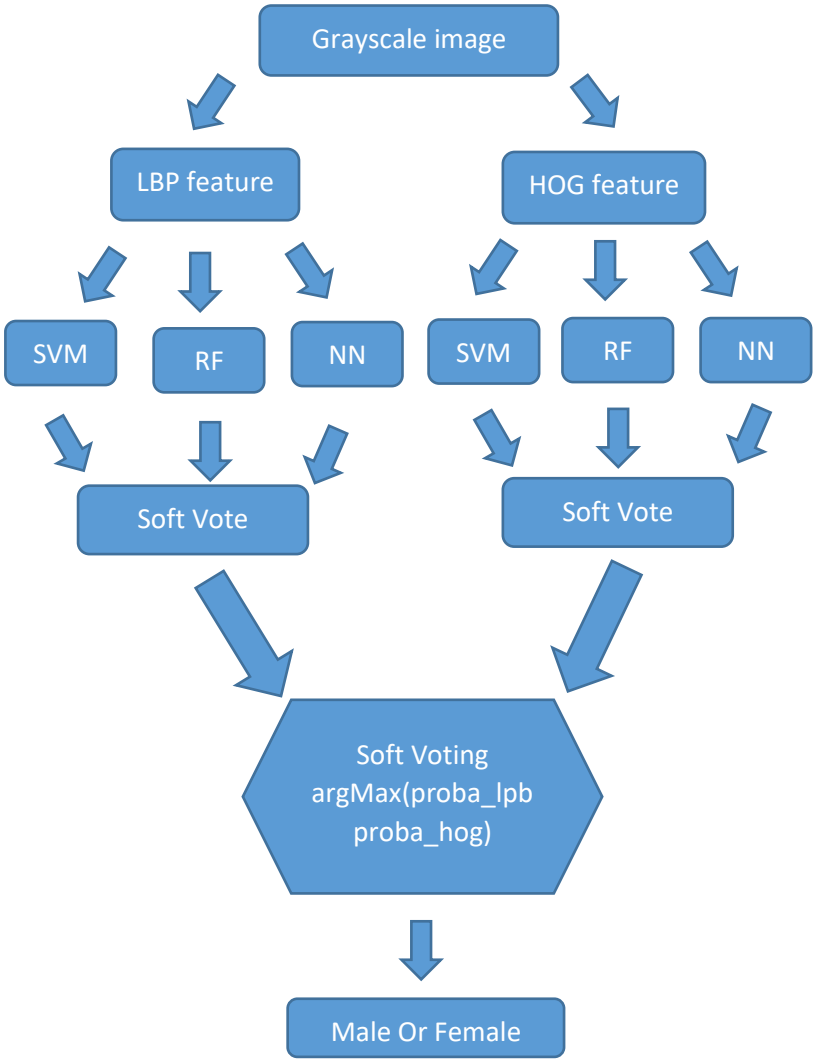
Support Vector Machine is a machine learning algorithm that is heavily used in classification problems as it is way more sophisticated than linear and logistic regression, better than shallow neural networks, and faster than deep neural networks like CNN. It is mainly used when training and inference time needed to be minimal or because of not having the required hardware for deep neural networks training.

SVM works by finding the most optimum hyperplane (int the n-dimension of features) having the largest gap to the feature vectors. If the algorithm failed to find the optimum hyperplane it projects the data to a higher dimension, finds the plane and then scales back to the original n-dimension we had.

To get with our final approach we go through many approaches and trials

- Frist approach:

  is to use Two of Texture-based features local binary patterns (LBP) and Histogram of Oriented Gradients (HOG). The preprocessing for two classifiers was to convert the image to grayscale then apply (16, 2) LBP as 16 is the neighboring pixels at a distance of 2 pixels from the central pixel. Get vector with 243 dimensions. On the same gray image apply HOG with cells with 8*8 pixels to get histogram with 3780 dimensions. After get the features apply three classifiers on each feature using soft vote to get one decision form each feature. The classifiers are Support Vector Machine (SVM), Random Forest (RF) and shallow Neural Network (NN). After get decision form each feature calculate the sum of tow decision prediction the get the max argument for each image to get the vote from the two features.



Without the final stage, HOG got accuracy between 64-72% and LPB got accuracy between 57-67% and was biased to the males. After the final voting stage, the accuracy became between 68-74%

- Second approach:
  Was to add COLD feature to the previous approach and now the final vote is between 3 features and reached an accuracy between 69-75%.

We tried to add the hinge feature to the previous approach that result in increasing in the time and didn't show a notable change in the accuracy.

- Third approach:
  Tried extract hinge feature after applying preprocessing to convert the image to binary to eliminate all the noise and different in light then used SVM classifier only and get high accuracy between 70-80%

- Fourth approach:
  Used COLD feature and apply PCA for this feature to reduce the dimensionality from 420 dimension to 70 dimension this number was selected after trials on validation set and that get more accuracy then use the all dimensions

Criteria used for choosing the best Model

Our criteria are to choose the feature which gets the highest accuracy (Hinge) and get its model then select the feature with the second higher accuracy (COLD with PCA) and get its model each model of them when use in the test each model gets predict probability sum the value of the two models then get the max argument. Apply Soft Voting between the two models to get a better model with an accuracy of 73-81%. With this sequential forward criteria, we choose the third-best model that was HOG feature classified with SVM and apply soft voting with the three models and get better accuracy between 72-82%. Adding more feature after that slow the model and does not make a markable change in the accuracy.

Performance Analysis Module

Cold Accuracy	Hinge Accuracy	Hog Accuracy	Hinge + Cold Accuracy	Soft Voting Accuracy
80%	78%	70%	85%	85%
81%	84%	69%	86%	88%
83%	85%	72%	89%	90%

Enhancements and Future work

To develop our model, we trained each feature alone then select higher accuracies and vote to get the class, to enhance this we can try sequential backward selection: that we try voting system with all features then try to elements the less effective features. And we can also begin to explore new features like: Pixel counting, Circle counting, Perimeter, Region properties.

One of the most important thing to enhance our model is to begin to use other languages like Arabic thus there is some features not so effective in English but we predict this features could be effective in languages like Arabic; for example, chain code feature that we couldn't use in English because each character is separate but in Arabic we can apply this feature on the whole word.
Finally, more data more accuracy then we aim to find more data and train our model on these data.

Work Load Distribution

	Raghad	Mohammed	Menna	Nada
Features (each one implement his/her features)	- Local Binary Patterns (LBP) - Histogram of Oriented Gradients (HoG) - COLD Feature	- Co-occurrence Matrices (GLCM) - Segmentation-based Fractal Texture Analysis (not implemented) - Hinge Feature	- Chain code-based features - Polygon-based features (not implemented) - COLD Feature	- Box counting FD - Skew angle - Hinge Feature
Implemented code	Image preprocessing	Image preprocessing		"pridect.py" script
report	Project Pipeline	Preprocessing Module	Performance Analysis Module	- Enhancements and Future work - Work Load Distribution

Papers

[1] Automatic analysis of handwriting for gender classification
[2] Improving Handwriting based Gender Classification using Ensemble Classification
[3] Writer identification using curvature-free features
[4] A New COLD Feature based Handwriting Analysis for Ethnicity/Nationality Identification.
[5] A method for automatic classification of gender based
on text- independent handwriting

References

[1] Shallow NN