

Analyzing Handwriting for Gender, Handedness and Combined Prediction using Machine Learning Algorithms

Ayushi Agarwal¹ and Mala Saraswat²

¹⁻²ABES Engineering College, Ghaziabad, India

Email: agarwalayushi028@gmail.com, mala.saraswat@abes.ac.in

Abstract—Studies related to the handwriting recognition has got good pace in recent time. Various research works were going on for recognition of handwriting in online and offline mode both. Manually recognition of handwriting is also an approach people are working on. A prediction based on the handwriting includes recognizing gender, age, handedness, behavior, basic human mood, nationality, ethnicity and many more. For this various predefined datasets are available. The available datasets are in many different languages. So fetching the required dataset of a particular language data is a tedious job to perform. In various other works the researchers had used different data as per their requirements. Here in this work of ours we are working with two different datasets. A new task that we had taken up here is we have created our own dataset. Thus here, the Machine learning algorithms were performed for predicting the traits firstly on IAM dataset of English data samples and another Real time dataset also in English writing scripts.

Index Terms— Machine Learning, Support Vector Machine, Algorithm, Image processing, Random Forest.

I. INTRODUCTION

The technology is becoming advance day by day and capable enough to work and perform activities similar to the human behavior. In expansion to this the researchers has got their interest in using the technology for understanding the human characteristics and use this study in designing and training the machines. Manufacturing of machine models with this theme has resulted in creation of robots and various self-evaluating gadgets where initially they are trained by humans but after then they independently without any human intervention. Leading to this and looking upon the real life time of the current phase the researchers had worked in understanding the human characteristics and traits with the help of technology which includes Machine learning and Deep learning and this research inventions has evolved a new tech field which now a days is known as Artificial Intelligence.

Here in this work of ours we have worked on various Machine Learning algorithms for predicting gender, handedness and combined handedness and gender. For achieving our goal we had implemented the algorithms on two different datasets. Firstly we had extracted handwriting manuscripts samples of IAM dataset of English manuscripts. Secondly we have created a real time dataset.

II. LITERATURE REVIEW

When comes about the related works regarding for the defined task of predicting the gender and handedness by handwriting recognition. It has been observed that there are differences in handwriting of each individual and specific difference among male and female as well as to handedness of either group of people. With respect to the characteristic differences among the handwriting of each gender group each can be predicted.

In the previous works, firstly by Bi, N., Suen, C. Y., Nobile, N., & Tan, J. (2019), [1], with the implementation based upon Kernel Mutual Information (KMI) and SVM for predicting gender got 72 percent successful by using ICDAR English and RDF Chinese datasets, in each dataset.

In another work by Ahmed, M., Rasool, A. G., Afzal, H., & Siddiqi, I. (2017), [2], by using ICDAR and QUWI datasets implementing with ANN, SVM and KNN discriminated gender got successful with 69 percent with KNN and 70 percent successful with SVM.

Gattal, A., Djeddi, C., Siddiqi, I., & Chibani, Y. (2018), [3], in their work they used oriented basic image features (oBIFs) histograms and SVM and 72 percent successfully predicted gender. Thus in the works of other writers as well where they had used the ML techniques and algorithms got maximum success up to 72 percent for predicting gender and approximately 89 percent for handedness. This much of high percent success also depend upon the large set of data as more the data for training and testing more is the accuracy rate.

While coming to the works for the defined tasks by DL algorithms and models [15] the accuracy rate got increased up to around 80 percent for gender prediction and 90 percent in handedness prediction and combined accuracy was about 83 percent with IAM English dataset which consists of 115,320 isolated words from the handwriting samples collected from 657 writers. In the other works with different approaches [11,12,13] had reached up to 80.72 percent maximum accuracy in case of gender prediction and 100 percent prediction accuracy for handedness [14] by using IAM dataset.

III. PROPOSED METHODOLOGY

For over recognition task from handwriting scripts as stated in the previous section that we have planned for predicting the traits with the help of two datasets, they are IAM dataset and Real-time dataset. With both the datasets image data processing steps have been followed. Few of the steps were avoided in case of IAM dataset because it is a predefined dataset and available on open sources for data extraction. Talking about the Real-time dataset, as it is our own collected and created data so here each step of image processing is followed to make the data or scanned handwriting scripts sample in machine under stable form. As this would enhance our model and give the desired prediction outputs.

Then with both the datasets experiments were carried out in three phases. First implementation is only for predicting the gender from the handwriting and second is the prediction of handedness that is whether the person is right handed person or the left handed person respectively. And the third part is the combined prediction where gender and handedness is predicted in combination. For doing this the data is treated with four parameters they are: RM- Right handed male, RF- Right handed female, LM- Left handed male and LF- Left handed female.

In various other works of research none of the researcher has created their own dataset. Most of the frequently used datasets for doing predictions with handwriting recognition are IAM dataset, KHATT dataset, QUWI dataset and others. They are works done with different languages as well. Here we have focused for English manuscripts as we found the common writing language to our locality is English. And finally after all pre-steps we have implemented our data with machine learning algorithms. We have selected four algorithms they are: SVM, KNN, RFC and NAÏVE BAYES. All are supervised machine learning algorithms. With this we have also found that altogether four algorithms implementation for gender, handedness and combined prediction is nearly non. Few had done works with two algorithms but not more than that.

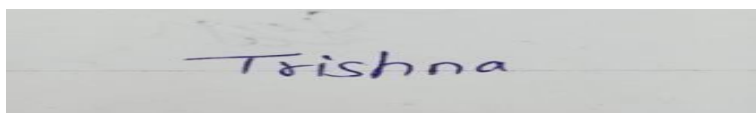


Fig.1. sample image of real data collected

IV. EXPERIMENTAL RESULTS

Here in this section you all will see the experimental results for gender, handedness and combined prediction in three sections. In the section 4.1, there are the results for only gender prediction. In section 4.2, there are results

for handedness prediction only. And in the last 4.3, there are results for combined prediction along with their respective resultant bar graphs.

For IAM dataset, according to the official information for this dataset it comprises of around 655 people's handwriting samples which were further partitioned as sentences collection and words collection along with their traits details so as to rectify our implemented prediction results. Here, we could not extract whole dataset due to the limitation of our system but we have successfully achieved around 4899 data samples out of which 1000 were female data sample and 3899 male. Out of which only 552 people were left handed in the ratio of 220 female and 332 male left handed people and rest all were right handed data. We have performed training with 3233 data samples and testing with around 900-1000 data in each phase.

Moving on to discuss about Real-time dataset, here the writing samples of 2900 pupils were collected. Among them 1900 were male gender pupil and 1000 female gender, 2600 right handed and 300 left handed. For male gender prediction training is done with 2900 data and 2000 data for testing. For female gender 2700 data is used for training and 2300 is used for testing. For right handedness 2800 data is used for training and 2420 data is used for testing. For left handedness 980 data is used for training and 300 data is used for testing. When comes to the combined prediction for right handed male 2690 data is used for training and 1473 data is used for testing. For left handed male 529 data is used for training and 180 used for testing. For right handed female 2430 data is used for training and 1155 data is used for training and lastly for left handed female 740 data is used for training and 280 data is used for testing.

In comparison to the previous works done the data so far used was large enough in quantity that the direct implementation could be performed. But here our extracted data and our real time data was not enough in comparison to previous works. So to expand our data we have used data augmentation in online mode. It is so because in online data augmentation extra storage space is not needed. It happens simultaneously while execution is processed.

A. Experimental Results for Gender Prediction

TABLE: I. GENDER PREDICTION WITH IAM DATASET

GENDER	ALGO USED	PRECISION	RECALL	F-SCORE
MALE GENDER	NAÏVE BAYES	71.42	85.46	77.81
	KNN	71.1	80.63	75.56
	RFC	70.36	80.64	75.15
	SVM	71.21	85.1	77.53
FEMALE GENDER	NAÏVE BAYES	62.15	41.56	49.81
	KNN	60.46	40.34	48.39
	RFC	59.14	40.1	47.79
	SVM	61.54	41.43	49.52

TABLE: II. AVERAGE ACCURACY WITH IAM DATASET FOR GENDER PREDICTION

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	68.82
KNN	66.34
RFC	66.10
SVM	68.65

TABLE: III. GENDER PREDICTION WITH REAL-TIME DATASETS

GENDER	ALGO USED	PRECISION	RECALL	F-SCORE
MALE GENDER	NAÏVE BAYES	70.56	81.54	75.65
	KNN	69.71	81.03	74.94
	RFC	69.56	81	74.84
	SVM	71.63	81.33	76.17
FEMALE GENDER	NAÏVE BAYES	60.49	40.03	48.17
	KNN	57.35	40.64	47.57
	RFC	59.03	39.09	47.03
	SVM	60.3	40.4	48.38

TABLE: IV. AVERAGE ACCURACY WITH REAL-TIME DATASET FOR GENDER PREDICTION

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	68.32
KNN	67.03
RFC	66.56
SVM	68.45

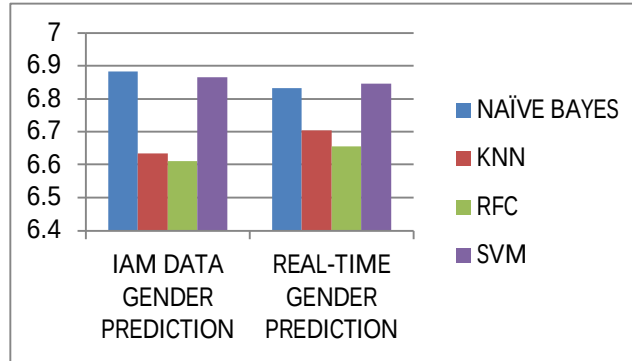


Fig: 2.Graph for Average Accuracy for Gender prediction with IAM and Real- time Datasets respectively

From the results observed, it can be observed that our own dataset has matched up with the previous works. And our own data set has also resulted good prediction in comparison to the IAM predefined dataset. The average accuracy achieved with both the datasets was approximately 69%. The Naïve Bayes and SVM algorithms has resulted good results.

B. Experimental Results for Handedness Prediction

TABLE: V HANDEDNESS PREDICTION WITH IAM DATASET

HANDEDNESS	ALGO USED	PRECISION	RECALL	F-S CORE
RIGHT HANDED	NAÏVE BAYES	81.43	73.96	77.51
	KNN	80.25	72.09	75.95
	RFC	82.76	75.05	78.71
	SVM	82.66	74.24	78.22
LEFT HANDED	NAÏVE BAYES	52.45	42.63	47.03
	KNN	52.05	41.64	46.26
	RFC	50.9	40.4	45.05
	SVM	54.43	43.73	48.49

TABLE: VI AVERAGE ACCURACY WITH IAM DATASET FOR HANDEDNESS PREDICTION

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	76.43
KNN	73.54
RFC	73.25
SVM	76.23

Here, with the IAM dataset the maximum accuracy achieved is approximately 77% and with real-time data set the maximum accuracy achieved is approximately 71%.

C. Experiment for Combined Prediction of Traits

In the next coming up summation of the observation is a combined prediction data. Both gender and handedness are predicted in combination. For this the attributes are segregated as four characters namely, RM that is right handed male. For doing the training of this particular category the data was available with us in good quantity, next comes the LM that is left handed male, RF that is right handed female and LF that is left handed female.

TABLE: VII HANDEDNESS PREDICTION WITH REAL-TIME DATASET

HANDEDNESS	ALGO USED	PRECISION	RECALL	F-SCORE
RIGHT HANDED	NAÏVE BAYES	88.56	70.98	78.8
	KNN	84.73	71	77.25
	RFC	85	70.34	76.97
	SVM	88.64	70.24	78.37
LEFT HANDED	NAÏVE BAYES	15.7	40.39	22.61
	KNN	15.25	39.98	22.07
	RFC	15.05	40.97	22.01
	SVM	15.64	40.4	22.55

TABLE: VIII AVERAGE ACCURACY WITH REAL-TIME DATASET FOR HANDEDNESS PREDICTION

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	71.54
KNN	69.51
RFC	69.22
SVM	71.35

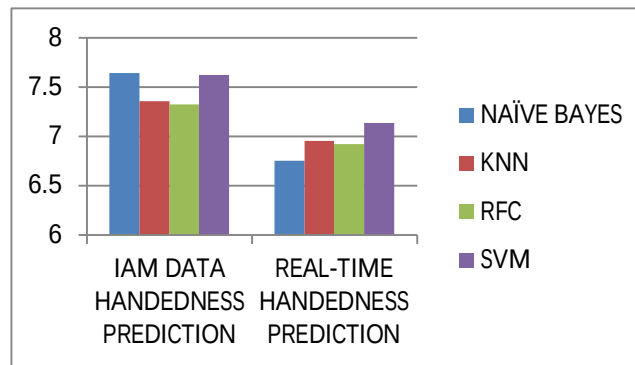


Fig: 3.Graph for Average Accuracy for Handedness prediction with IAM and Real-time Datasets respectively

While working for creating the real-time dataset it was observed that the right handed data is easily available, while struggle comes when comes to the left handed data collection. Still we managed and collected good amount of left handed handwriting sample for training and testing.

For the IAM dataset the data is already available to us in different folders according to their prediction of trait. So, we do not have to struggle for segregating the data according to the feature for which we were doing implementation.

TABLE: IX. COMBINED PREDICTION WITH IAM DATASET

PREDICTION	ALGO USED	PRECISION	RECALL	F-SCORE
RIGHT HANDED MALE(RM)	NAÏVE BAYES	64.4	71.64	67.82
	KNN	63.76	71.09	67.22
	RFC	63.62	71.32	67.25
	SVM	64.56	71.56	67.88
LEFT HANDED MALE(LM)	NAÏVE BAYES	50.8	48.95	49.85
	KNN	51	46.72	48.76
	RFC	51	48	49.45
	SVM	51.24	50	49.99
RIGHT HANDED FEMALE(RF)	NAÏVE BAYES	59.64	54.46	56.93
	KNN	57.34	52.09	54.58
	RFC	57.02	52.61	54.72
	SVM	59.1	52.6	55.66
LEFT HANDED FEMALE(LF)	NAÏVE BAYES	61.22	33.33	43.16
	KNN	60.59	34.43	43.9
	RFC	60.9	34.52	44.06
	SVM	61.64	34.64	44.35

TABLE: X. AVERAGE ACCURACY FOR COMBINED PREDICTION WITH IAM DATASET

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	74.60
KNN	74.34
RFC	74.21
SVM	75.65

TABLE: XI. COMBINED PREDICTION WITH REAL-TIME DATASET

PREDICTION	ALGO USED	PRECISION	RECALL	F-S CORE
RIGHT HANDED MALE(RM)	NAÏVE BA YES	54.75	35.25	42.89
	KNN	53.02	34.03	43.79
	RFC	53.25	34.36	41.76
	SVM	54.46	35.89	43.26
LEFT HANDED MALE(LM)	NAÏVE BA YES	15.96	12.96	14.3
	KNN	13.12	14.63	13.83
	RFC	11.69	14.57	13.13
	SVM	14.63	12.56	13.51
RIGHT HANDED FEMALE(RF)	NAÏVE BA YES	40.34	40.4	40.36
	KNN	39.09	39.39	39.23
	RFC	39.35	39.2	39.27
	SVM	40.65	40.62	40.63
LEFT HANDED FEMALE(LF)	NAÏVE BA YES	6.74	23.21	10.44
	KNN	6.54	21	9.97
	RFC	5.54	21.05	8.77
	SVM	3.76	25.8	6.95

TABLE: XII. AVERAGE ACCURACY FOR COMBINED PREDICTION WITH REAL-TIME DATASET

ALGORITHMS USED	AVERAGE ACCURACY
NAÏVE BAYES	72.12
KNN	70.64
RFC	70.22
SVM	72.09

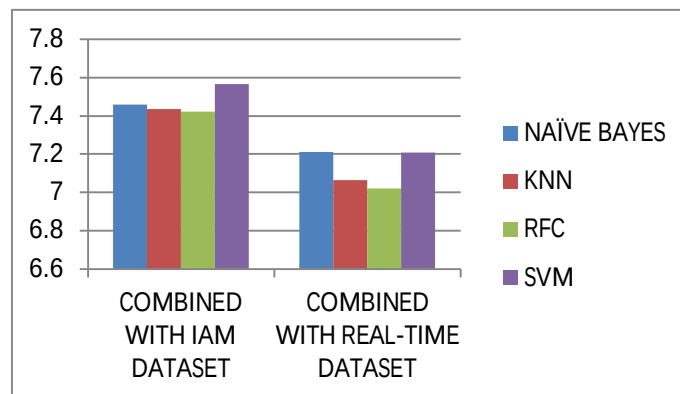


Fig: 4.Graph for accuracy for the combined prediction with IAM and Real-time dataset

Here, the average accuracy with the IAM dataset is approximately 76% where with the Real-time dataset the average accuracy is approximately 72%. Further moving on in all the cases SVM and NAÏVE BAYES algorithm has showed good results. Our real-time dataset have nearly matched the predefined dataset's performance. Online data augmentation has given more support for the good prediction rate. With ML algorithms the accuracy rate for SVM algorithm is good in each case in comparison with other algorithms of machine learning. After SVM, its Naïve Bayes followed by KNN and then RFC. For RFC it was always stated that more the number of trees more accurate will be the result. But if we observe here for our prediction task it was least accurate for the purpose. For KNN, it was mentioned that it is good with missing data that it if some information is incomplete than also it would perform well. But according to our observation it is not too accurate

for prediction tasks. Talking about the Naïve Bayes and SVM, these two supervised machine learning algorithms are good for performing tasks in comparison to other algorithms. But their accuracy rate is not the limit, it can be enhanced if data augmentation performed or the same data is implemented with deep learning techniques.

V. CONCLUSION

Real datasets can also be a good approach to deal with the prediction rather than only utilizing the predefined data set as the system characteristics do matters for dealing with the huge data. Dealing with real data has provided good understanding for image processing, handwriting recognition, data manipulation, and feature extraction and data augmentation for online and offline mode.

Yes there is capacity to increase the accuracy if the deep learning models are implemented for the prediction since they more compatible to deal with images. This we have covered in our next part of the paper.

REFERENCES

- [1] Bi, N., Suen, C. Y., Nobile, N., & Tan, J. (2019). A multi-feature selection approach for gender identification of handwriting based on kernel mutual information. *Pattern Recognition Letters*, 121, 123-132.
- [2] Ahmed, M., Rasool, A. G., Afzal, H., & Siddiqi, I. (2017). Improving handwriting based gender classification using ensemble classifiers. *Expert Systems with Applications*, 85, 158-168.
- [3] Gattal, A., Djeddi, C., Siddiqi, I., & Chibani, Y. (2018). Gender classification from offline multi-script handwriting images using oriented basic image features (oBIFs). *Expert Systems with Applications*, 99, 155-167.
- [4] Al Maadeed, S., & Hassaine, A. (2014). Automatic prediction of age, gender, and nationality in offline handwriting. *EURASIP Journal on Image and Video Processing*, 2014(1), 1-10.
- [5] Marzinotto, G., Rosales, J. C., El-Yacoubi, M. A., Garcia-Salicetti, S., Kahindo, C., Kerhervé, H., ... & Rigaud, A. S. (2016). Age-related evolution patterns in online handwriting. *Computational and mathematical methods in medicine*, 2016.
- [6] Morera, Á., Sánchez, Á., Vélez, J. F., & Moreno, A. B. (2018). Gender and handedness prediction from offline handwriting using convolutional neural networks. *Complexity*, 2018.
- [7] Maken, P., Gupta, A., & Gupta, M. K. (2019). A study on various techniques involved in gender prediction system: a comprehensive review. *Cybernetics and Information Technologies*, 19(2), 51-73.
- [8] Gattal, A., Djeddi, C., Bensefia, A., & Ennaji, A. (2020, June). Handwriting based gender classification using cold and hinge features. In *International Conference on Image and Signal Processing* (pp. 233-242). Springer, Cham.
- [9] Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34-42).
- [10] Maken, P., & Gupta, A. (2021). A method for automatic classification of gender based on text-independent handwriting. *Multimedia Tools and Applications*, 1-30.
- [11] Rahmanian, M., & Shayegan, M. A. (2021). Handwriting-based gender and handedness classification using convolutional neural networks. *Multimedia Tools and Applications*, 1-24.
- [12] Siddiqi, I., Djeddi, C., Raza, A., & Souici-Meslati, L. (2015). Automatic analysis of handwriting for gender classification. *Pattern Analysis and Applications*, 18(4), 887-899.
- [13] Liwicki, M., A.Schlapbach, P.Loretan, H.Bunke. Automatic Detection of Gender and Handedness from On-Line Handwriting. – Journal of Social Psychology, 2007, No March, pp. 179-183
- [14] Liwicki, M., A.Schlapbach, H.Bunke. Automatic Gender Detection Using On-Line and Off-Line Information. – Pattern Analysis and Applications, Vol. 14, 2011, No 1, pp. 87-92.
- [15] Ángel Morera, Ángel Sánchez, José Francisco Vélez, and Ana Belén Moreno. Gender and Handedness Prediction from Offline Handwriting Using Convolutional Neural Networks. Hindawi Complexity, Volume 2018.
- [16] M. Liwicki, A. Schlapbach, and H. Bunke, “Automatic gender detection using on-line and off-line information,” PAA. Pattern Analysis and Applications, vol. 14, no. 1, pp. 87–92, 2011.
- [17] M. Liwicki, A. Schlapbach, P. Loretan, and H. Bunke, “Automatic detection of gender and handedness from on-line Handwriting,” in Proceedings of the 13th Conf. Intl Graphonomics Society, 2007.
- [18] N. Bouadjene, H. Nemmour, and Y. Chibani, “Histogram of Oriented Gradients for writer’s gender, handedness and age prediction,” in Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, INISTA 2015, IEEE, Madrid, Spain, August 2015.
- [19] IAM Homepage, <http://www.fki.inf.unibe.ch/databases/iamon-line-handwriting-database>.