CHAPTER-7

HYBRIDIZATION OF FEATURES AND CLASSIFICATION TECHNIQUES FOR GENDER CLASSIFICATION AND WRITER IDENTIFICATION SYSTEM

The development of the gender classification and writer identification system based on handwritten text in the Gurumukhi script is indubitably a demanding application for the handwriting and forensic-based research communities. As it is a novel assignment in the province of pattern recognition, the success of the systems will certainly influence and motivate the researchers for the development of applications such as left or right-handedness, stress, age, and nationality identification, etc. In this chapter, we are presenting the gender classification system and writer identification system using hybridization of both feature extraction and classification techniques which is a victorious and effective approach for handling the computational complexity of the data and enhancing the classification results. There are many techniques available for the hybridization of classification methods and in this chapter, the majority voting scheme has been implemented.

This is conclusively a unique perspective for designing such an application in comparison to the state-of-the-art work. The target of feature fusion is to explore a huge feature vector for recuperating the classification process and further hybridizing classification methods have certainly reinforced the accuracy rate. Section 7.2 deals with the performance of gender classification and writer identification without using the hybridization technique. Section 7.3 presents the concept of hybridization of feature extraction and classification techniques. Section 7.4 shows the experimental results received for the gender classification system with hybridization of both feature extraction and classification techniques followed by section 7.5 that represents the performance of the writer identification system with hybridization of both feature extraction and classification techniques. The chapter ends with the comparative analysis of the results attained in section 7. 6 followed by the discussion and conclusion.

7.1 INTRODUCTION

This chapter presents a unique perspective of evaluating gender classification and writer identification systems. In this method, feature extraction methods namely, Zoning, Diagonal, Transition, and Peak extent based method have been implemented followed by the classification of ANN, MLP, SVM, k-NN successively. Next, the hybridization of both feature extraction and classification techniques has been implemented to explore and to improve the classification results. Performance parameters are also evaluated to characterize the quality and capability of the system.

7.2 SYSTEM PERFORMANCE FOR GENDER CLASSIFICATION AND WRITER IDENTIFICATION

7.2.1 Dataset Collection

To study the strength of this novel combination for the experiment, offline handwritten samples of 100 female writers and 100 male writers have been collected i.e., a total of 200 writers has been taken. Each writer has written 35 primary characters in the Gurumukhi script, 10 sheets per writer, so the dataset consisting of 200×35×10=70,000 Gurumukhi characters, with 35,000 samples of female writers and 35,000 samples of male writers. After the data collection phase, we have Pre-processing phase that has performed digitization, normalization, and thinning on the offline handwritten Gurumukhi characters as shown in Figure 7.1. The transformation of digitized image into a thinned image has been accomplished using a number of phases and finally thinned images for further experimental achievements are generated that are shown in Figure 7.1.

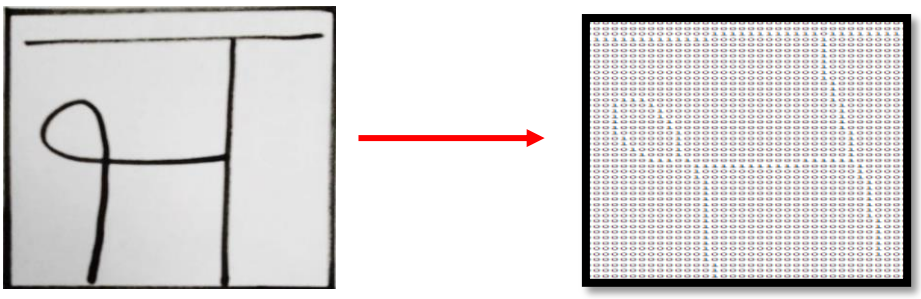

Figure 7.1. Transformation of Digitized image to Thinned image

7.2.2 Implementation of Feature Extraction and Classification Techniques

This section aims to experimentally evaluate the performance of feature extraction and classifiers without implementing hybridization techniques. So here in this subsection, Zoning, Diagonal, Transition, and Peak Extent-based feature extraction methods have been implemented individually on the processed Gurumukhi characters to extract the hidden features. These feature extraction techniques have already been discussed in detail in section 3.2.3. In the zoning feature extraction method, the character image is divided into equal-sized zones followed by the diagonal method that is based on locating the pixels on the diagonals, and then foreground pixels along the diagonals are added. The transition method will count the transition from white to black pixels and vice versa in both directions i.e., left to right and top to bottom. Peak Extent based feature extraction method will retrieve pixels on the horizontal and vertical peak extents.

For the classification, Artificial Neural Network, Decision Tree, Random Forest, Adaptive Boosting, and Multi-layered Perceptron have been implemented. Table 7.1 depicts a maximum gender classification accuracy of 90.92% with the Peak Extent-based feature extraction technique and Adaptive Boosting classification technique and the graphical view is shown in Figure 7.2.

Table 7.1. Gender Classification Accuracy

Feature Extraction and Classification Techniques	ANN (C1)	Decision Tree (C2)	Random Forest (C3)	Adaptive Boosting (C4)	MLP (C5)
Zoning (F1)	89.29%	60.35%	83.79%	85.25%	86.76%
Diagonal (F2)	87.90%	60.06%	83.04%	89.35%	86.69%
Transition (F3)	89.89%	61.25%	78.10%	81.13%	84.59%
Peak Extent (F4)	90.67%	66.77%	90.38%	**90.92%**	87.89%

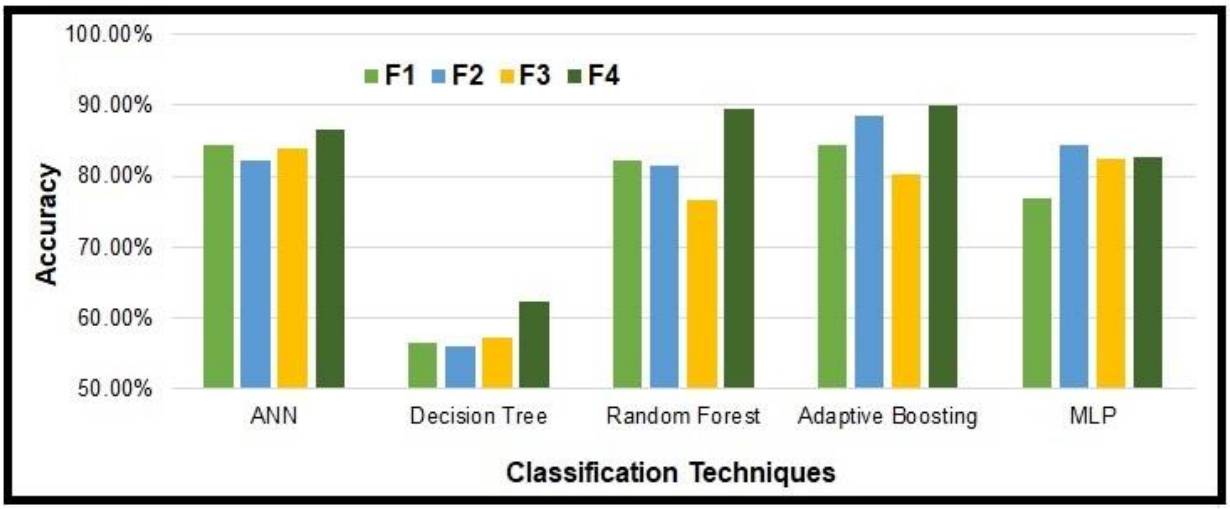

Figure 7.2. Gender Classification Accuracy without hybridization of Feature Extraction Techniques

For the identification of the writer, maximum accuracy of 90.02% has been revealed when implementing peak extent-based feature extraction approach with adaptive boosting classification as shown in Table 7.2 and the graphical view is shown in Figure 7.3. The goal here is to retrieve the results individually and then compare the accuracy achieved without implementing hybridization followed by evaluating the results with hybridization of both feature extraction and classification techniques. The comparison of the results shows the need, importance, and effect of hybridization on the accuracy rate.

Table 7.2. Writer Identification Accuracy

Feature Extraction and Classification Techniques	ANN (C1)	Decision Tree (C2)	Random Forest (C3)	Adaptive Boosting (C4)	MLP (C5)
Zoning (F1)	84.38%	56.40%	82.15%	84.41%	76.91%
Diagonal (F2)	82.15%	56.13%	81.41%	88.47%	84.47%
Transition (F3)	84.01%	57.24%	76.57%	80.33%	82.47%
Peak Extent (F4)	86.61%	62.40%	89.59%	**90.02%**	82.75%

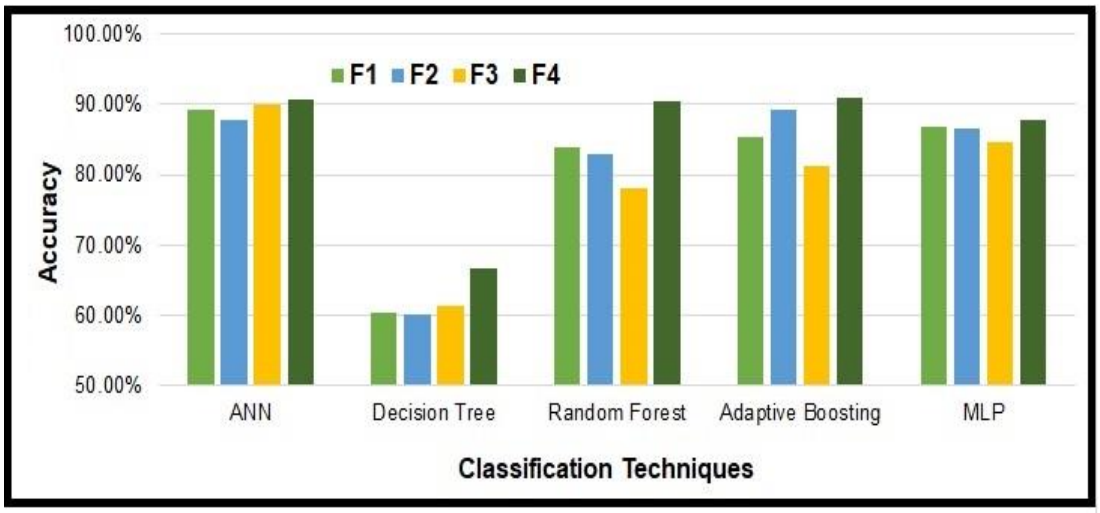

Figure 7.3. Writer Identification Accuracy without hybridization of Feature Extraction Techniques

7.3 HYBRIDIZATION OF FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES

This section describes the advantage of a novel computation scenario i.e., hybridization of feature extraction and classification techniques, which is the key concept of this chapter. Hybridization of feature extraction techniques has been proven to amazing results for machine learning and pattern recognition-based applications. It is also sometimes called a fusion of feature extraction algorithms. Higher the feature values explored for a character, more evidently and precisely the classifier will perform the classification with a large number of feature values and hence higher accuracy rate can be achieved in the classification process. If the Zoning phase (F1) produces 85 feature values and the Transition method (F2) produces 85 feature values and F1+F2 will retrieve 170 feature values for every character that will be fed to the classifiers so that no instance of the character will be left unstipulated. With 170 features for a character in comparison to 85 features has enhanced the chances of identification and classification.

Khanduja *et al.* (2015) presented a hybrid approach for generating a feature vector for the Devanagari script on CASIA and HWDB datasets by implementing loop feature and curve-fitting based feature extraction method with neural network classifier and revealed 93.4% accuracy in recognition of character. Goel and Vishwakarma (2016) presented hybridization of Discrete Wavelet Transform and

Discrete Cosine Transform to generate a feature vector with an SVM classifier and produce a promising accuracy rate. Here Table 7.3 presents the number of features extracted from Zoning, Diagonal, Transition, and Peak Extent methods separately and Table 7.4. presents feature values obtained after hybridization. Here F1+F2 means hybridizing feature values extracted from the Zoning and Diagonal methods similarly F1+F2+F3+F4 means hybridizing feature values extracted from the Zoning, Diagonal, Transition and Peak extent methods, etc. Hybridization of classification techniques ensembles the classification results and generates the maximum strength of the classifiers, which in this proposed experiment, has been achieved using a majority voting scheme. The benefits of both majority voting and hybridizing classification process have already been discussed in section 6.3.

Table 7.3. Number of Features extracted without hybridization of Features

Feature Extraction Techniques	Number of features obtained
Zoning (F1)	85
Diagonal (F2)	85
Transition (F3)	85
Peak Extents (F4)	**170**

Table 7.4. Number of Features extracted with hybridization of Features

Hybridization of Feature Extraction Techniques	Number of Features
F1+ F2	85+85=170
F1+ F3	85+85=170
F1+ F4	85+170=255
F2+ F3	85+85=170
F2+ F4	85+170=255
F3+ F4	85+170=255
F1+ F2+ F3	85+85+85=255
F1+ F2+ F4	85+85+170=340
F1+ F3+ F4	85+85+170=340
F2+ F3+ F4	85+85+170=340
F1+ F2+ F3+ F4	**85+85+85+170=425**

7.4 SYSTEM PERFORMANCE FOR GENDER CLASSIFICATION BASED ON HYBRIDIZATION OF FEATURES AND CLASSIFICATION TECHNIQUES

With hybridization of Zoning, Diagonal, Transition and Peak Extent Based feature extraction methods and hybridization of classification methods such as ANN, Decision Tree, Random Forest, Adaptive Boosting, and MLP, maximum accuracy of 94.27% has been attained for gender classification with hybridization of feature and classification techniques with majority voting as shown in Table 7.5 and graphically it is represented in Figure 7.4. In comparison to the implementation which was done before hybridization, i.e. classifiers when executed individually, then the maximum accuracy retrieved was recorded as 90.92% in Table 7.1. So hybridization has strengthened the accuracy of the classification as depicted in Table 7.5.

Table 7.5. Gender Classification Accuracy with hybridization of Feature Extraction and Classification Techniques

Features Extraction & Classification techniques	ANN (C1)	Decision Tree (C2)	Random Forest (C3)	Adaptive Boosting (C4)	MLP (C5)	**C1+C2+C3+C4+C5**
Zoning (F1)	89.29%	60.35%	83.79%	85.25%	86.76%	90.29%
Diagonal (F2)	87.90%	60.06%	83.04%	89.35%	86.69%	87.90%
Transition (F3)	89.89%	61.25%	78.10%	81.13%	84.59%	89.89%
Peak Extent (F4)	90.67%	66.77%	90.38%	90.92%	87.89%	92.67%
F1+F2+F3+F4	91.27%	70.76%	91.51%	92.12%	89.68%	**94.27%**

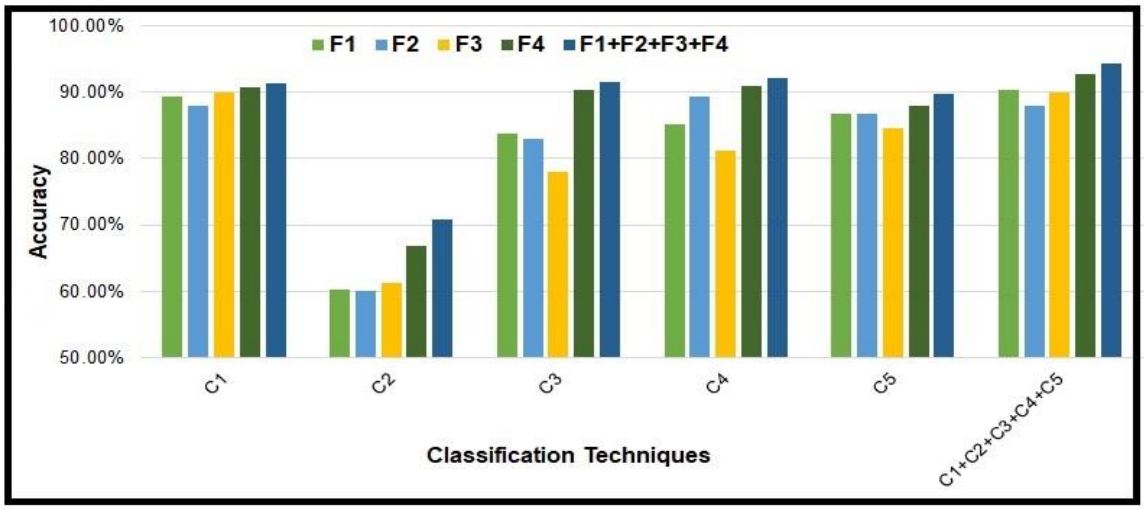

Figure 7.4. Gender Classification Accuracy with hybridization of feature extraction and classification techniques

7.5 SYSTEM PERFORMANCE FOR WRITER IDENTIFICATION BASED ON HYBRIDIZATION OF FEATURES AND CLASSIFICATION TECHNIQUES

With hybridization of Zoning, Diagonal, Transition, and Peak Extent based feature extraction methods and hybridization of classification methods such as ANN, Decision Tree, Random Forest, Adaptive Boosting and MLP, maximum accuracy of 91.93% has been attained for writer identification with hybridization of feature extraction and classification techniques using majority voting as shown in Table 7.6 and in Figure 7.5 and without hybridization, writer identification accuracy recorded was 90.02% with Peak extent based feature sand Adaptive boosting classifier as shown in Table 7.2.

Table 7.6. Writer Identification Accuracy with hybridization of Feature Extraction and Classification Techniques

Features Extraction & Classification techniques	ANN (C1)	Decision Tree (C2)	Random Forest (C3)	Adaptive Boosting (C4)	MLP (C5)	**C1+C2+C3+ C4+C5**
Zoning (F1)	84.38%	56.40%	82.15%	84.41%	76.91%	87.07%
Diagonal (F2)	82.15%	56.13%	81.41%	88.47%	84.47%	88.33%
Transition (F3)	84.01%	57.24%	76.57%	80.33%	82.47%	87.91%
Peak Extent (F4)	86.61%	62.40%	89.59%	90.02%	82.75%	91.04%
F1+F2+F3+F4	88.10%	66.13%	90.70%	91.21%	86.36%	**91.93%**

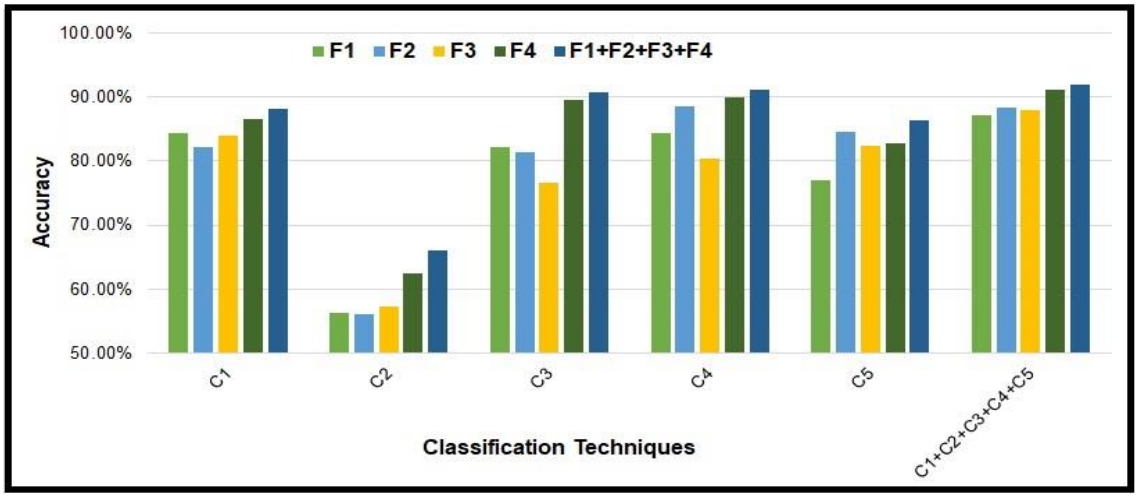

Figure 7.5. Writer Identification Accuracy with hybridization of Feature Extraction and Classification Techniques

7.6 COMPARATIVE ANALYSIS

In Table 7.1 and Table 7.5, we have presented the accuracy rate for the gender classification system, without using the hybridization technique and later on the hybridized approach both for feature extraction and classification using a majority voting scheme. The accuracy rate for the gender classification system has been enhanced from 90.92% to 94.27% Again, the accuracy achieved for the identification of the writer has been now enhanced and boost up from 90.02% to 91.93%. For writer identification, experimental results are presented in Table 7.2 and Table 7.6. The graphical representation of the results imparting the strength of hybridization on the writer identification and gender classification accuracy rate is described in Figures 7.1 to 7.5. Final results for gender classification and writer identification using hybridization of feature extraction and classification techniques are summarized in Table 7.7.

Table 7.7. Comparative Analysis of results achieved for Gender Classification and Writer Identification System

Nature of System	Feature Extraction Techniques	Classification Techniques	Number of Writers	Accuracy Achieved
Gender Classification	Zoning (F1), Diagonal (F2), Transition (F3), Peak Extent(F4)	ANN, MLP, DT, RF, AdaBoost	200 with 200×35×10=70,000 Gurumukhi samples	**90.92%**
Writer Identification				**90.02%**
Gender Classification	Hybridization of Zoning (F1), Diagonal (F2), Transition (F3, Peak Extent(F4)	Hybridization of ANN, MLP, DT, RF, AdaBoost using majority voting scheme	200 with 200×35×10=70,000 Gurumukhi samples	**94.27%**
Writer Identification				**91.93%**

7.7 DISCUSSION AND CONCLUSION

In this chapter, we implemented a novel approach i.e., hybridization of both the feature extraction and classification techniques. In the previous chapters, we used hybridization of feature extraction techniques in chapter 5, and hybridization of classification techniques has been successfully performed in chapter 6 but in this chapter, we used a novel computing paradigm of implementing and experiencing the

results with hybridization of both the approaches. Such execution of novel combination really revealed better results as shown in Table 7.7.

As in our previous experiments, this experiment has been implemented using Zoning, Diagonal, Transition, and Peak Extent based feature extraction methods and for classification techniques, artificial neural network, Decision Tree, Random Forest, Adaptive Boosting, and Multi-layer Perceptron have been experienced and the maximum accuracy rate for gender classification and writer identification has been reported as **94.27% and 91.93%** for gender classification and writer identification system respectively which is a promising and successful task for 200 writers and 70,000 Gurumukhi characters in the proposed experiment.