CHAPTER-6

GENDER CLASSIFICATION AND WRITER IDENTIFICATION SYSTEM BASED ON CURVE FITTING AND INTERSECTION & OPEN-END POINT FEATURES

The development of gender classification and writer identification systems mainly depends on the extraction of hidden features from the handwritten characters of writers. In chapter 4, we have efficiently explored Zoning, Diagonal, Transition, and Peak extent methods for extracting the features and attained satisfactory results. In this chapter, we have implemented two influential, novel, and effective techniques of extracting features i.e., the effect of curve fitting & Intersection and Open-End Point based feature extraction methods, which is discussed in detail in section 6.2. There is another novel accomplishment in this chapter i.e., the hybridization of classification techniques using the majority voting scheme, which is described in sections 6.3 and 6.4. For the classification, K-NN, SVM, Decision Tree, and Random Forest methods have been implemented. In section 6.5, the experiment has been performed for gender classification with hybridization of classification techniques based on curve fitting and Intersection & open endpoint features. In section 6.6, the experiment has been performed for the writer identification with hybridization of classification techniques based on curve fitting and Intersection & open endpoint features. Comparative analysis has been analyzed in section 6.7 and final results are summarized at the end of the chapter in section 6.8 in Table 6.11.

6.1 INTRODUCTION

In this chapter, we have been implemented two promising feature extraction strategies for the gender classification and writer identification system as discussed in section 3.2. For the experimental evaluation, a corpus having offline handwritten data samples from 200 writers has been generated, out of which 100 are female writers and 100 are male writers. So, the representation of the dataset comprising 70,000 Gurumukhi characters is shown in Table 6.1.

Table 6.1. Dataset Description for Gender Classification

Number of Writers	Number of Characters of Gurumukhi script	Number of Specimen	Total Samples
100 female writers	35	10	35,000 characters
100 male writers	35	10	35,000 characters
Total Sample Collection			70,000 Gurumukhi characters

A novel scenario called hybridization of classification techniques, namely, K-NN, random forest, SVM and decision tree have been implemented using the majority voting scheme, on the feature values generated by curve fitting and intersection & open-endpoint based feature extraction methods.

6.2 CURVE FITTING AND INTERSECTION & OPEN-END POINT FEATURES

6.2.1 Curve Fitting Features

Curve fitting is the process of constructing a mathematical function and structuring a curve that reveals the best fit among a number of foreground pixels of a character image. A fitted curve can be used as a tool for data visualization. It is based on either interpolation or smoothing i.e., exact fit or generating smooth function. The curve fitting approach is suited to the fitting of the best curve, finding relationships among the variables and infers values for a function with no data availability. Extrapolation means fitting of curve outside the scope of experiential data and is a matter of uncertainty.

A parabola is a plane curve with an angular shape that represents the path of something throwing forward and high in the air and falls back to the ground representing the paths of projectiles, as shown in Figure 6.1(a). It is also mirrored symmetrical and is U-shaped. It consists of focus also called a point and a line called the directrix. If a> 0, then the equation of the parabola is $f(x) = a + bx + cx^2$. Then the parabola opens upward, and if a<0, then the parabola opens downward. The shape is similar to the shape of bridges and arches. These shapes shown in Figure 6.1(a) can be used to represent the best fit as shown in Figure 6.1(b) to extract meaningful information for the offline handwriting-based applications.

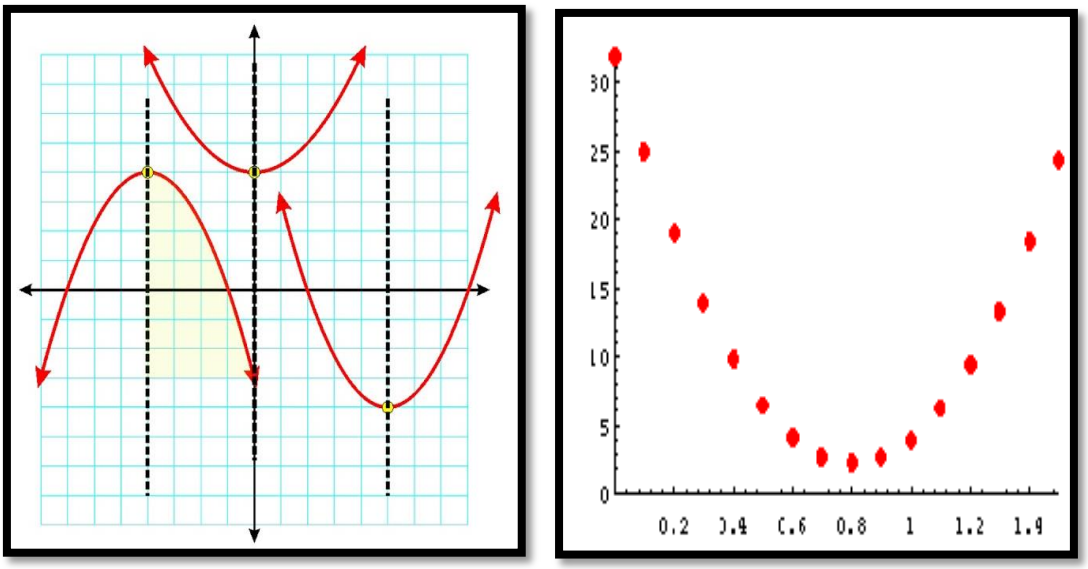

Figure 6.1(a) Fitting of a Parabola Curve (b) Best fit curve

For the implementation of the proposed experiment, the thinned image of a character is divided into zones and then the parabola is formed fitted to the series of ON pixels called foreground pixels in each zone using the Least Square Method (LSM).

A parabola $y = a + bx + cx^2$ is uniquely defined by three parameters: *a*, *b* and *c*. Values of *a*, *b* and *c* are calculated by solving the following equations obtained from LSM.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad (1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad (2)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad (3)$$

Following are the successive steps for evaluating features using Curve fitting based method:

Step I: Divide the thinned image into a number of equal-sized zones.

Step II: For each zone, a parabola is created using the least square method with the evaluation of the values of *a*, *b* and *c*.

Step III: Corresponding to the zone that does not have a foreground pixel, set the values of *a*, *b*, and *c* as zero.

Step IV: Normalize the features in the range of [0, 1] by using the formula,

$$\text{Normalized feature } NV_i = \frac{(\text{Actual feature } V_i - \text{min of actual feature vector})}{(\text{max of actual feature vector} - \text{min of actual feature vector})}$$

Table 6.2 and Table 6.3 shows the results of gender classification and writer identification accuracy with curve fitting based features and linear SVM, K-NN, Decision tree and Random forest. Figure 6.2 shows the curve fitting features.

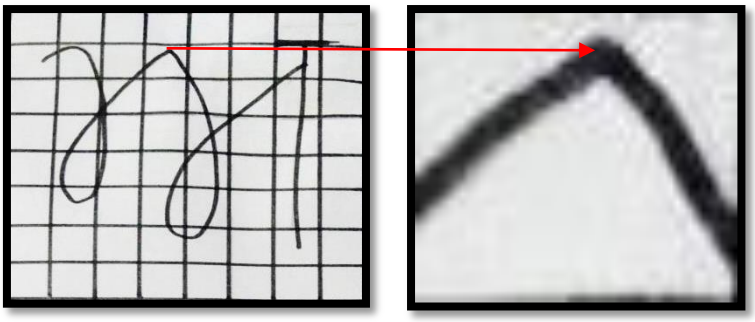

Figure 6.2. Curve Fitting Features

Table 6.2. Gender Classification Accuracy with curve fitting feature

Classification Techniques	Curve fitting based features		
	Accuracy (%)	TPR (%)	FPR (%)
Linear SVM (C1)	**89.28**	88.27	0.36
K-NN (C2)	89.2	89.11	0.26
Decision Tree (C3)	89.28	88.21	0.52
Random Forest (C4)	86.21	85.42	0.26

Table 6.3. Writer Identification Accuracy with curve fitting feature

Classification Techniques	**Curve Fitting Based Features**		
	Accuracy (%)	**TPR (%)**	**FPR (%)**
Linear SVM (C1)	**86.51**	85.53	0.41
K-NN (C2)	86.43	86.35	0.30
Decision Tree (C3)	86.51	85.47	0.60
Random Forest (C4)	83.54	82.77	0.30

For the gender classification, by using the curve fitting based method, gender classification accuracy of 89.28% has been achieved and for writer identification, an accuracy of 86.51% has been achieved with linear SVM classification.

6.2.2 Intersection and Open-End Point Features

The intersection point is the one in which pixel is having one pixel in the neighbourhood and the open endpoint is the pixel in which there is more than one pixel in the neighbourhood, Kumar *et al.,* 2014. So, after the zoning process, we have calculated pixels based on the intersection and open endpoint features, Dargan and Kumar (2021) as shown in Figure 6.3.

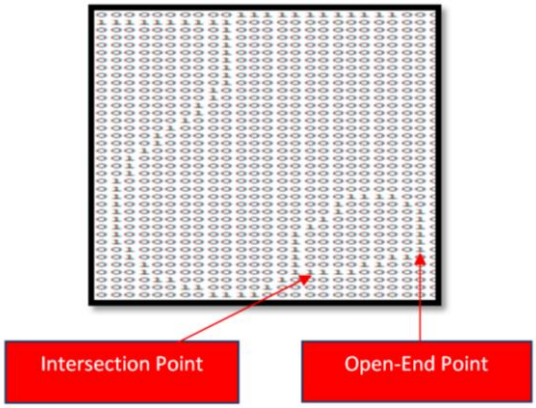

Figure 6.3. Intersection and Open-End Point features

The following steps have been implemented for extracting these features.

Step I: Divide the thinned image of a character into *n* zones.

Step II: Finding the number of intersections and open-end points for each zone.

Step IV: Results are shown in Table 6.4 and Table 6.5.

Table 6.4. Gender Classification Accuracy with Intersection and Open-End Point features

Classification Techniques	Intersection and Open-End Point Features		
	Accuracy (%)	TPR (%)	FPR (%)
Linear SVM (C1)	**88.54**	86.45	0.37
K-NN (C2)	88.50	86.41	0.28
Decision Tree (C3)	87.94	85.49	0.56
Random Forest (C4)	84.36	82.42	0.28

Table 6.5 Writer Identification Accuracy with Intersection and Open-End Point features

Classification Techniques	Intersection Open End Point Features		
	Accuracy (%)	TPR (%)	FPR (%)
Linear SVM (C1)	**81.83**	79.90	0.41
K-NN (C2)	81.79	79.86	0.31
Decision Tree (C3)	81.28	79.01	0.61
Random Forest (C4)	77.97	76.17	0.31

6.2.3 Experimental Results and Discussion

In this subsection, based on the above experimental implementations, results of gender classification and writer identification systems have been analyzed. With curve fitting-based features, maximum gender classification accuracy of 89.28% has been achieved and maximum writer identification accuracy of 86.51% has been achieved with the linear SVM classification technique. Similarly, by implementing Intersection and Open endpoint-based features, the maximum gender classification accuracy of 88.54% and 81.83% for writer identification have been retrieved using linear SVM. In section 6.3, the impact of hybridization of classification techniques with these features on the gender classification and writer identification accuracy rate has thoroughly been discussed by using a majority voting scheme.

6.3 MAJORITY VOTING SCHEME

The majority stands for more than half the votes. It can be taken as plurality which means a subset whose proportion is larger than half of the votes cast. It is an iterative process to ensemble the classification process. It is a dual decision rule used in decision-making and selects alternatives that have a majority. The majority of votes with the classifiers always win (Abbas *et al.*, 2021; Kaur and Kumar, 2021; Nasir and Siddiqi, 2020; Cilia *et al.*, 2020). Toman *et al.* (2011) presented majority voting-based hybrid classification in recognition of offline bangle numerals with a 97.16% accuracy rate. Hajdu *et al.* (2013) worked on generalizing and creating ensembles of many individual classifiers to enhance the accuracy rate and by considering probability terms in constraining the framework and thus shifting from majority

voting to generalize voting. Arora *et al.* (2010) presented a performance comparison of ANN and SVM based on a majority voting scheme and achieved high rates. Nasir and Siddiqi (2021) developed a model for writer identification on Devanagari script, based on CNN, and decision on various samples was combined using a majority voting scheme. Figure 6.4. shows the working of the majority voting scheme.

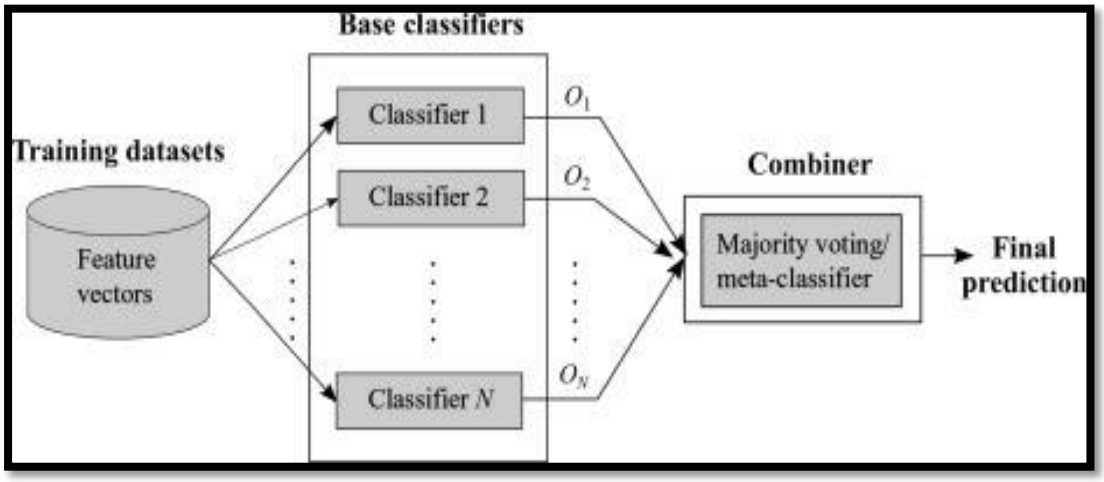

Figure 6.4. Majority Voting Scheme

Bhattacharya and Chaudhuri (2003) presented the implementation of a majority voting scheme in the recognition of handwritten numerals. Kumar *et al.* (2012) proposed majority voting-based recognition of handwritten numerals on four neural network classifiers and achieved success rates, (Rahman *et al.*, 2002).

6.4 HYBRIDIZATION OF CLASSIFICATION TECHNIQUES

Due to an increase in size and complexity of the data, hybridization of parametric and non-parametric classification techniques is really a boon to perform the robust, agile, and optimize classification. The huge growth in the data poses a need for the development of an efficient classification technique that can perform competent data analysis. Hybridization of classification techniques facilitates taking optimal decisions with less computational time and less computational cost. Hybrid classifiers are less susceptible to the choice of parametric models. So, hybridization will overcome the limitations and weakness of the classifiers and help in boosting the outputs i.e., the strength of classifiers.

Hybrid classification is a technique to achieve more robust and accurate classification by measuring consensus among the participating classifiers. Many

researchers have been successfully developing applications based on the hybridization approach for enhancing, improving, and optimizing classification results, (Delimata and Suraj, 2013). Rahman *et al.* (2002) presented state-of-the-art work on a majority voting scheme and its variations. Chaudhuri *et al.* (2009) proposed an ensemble of classification methods based on parametric and non-parametric classification and focused on the idea of boosting the accuracy with hybridization. Kumar *et al.* (2018) worked on the recognition of medieval handwritten Gurumukhi characters in the manuscripts using hybridization of K-NN, SVM, Random Forest, and Decision Tree classifiers with a majority voting scheme and achieved 95.91% accuracy. Sherwani *et al.* (2020) presented a detailed analysis of the comprehensive review on hybridization of classification techniques using recurrent neural networks (RNNs) algorithm, Backpropagation neural network (BPNN), and Levenberg Marquardt (LM) algorithms. Wong *et al.* (2020) presented a hybrid classification of naive Bayesian classifier and decision tree induction based on instance filtering and achieved successful results. Ghosh *et al.* (2020b) presented an amalgamation of SVM, k-NN, and principal component analysis to recognize the handwritten digit.

The majority can also be referred to as the winning margin, and help to create and maintain stable majority control. The majority means more than half out of the total and voting means to elect the best. This method is used to determine a single winner in a contest. Hybridization of classification algorithms can be done in two modes, hard and soft voting schemes. Majority voting is also called hard voting that has been implemented in this proposed experiment in which every classifier vote for a class and the predicted target result of the classification will be the mode of the distribution of individually predicted labels.

Kumar *et al.* (2012) presented handwritten numeral identification based on Daubechies wavelet transform, multilayer feed-forward, k-NN, linear discriminant analysis, and majority voting method and attained successful results. The approach is called hybrid because, in this process, the basic algorithms are involved in Pre-processing and induction process.

Advantages of hybridization of classification techniques:

- An instance misclassified by one technique may be identified by other and hence boosting the accuracy rate.

- In case of uncertainty with the model strength or assumptions, it is always safe and secure to use a fusion of classifiers, which provides a safeguard against potential deviations in concern with parametric model assumptions.

- When the factual population distributions are extreme from the unspecified parametric models, hybrid classifiers can substantially advance the performance of parametric methods and capitulate misclassification rates, which are analogous to those of the nonparametric classifiers.

- With hybridization, misclassified instances are very few.

- The method is widely helpful for the applications such as handwriting identification, cancer diagnosis, terrorism prediction, classifying engineering students, and selection of students for business school, etc.

6.5 SYSTEM PERFORMANCE BASED ON CURVE FITTING FEATURES AND HYBRIDIZATION OF CLASSIFICATION TECHNIQUES

6.5.1 Experimental Results for Gender Classification System

The development of a framework for the gender classification system with 200 writers having 100 female writers and 100 male writers has been presented by using curve-fitting features. In the proposed experimental work, four classification techniques, namely K-NN (C1), Random Forest (C2), SVM (C3), Decision Tree (C4) have been experienced using majority voting scheme and results of gender classification and writer identification are shown in Table 4.2 and 4.3 as shown in Table 6.6. Experiment results revealed that after implementing hybridization on the classification techniques using majority voting, the maximum accuracy of 90.57% has been achieved, which was 89.28% before hybridization.

Table 6.6. Gender Classification Accuracy with hybridization of classification techniques and Curve Fitting Based Features

Hybridization of Classification Techniques	Curve Fitting Based Features		
	Accuracy (%)	TPR (%)	FPR (%)
C1+ C2	89.50	88.41	0.44
C1+ C3	90.13	89.05	0.36
C1+ C4	89.08	88.00	0.44
C2+ C3	90.09	89.01	0.36
C2+ C4	88.87	87.8	0.62
C3+ C4	89.78	88.71	0.26
C1+ C2+ C3	90.23	89.38	0.17
C1+ C2+ C4	89.63	88.43	0.17
C1+ C3+ C4	90.27	89.19	0.26
C2+ C3+ C4	90.43	89.34	0.17
C1+ C2+ C3+ C4	90.57	89.47	0.36

6.5.2 Experimental Results for Writer Identification System

The accuracy achieved for writer identification system based on the handwritten samples of 200 writers using hybridization of classification techniques along with the features extracted from curve-fitting based method which was 86.51% before hybridization. Thus, hybridization has enhanced the accuracy for both gender and writer identification system is shown in Table 6.7.

Table 6.7. Writer Identification Accuracy with hybridization of classification techniques and Curve Fitting Features

Hybridization of Classification Techniques	Curve Fitting Based Features (F2)		
	Accuracy (%)	TPR (%)	FPR (%)
C1+ C2	86.72	85.67	0.51
C1+ C3	87.34	86.29	0.41
C1+ C4	86.32	85.27	0.51
C2+ C3	87.30	86.25	0.41
C2+ C4	86.11	85.08	0.71
C3+ C4	87.00	85.96	0.30
C1+ C2+ C3	87.43	86.61	0.20
C1+ C2+ C4	86.85	85.69	0.20
C1+ C3+ C4	87.47	86.42	0.30
C2+ C3+ C4	87.63	86.57	0.20
C1+ C2+ C3+ C4	**87.76**	86.70	0.41

The accuracy achieved for writer identification system based on the handwritten samples of 200 writers using hybridization of classification techniques along with the features extracted from curve-fitting based method which was 86.51% before hybridization. Thus hybridization has enhanced the accuracy for both gender and writer identification system is shown in Table 6.7

6.6 SYSTEM PERFORMANCE BASED ON INTERSECTION & OPEN-END POINT FEATURES AND HYBRIDIZATION OF CLASSIFICATION TECHNIQUES

6.6.1 Experimental Results for Gender Classification System

In this subsection, the gender classification accuracy has been evaluated with intersection and open end point features along with the hybridization of the classifiers. The results achieved are shown in Table 6.8 in which maximum accuracy of 88.88% has been achieved.

Table 6.8. Gender Classification Accuracy with hybridization of Classification Techniques and Intersection and Open-End Point based features

Hybridization of Classification Techniques	Intersection and Open-End Point Based Features		
	Accuracy (%)	TPR (%)	FPR (%)
C1+ C2	88.55	86.55	0.46
C1+ C3	88.84	85.49	0.37
C1+ C4	87.31	85.81	0.46
C2+ C3	88.84	86.78	0.37
C2+ C4	87.22	85.07	0.66
C3+ C4	88.06	85.98	0.28
C1+ C2+ C3	88.73	86.64	0.19
C1+ C2+ C4	87.98	86.31	0.19
C1+ C3+ C4	88.75	86.47	0.28
C2+ C3+ C4	**88.88**	86.78	0.19
C1+ C2+ C3+ C4	88.63	86.54	0.37

Here for gender classification system, maximum accuracy of 88.88% has been achieved using intersection and open-endpoint-based features along with hybridization of C2, C3 and C4 which was previously 88.54% without using hybridization of classification techniques.

6.6.2 Experimental Results for Writer Identification System

Similarly, the accuracy rate achieved for identifying the writer has been presented in this section. Based on the hybridization of classification techniques i.e., C2, C3, and C4, and using Intersection and Open endpoint features, writer identification accuracy of 82.14% has been achieved as presented in Table 6.9.

Table 6.9. Writer Identification Accuracy with hybridization of classification techniques and Intersection and Open end Point based features

Hybridization of Classification Techniques	Intersection and Open-End Point Based Features		
	Accuracy (%)	TPR (%)	FPR (%)
C1+ C2	81.84	79.99	0.51
C1+ C3	82.11	79.01	0.41
C1+ C4	80.69	79.31	0.51
C2+ C3	82.11	80.20	0.41
C2+ C4	80.61	78.62	0.72
C3+ C4	81.39	79.46	0.31
C1+ C2+ C3	82.01	80.07	0.21
C1+ C2+ C4	81.31	79.77	0.21
C1+ C3+ C4	82.02	79.92	0.31
C2+ C3+ C4	**82.14**	80.20	0.21
C1+ C2+ C3+ C4	81.91	79.98	0.41

6.7 COMPARATIVE ANALYSIS

Maximum classification accuracy of 90.57% has been achieved with curve fitting-based features and hybridization of classifiers C1+C2+C3+C4 using the majority voting scheme. A true positive rate of 89.47% and a false-positive rate of 0.36% have been realized. Earlier the gender classification accuracy was reported as 89.28% with linear SVM and curve-fitting based features. So, hybridization shows an improvement in gender classification accuracy rates. Complete results for gender classification with curve fitting and intersection and open-endpoint-based features and hybridization of classification techniques are as shown in Table 6.10.

Table 6.10. Gender Classification Accuracy with hybridization of Feature Extraction Techniques and Curve Fitting Features and Intersection Open-End Point Features

Hybridization of Classification Techniques	Intersection and Open-End Point Based Features			Curve Fitting Based Features		
	Accuracy (%)	TPR (%)	FPR (%)	Accuracy (%)	TPR (%)	FPR (%)
C1+ C2	88.55	88.41	0.46	89.50	88.41	0.44
C1+ C3	88.84	89.05	0.37	90.13	89.05	0.36
C1+ C4	87.31	88.00	0.46	89.08	88.00	0.44
C2+ C3	88.84	89.01	0.37	90.09	89.01	0.36
C2+ C4	87.22	87.8	0.66	88.87	87.8	0.62
C3+ C4	88.06	85.98	0.28	89.78	88.71	0.26
C1+ C2+ C3	88.73	86.64	0.19	90.23	89.38	0.17
C1+ C2+ C4	87.98	86.31	0.19	89.63	88.43	0.17
C1+ C3+ C4	88.75	86.47	0.28	90.27	89.19	0.26
C2+ C3+ C4	88.88	86.78	0.19	90.43	89.34	0.17
C1+ C2+ C3+ C4	88.63	86.54	0.37	**90.57**	89.47	0.36

For the writer identification, results achieved are presented in Table 6.11 with both curve fitting and intersection and open-endpoint-based features and hybridization of classification techniques.

Table 6.11. Writer Identification Accuracy with hybridization of Feature Extraction Techniques and Curve Fitting Features and Intersection Open-End Point Features.

Hybridization of Classification Techniques	Intersection and Open-End Point Based Features			Curve Fitting Based Features		
	Accuracy (%)	TPR (%)	FPR (%)	Accuracy (%)	TPR (%)	FPR (%)
C_1+ C_2	81.84	79.99	0.51	86.72	85.67	0.51
C_1+ C_3	82.11	79.01	0.41	87.34	86.29	0.41
C_1+ C_4	80.69	79.31	0.51	86.32	85.27	0.51
C_2+ C_3	82.11	80.2	0.41	87.30	86.25	0.41
C_2+ C_4	80.61	78.62	0.72	86.11	85.08	0.71
C_3+ C_4	81.39	79.46	0.31	87.00	85.96	0.30
C_1+ C_2+ C_3	82.01	80.07	0.21	87.43	86.61	0.20
C_1+ C_2+ C_4	81.31	79.77	0.21	86.85	85.69	0.20
C_1+ C_3+ C_4	82.02	79.92	0.31	87.47	86.42	0.30
C_2+ C_3+ C_4	**82.14**	80.20	0.21	87.63	86.57	0.20
C_1+ C_2+ C_3+ C_4	81.91	79.98	0.41	**87.76**	86.70	0.41

Maximum writer identification accuracy of **87.76%** has been attained with curve fitting based features and hybridization of classifiers C1+C2+C3+C4 with true

positive rate 86.70% and false positive rate 0.41%, which was earlier 88.54% with linear SVM using intersection and open-endpoint-based features.

6.8 DISCUSSION AND CONCLUSION

This chapter presents the development of gender classification and writer identification system using curve fitting and intersection & open-end based feature extraction methods followed by the hybridization of classification techniques. Firstly, implementation of curve fitting-based features has been practiced with the classification techniques individually, namely linear-SVM, k-NN, Decision Tree, and Random Forest, then using hybridization of classification using majority voting scheme has been experienced. It has been analyzed that for 200 writers i.e., 70,000 Gurumukhi characters with 35000 female and 35000 male characters when implementing hybridization of all classification methods, C1+C2+C3+C4 and curve-fitting based feature, the maximum gender classification accuracy of 90.57% has successfully been achieved as compared to 89.28% which was before hybridization. With C1+C2+C3+C4 and curve-fitting-based feature, maximum writer Identification accuracy of 87.6% has been realized, and before hybridization, it was 86.51%. Table 6.12 shows the results achieved without hybridization. So improved accuracy rates have been experienced which are promising and satisfactory as depicted in Table 6.13 with this novel exploration of approaches.

Table 6.12. Summarized view of results achieved for gender classification and writer identification system without hybridization of classification techniques

<table>
<tr><th>Nature of System</th><th>Feature Extraction Techniques</th><th>Classification Techniques</th><th>Number of writers</th><th>Results</th></tr>
<tr><td>Gender Classification System</td><td rowspan="2">Curve Fitting Based Method</td><td rowspan="4">Without Hybridization of classification techniques, namely Linear SVM, K-NN, Decision Tree, Random Forest uses the majority voting scheme</td><td rowspan="4">Total 200 writers, 70000 Gurumukhi characters, 35000 male samples and 35000 female samples</td><td>89.28%</td></tr>
<tr><td>Writer Identification System</td><td>86.51%</td></tr>
<tr><td>Gender Classification System</td><td rowspan="2">Intersection and Open-end point method</td><td>88.54%</td></tr>
<tr><td>Writer Identification System</td><td>81.83%</td></tr>
</table>

Table 6.13. Summarized view of results achieved for gender classification and writer identification system with hybridization of classification techniques

Nature of System	Feature Extraction Techniques	Classification Techniques	Number of writers	Results
Gender Classification System	Curve Fitting Based Method	With Hybridization of classification techniques, namely Linear SVM, K-NN, Decision Tree, Random Forest uses the majority voting scheme	Total 200 writers, 70000 Gurumukhi characters, 35000 male samples and 35000 female samples	90.57%
Writer Identification System				87.76%
Gender Classification System	Intersection and Open-end point method			88.88%
Writer Identification System				82.14%