

# Historical Kannada Document Classification using Symbolic Representation

**Manjunath B**

Dept. of MCA, Maharaja Institute of Technology Mysore  
Mandya, India

**Sharathkumar Y H**

Dept. of ISE, Maharaja Institute of Technology Mysore  
Mandya, India

## **Abstract**










*In India, the Kannada is one of the historical and formal languages in Karnataka. The historical Kannada documents gives us information about education, legislation, culture and traditions that have been practiced. Getting such information from stone carvings and palm leaves and other sources are improves our knowledge of Kannada language. Extracting information from historical records is very challenging because of poor quality, variability, contrast and envelope of characters. In this work, the given document is pre-processed using connected component analysis. The features like HOG and SIFT are extracted. The features are stored in knowledgebase by aggregating and representing as inter-value type data. Symbolic classifier is introduced for the classification purpose. Experiment was conducted on our database to verify the performance of the presented method.*

**Keywords:** Historical Kannada, Symbolic classifier, Ancient scripts, Inscription.

## **1. Introduction**

Kannada is one of the ancient languages of India. Kannada has its own style in scripting and originated before 230BC in the form of ancient scripts, such as epigraphs or inscriptions. Ancient scripts are the primary resources to enhance our knowledge about ancient civilization, which includes traditions, education, legislation, medicine and so on. Ancient scripts are totally different than the current scripts, it is shown in table 1 and To read and identify ancient scripts is not straightforward task because of scripting style but we can identify ancient scripts by epigraphers. This manual recognition method is a time-consuming and tedious task. It is a good idea to develop OCR to automatically identify Kannada ancient scripts to alleviate the difficulties of the manual recognition method. The OCR (Optical Character Recognition) is an essential part of a manuscript image processing method. The LBP, Gabor filter and GLTP features with symbolic classifier to recognize ancient Kannada characters. This automated epigraphs recognition system, it reads the epigraphs, extract the significant features, based on the extracted features performs the classification, and finally recognition the epigraphs. The epigraphs have set of ancient characters. In this presented work, our efforts has been made to recognize the Kannada characters as shown in table 1 through classification and recognition techniques. The paper is structured as follows: in section 2, the related work is presented, in section 3, gives the details proposed recognition system, in section 4, about Symbolic representation, in section 5, illustrates the results and discussion and finally section 6 conclusion.

Table1: Evolution of Character 'Aha' from 3rd B.C to Present Kannada.

| Sample Scripting Style  | Period (Century)                    | Rulers Name          |
|---|-------------------------------------|----------------------|
|  | 3 <sup>rd</sup> BC                  | Ashoka               |
|  | 2 <sup>nd</sup> AD                  | Shaatavaahana        |
|  | 4 <sup>th</sup> -5 <sup>th</sup> AD | Kadamba              |
|  | 6 <sup>th</sup> AD                  | Baadami<br>Chaalukya |
|  | 9 <sup>th</sup> AD                  | Raashtrakuta         |
|  | 10 <sup>th</sup> AD                 | Kalyana<br>Chaalukya |
|  | 12 <sup>th</sup> AD                 | Hoysala              |
|  | 15 <sup>th</sup> AD                 | Vijayanagara         |
|  | 18 <sup>th</sup> AD                 | Mysore               |
|  | After 18 <sup>th</sup><br>AD        | Present Kannada      |

## 2. Related Work

The authors proposed symbolic representation [1] technique to extract symbolic features from 2D shapes images and numerous experiments conducted on good number of dataset with good results. Ehtesham et al.[2] presented symbol classifier to recognize character of Gujarati language by using multiple kernel learning. Here 3 different feature depictions applied for symbol images of Gujarati script and got good results. D S Guru et al.[3] proposed a method symbolic classifier for a text document, experiment was performed on a vehicle Wikipedia database for capturing the features and revealing the results obtained with the existing results, it takes relatively less time for text classification. In [4] reported symbolic classifiers for classify text documents by using Symbolic clustering approaches for different measures. The authors [5] presented MKFC-Means method for symbolic features of text documents, proposed method frequency vector, mean and standard deviation were considered for clustering the standard dataset. Writer dependent features discussed [6] in online signature verification application by different filter based features collection techniques. Experiment was conducted on standard database MCYT and discussed about importance of symbolic representation, feature vector and relevancy in signature verification problems. A symbolic classifier [7] has used to classification of unlabeled text documents, this work translate the imbalance classes into multiple smaller subclasses by class-wise clustering. After that each subclass denoted as feature vector form and stored in a knowledgebase, the results of proposed technique is better than the other existing methods. To solve the problem of image classification and authentication in document by using feature values from the image documents by class 1 classification built on the symbolic representation is proposed [8]. The authors [9] presented work on recognition method for dissimilarities in font

style and size of Kannada character set by using symbolic representation. The Hybrid approach has been reported in [10] for recognition of Amazigh character, in this work the Hough transformation used to extract the directional primitives from pre-processing character image and then trained the Hidden Markov Models by using directional primitives for recognizing the Amazigh character. In [11], the authors presented a system to classify Greek Inscriptions with different classification techniques. Authors [12] introduced OCR system, which has pre-processing, segmentation, feature extraction and classification for different font sizes of Devnagari characters using different methods for each stage and the recognition rate found to be quite high. Identifying the ancient inscription using methods of image processing, pattern recognition has been reported [13]. The authors proposed [14] a recurrent CNN for text classification and applied a recurring structure to capture related dataset. Using a max-pooling layer that automatically determines which words play an important role in text classification to capture key points in the text. A text classification from essay dataset by using CNN and RNN approach has been reported [15]. The authors conclusion was RNN done better than CNN for essay dataset. Garz et al. [16] presented, identifying text areas and decorative features in ancient scripts and robust method was encouraged by objects recognition method. SIFT descriptors were chosen to identify interest regions, which is used for localization. The authors [17] used local features for effective layout analysis of olden scripts. Here identifies and localizes the layout units in scripts. Hence, the textual units were disassembled into sections and part based finding has done, which employs local gradient features from object recognition fields, SIFT to define these structures. The authors proposed [18] a system to detect and recognize the text regions from scanned images by maximally stable external sections and a trained Convolutional Neural Networks, pre-processes the image, after that find out MSERs, and resulted values saved individually. Then classifier is trained by using resulted values for analyses the characters successfully with error rate less.

### **3. Proposed Method**

The planned system, shown in Figure 1, portrayed in Four stages. Firstly, the pre-processing stage, attempts to Division the images from the background, In stage two the images will be divided into 4,16,32 blocks. For each block we extracted features like Scale Invariant Feature Transform (SIFT) and Histogram of Gradient (HOG). The exacted features are fed into Symbolic classifier.

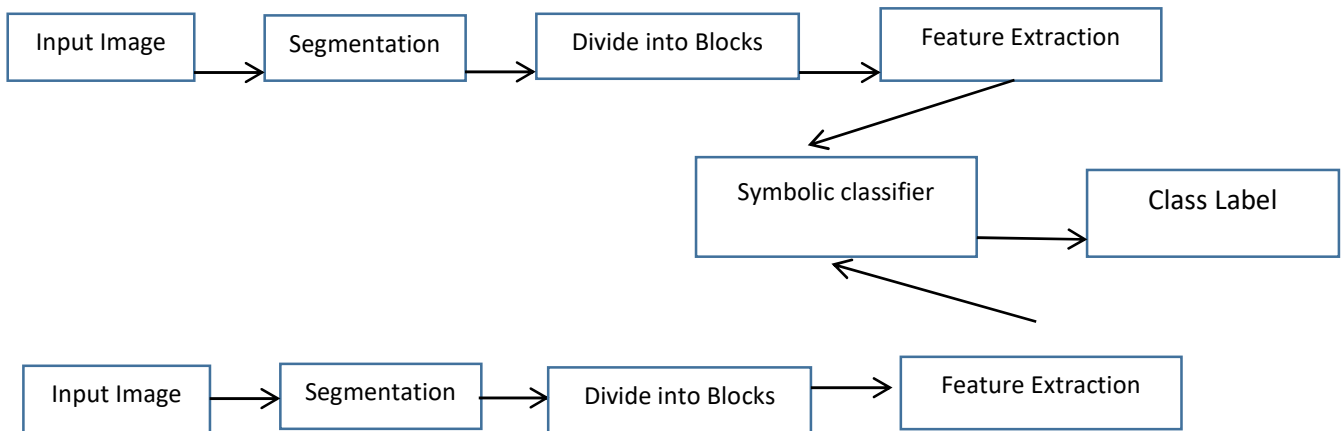


Figure 1: shows the block diagram of the proposed method

### 3.1 Pre-processing

In the pre-processing phase, we apply connected component mechanics to identify the elements in the image manuscript. After that, the bounding box is drawn and filled with color for each of the identified components. Then apply horizontal projection so that we can more accurately identify each text line using these methods.

The steps are in pre-processing phase

- Using Sauvolas approach to pre-process the historical image.
- To use connected component mechanics to identify the elements in the binary image.
- To draw bounding box for each recognized component and fill with color.
- A horizontal projection is applied on it and identify each text line, which is shown in figure 2

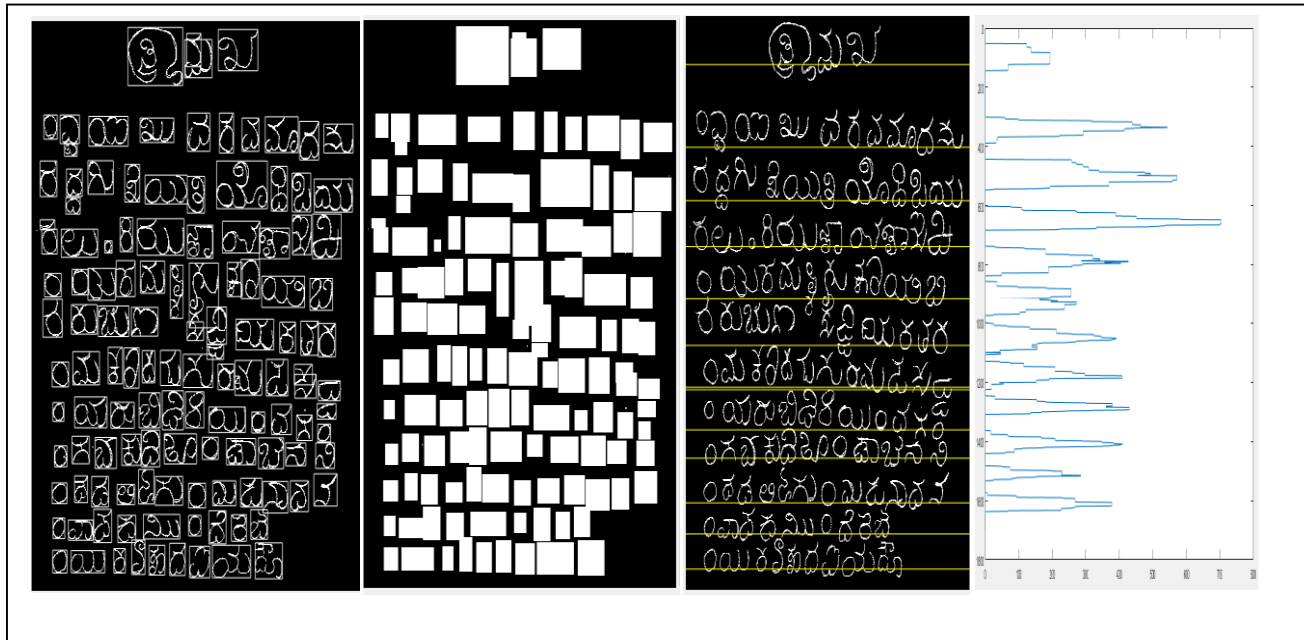


Figure. 2. Identify text line in the image

### 3.2 Feature Extraction

For a segmented image we extract scale invariant feature transforms and Histogram of Gradients. The following subsection gives an introduction to all the above features.

#### 3.2.1 Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) [6] algorithm has four important phases, Extrema detection: The stable points are identified in this phase and these points can be done by searching for fixed features across all the possible scales. Key point Detection: The algorithm first finds the  $\alpha N^2$  extrema then further refines them to  $\alpha N^2$  key points, which is key point of the image. To detect the extrema on edges, a  $2 \times 2$  matrix is generated and some simple calculations are performed on it to generate a ratio of principle of curvatures and this quantity is compared with a threshold value to decide whether an extrema is to be rejected or not. Orientation Assignment: The magnitude  $m_i^j(x, y)$  and orientation  $\theta_i^j(x, y)$  for each point  $L_i^j(x, y)$  can be computed by:

$$m_i^j(x, y) = \sqrt{(L_i^j(x + 1, y) - L_i^j(x - 1, y))^2 + (L_i^j(x, y + 1) - L_i^j(x, y - 1))^2} \quad (1)$$

$$\theta_i^j(x, y) = \tan^{-1} \frac{(L_i^j(x, y + 1) - L_i^j(x, y - 1))}{(L_i^j(x + 1, y) - L_i^j(x - 1, y))} \quad (2)$$

The input is the set of vector in neighbourhood of each keypoint this information is consolidated to form descriptor. It contains information in a  $2x \times 2x$  neighbourhood of keypoints.

### **3.2.2 Histogram of Oriented Gradients (HOG)**

In HOG feature extraction process [7], the feature extraction is done from a local region with  $16 \times 16$  pixels. The 8 orientations Histogram of Gradients are calculated from of  $4 \times 4$  cells. The total number of HOG features become  $128 = 8 \times (4 \times 4)$ . The HOG feature extraction is done from  $16 \times 16$  local regions. The cells are the local regions that are divided into small spatial area.  $4 \times 4$  pixels is the size of each cell. 8 orientations Histogram of gradients are calculated from each local cell. Hence, the total number of HOG features will become  $128 = 8 \times (4 \times 4)$  and it builds a HOG feature vector.

### **3.3 Symbolic Approach.**

The recent developments in the area of symbolic data analysis have proven that the real life objects can be better described by the use of symbolic data, which are extensions of classical crisp data [3]. Symbolic data appear in the form of continuous ratio, discrete absolute interval and multivalued, multivalued with weightage, quantitative, categorical, etc. The concept of symbolic data analysis has been extensively studied in the field of classification and it has been proved both theoretically and experimentally that the classification approaches based on symbolic data outperform conventional classification techniques. In image representation, the features of each class possess significant variations. Therefore, we felt that it would be more meaningful to capture these variations in the form of interval-valued features and to provide an effective representation for classification. In this work while fixing inter valued feature we considered minimum and maximum of feature values.

#### **3.3.1 Symbolic Representation.**

After features extraction are further represented using symbolic data. As scripts possess high intra-class variations in each class, to preserve these intra-class variations is very difficult through conventional data representation. Hence, in this work we have used an unconventional data representation that is symbolic data, in particular interval valued data. Symbolic data has an ability

to preserve the variations among the data effectively. The features are represented using interval valued [3] as follows.

Let  $[X_1, X_2, X_3, \dots, X_n]$  be a set of  $n$  samples of class say  $C_k$  where  $k=1, 2, 3, \dots, N$ . Where,  $N$  denotes the number of classes.  $F_j = [f_{j1}, f_{j2}, f_{j3}, \dots, f_{jd}]$  Representing  $d$  features characterizing the sample  $X_j$  of the class  $C_k$ .

Let  $\mu_{ks}$ ,  $s=1, 2, 3, \dots, d$  be the mean of the  $s$ th feature values obtained from all  $n$  samples of class  $C_k$ . That is,

$$\mu_{ks} = \frac{1}{n} \sum_{i=1}^n f_{js} \quad (3)$$

Similarly, Let  $\sigma_{ks}$ ,  $s=1, 2, 3, \dots, d$  be the standard deviation of the  $s$ th feature values obtained from all  $n$  samples of class  $C_k$ . That is,

$$\sigma_{ks} = \left[ \frac{1}{n} \sum (f_{js} - \mu_{ks})^2 \right]^{1/2} \quad (4)$$

Now, we propose an interval value feature to capture the intra-class variations in each  $s$ th feature of the  $k$ th class  $[f_{ks}^-, f_{ks}^+]$ , where

$$f_{ks}^- = \mu_{ks} - \sigma_{ks} \text{ and } f_{ks}^+ = \mu_{ks} + \sigma_{ks} \quad (5)$$

For all the class  $C_k$  the animals are represented by using the above interval valued data. These interval valued data for each class  $C_k$  is stored in the knowledge base for the further classification purpose. Therefore there are  $N$  number of symbolic features are there in the knowledgebase i.e.,

$$RF_k = \{[f_{k1}^-, f_{k1}^+], [f_{k2}^-, f_{k2}^+], \dots, [f_{km}^-, f_{km}^+]\} \quad (6)$$

### 3.3.2 Symbolic Classifier.

Now the features which are represented in interval valued data is used for classification using a symbolic classifier [3]. During classification, a test sample described using  $d$  feature values will compare with corresponding interval type features stored in the knowledgebase.

Let  $F_j = [f_{j1}, f_{j2}, f_{j3}, \dots, f_{jd}]$  be the features of the test sample. Let  $RF_k$  where  $k=1, 2, 3, \dots, N$  be the representatives symbolic feature vectors in the knowledgebase. During classification a feature value of the test sample is compared with the respective intervals to check whether it lies between

those intervals. The test sample of script is classified to a class based on the maximum acceptance count say  $M_c$ .

Acceptance count  $M_c$  is given by,

$$M_c = \sum_{s=1}^d C(f_{ts}, [f_{ks}^-, f_{ks}^+]) \quad (7)$$

$$\text{Where, } C(f_{ts}, [f_{ks}^-, f_{ks}^+]) = f(x) = \begin{cases} 1, & \text{if } f_{ts} \geq f_{ks}^- \text{ and } f_{ts} \leq f_{ks}^+ \\ 0 & \text{otherwise} \end{cases}$$

Since our dataset is too large, the probability of test sample to get the same acceptance count for more than one class is possible. To overcome this problem we recommend the use of similarity measure, which computes similarity value between a test sample of animal and each of the conflicting classes [4].

#### 4. Results

In order for experimentation, The dataset of Kannada's historical letters are shown in figure 3. So as to substantiate the proficiency of the proposed methodology, we completed broad trials on various Character dataset viz. Ashok, Kadamba, Hoysala and Mysuru Scripts. Each character dataset contains 25 images shown in figure 5. In this section. We aimed to learning the classification accurateness under different features of SIFT and HOG. We picked images arbitrarily from the dataset and experiment is carry-out in a dataset of 4 classes under 70, 50 and 30 different training models from each class. In addition, to exhibit the performances of classifiers and the testing is showed on Min- Max and Mean-Std Deviation representation by changing the training samples.



**Figure. 3 .** The dataset of Kannada's historical letters.





**Figure. 4.** Character dataset of Kannada.

The experimentation is conducted more than five times by dividing the images into 4,16,32 blocks and images picked randomly. The experimentation is conducted on classes by varying training samples from 30 to 70 percent of database. The results obtained for block 4 using HOG Features is shows Figure 5 and using SIFT Features is shows in Figure 6 and Fusion of both the features is shown in Figure 7. The results obtained for block 16 using HOG Features is shows Figure 8 and using SIFT Features is shows in Figure 9 and Fusion of both the features is shown in Figure 10. The results obtained for block 32 using HOG Features is shows Figure 11 and using SIFT Features is shows in Figure 12 and Fusion of both the features is shown in Figure 13. In figure 14 shows accuracy of symbolic classifier. It can be noticed that the symbolic classifier classifiers achieves relatively higher accuracy of 91 percent, When 70 percent of the database system are used for training.

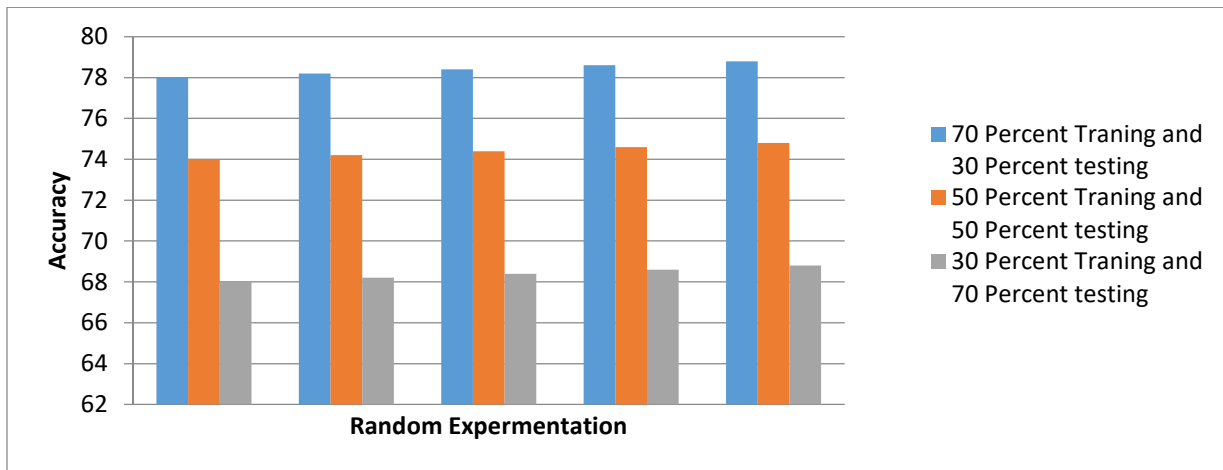
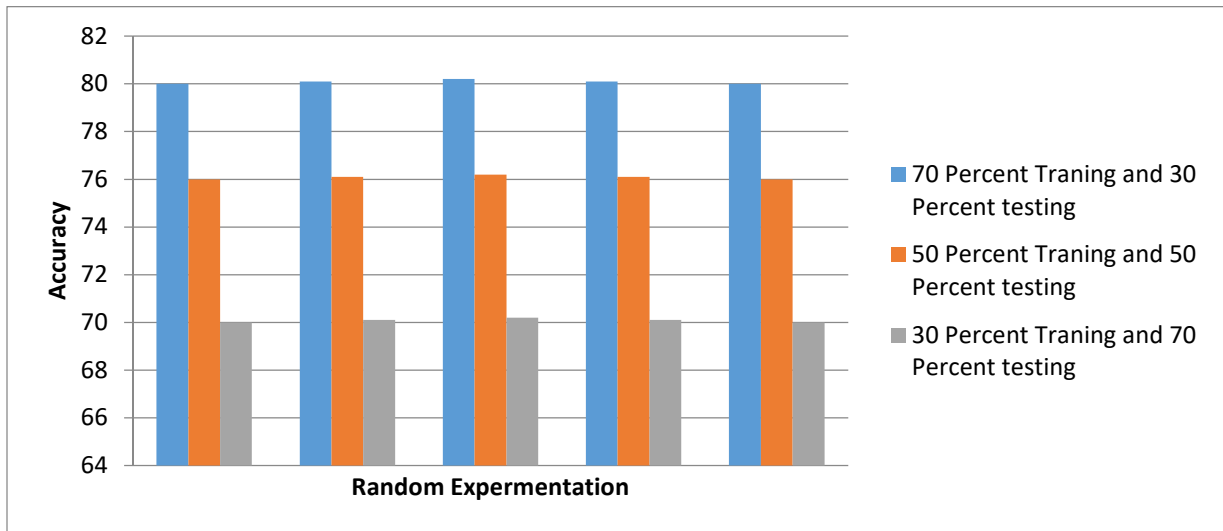
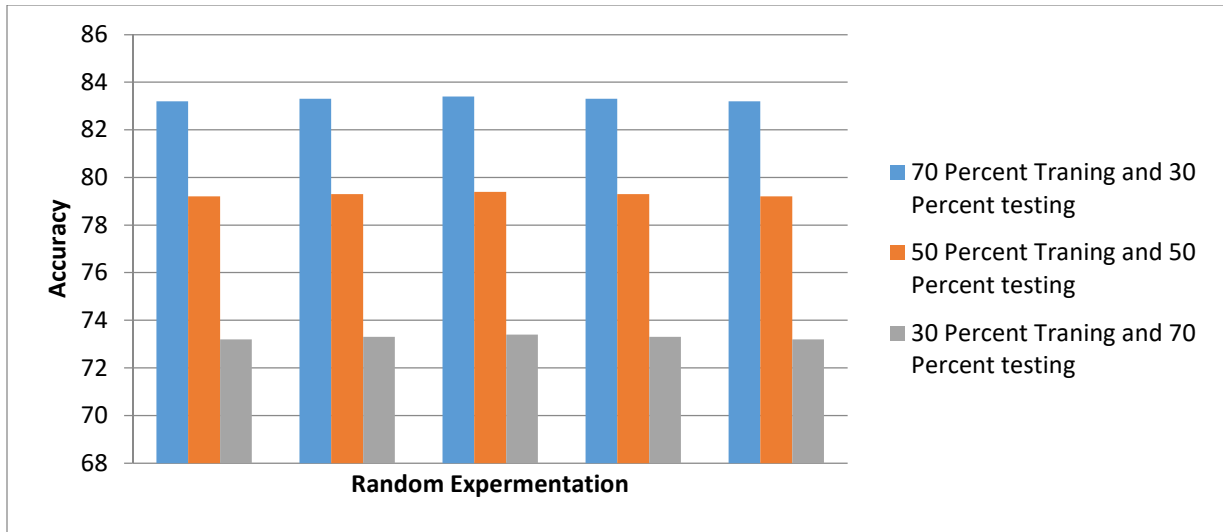
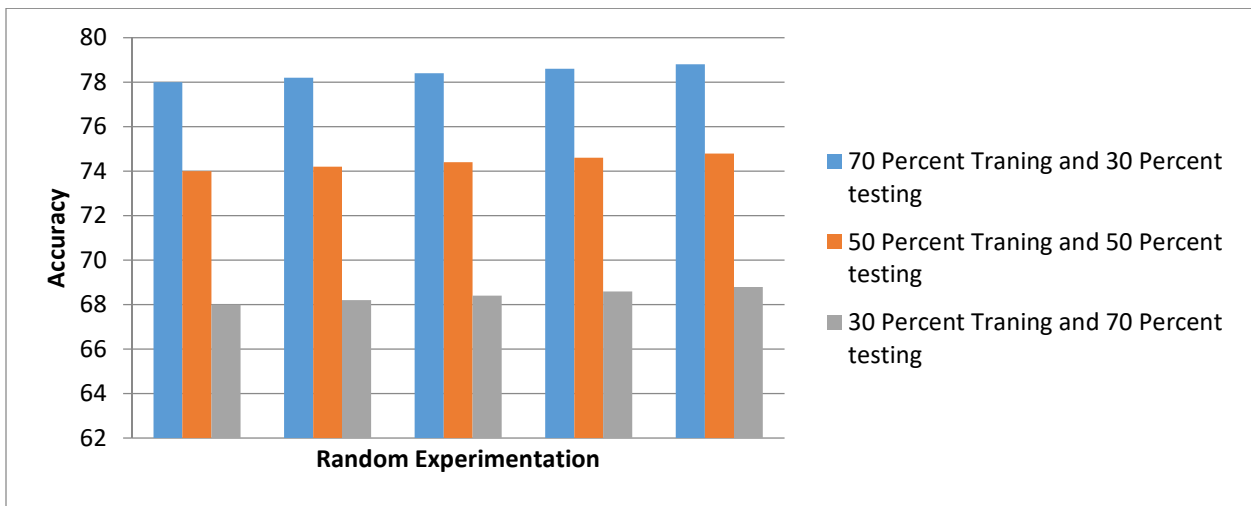
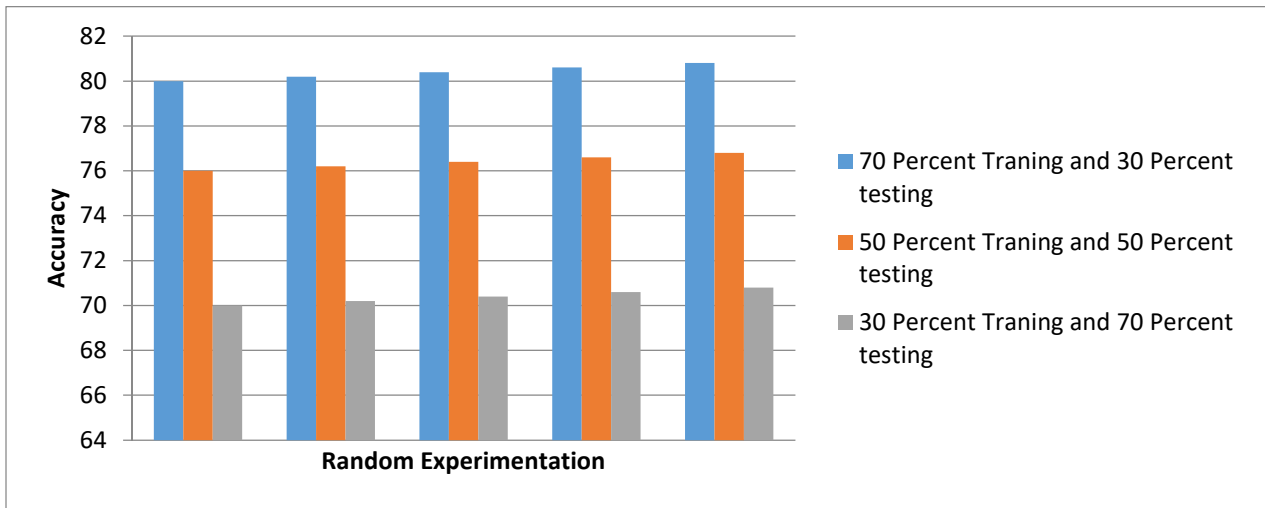
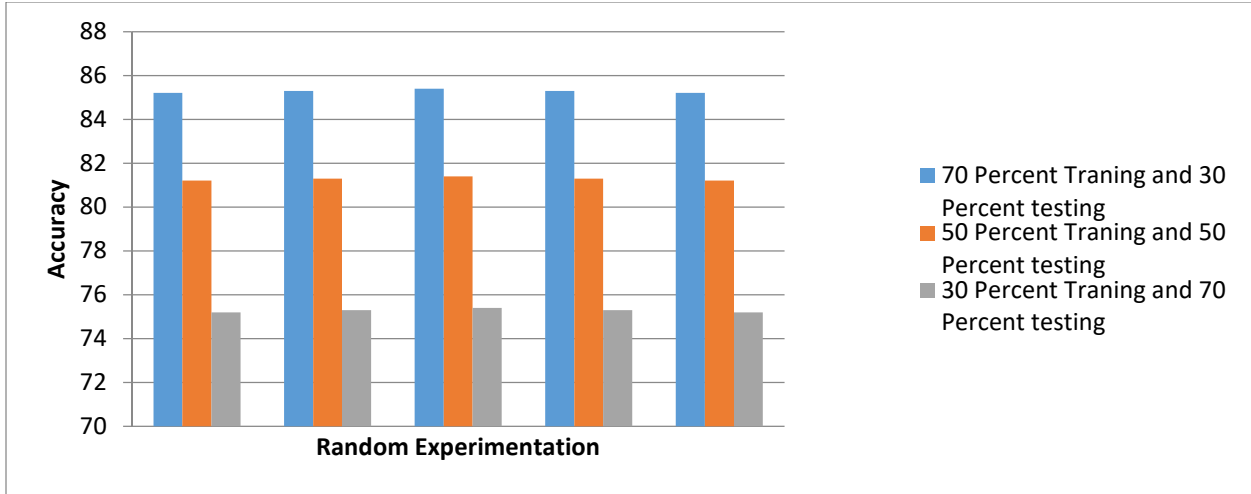
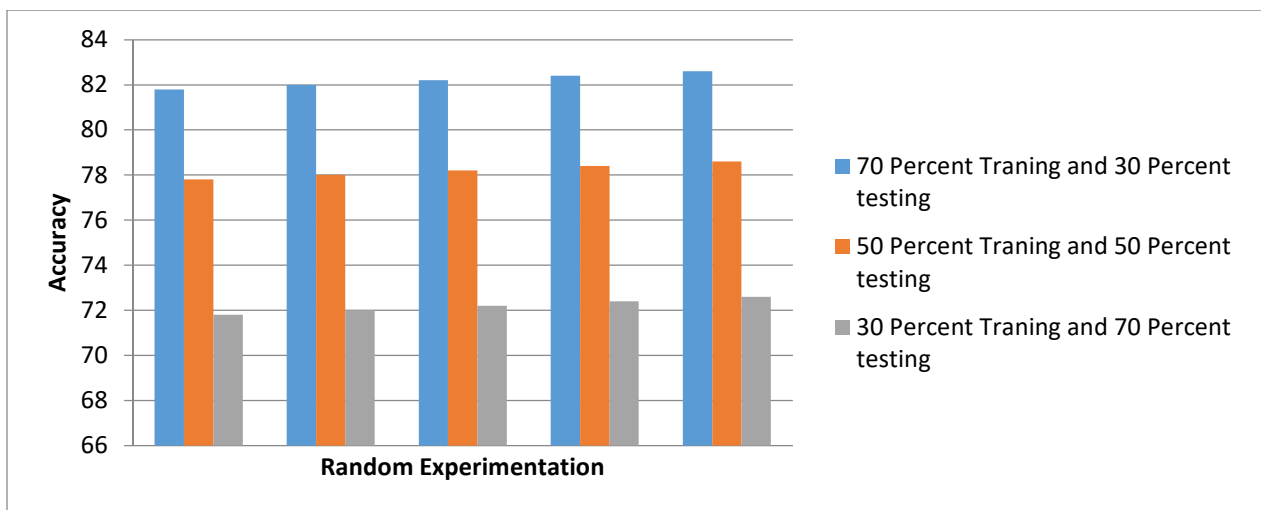
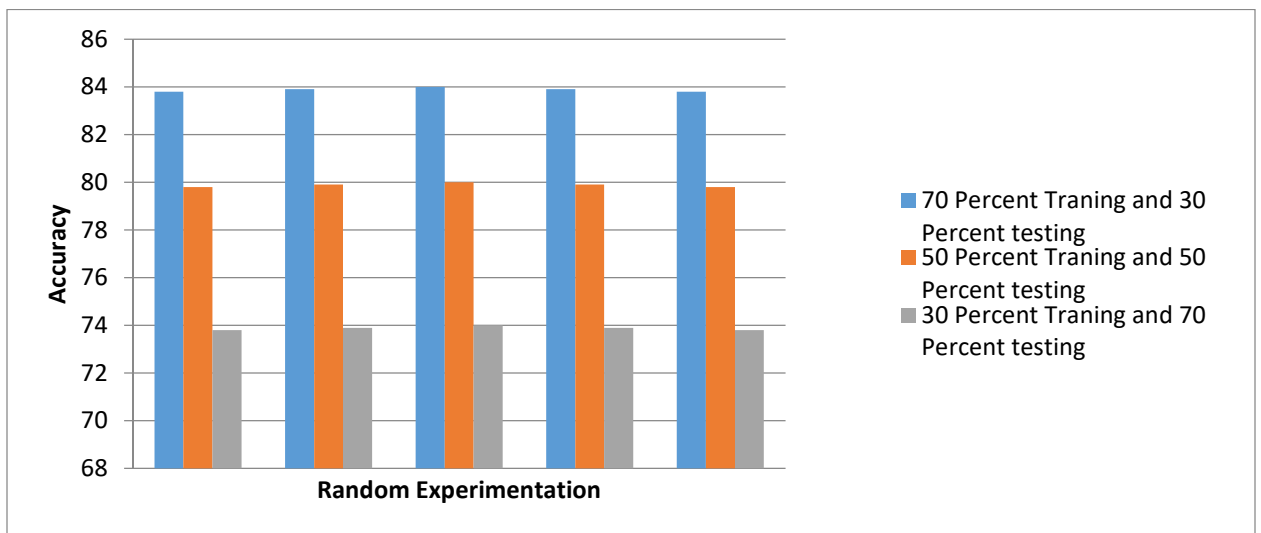
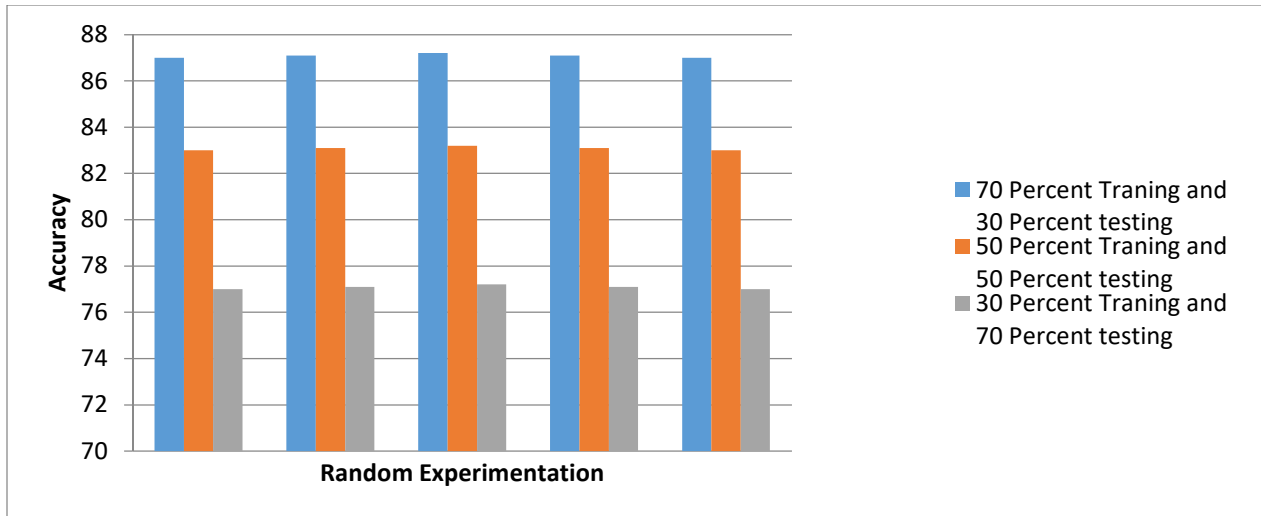


Figure 5: shows the Accuracy of HOG Features in Block 4 Partition



**Figure 6: shows the Accuracy of SIFT Features in Block 4 Partition**



**Figure 7: shows the Accuracy of Fusion Features in Block 4 Partition**

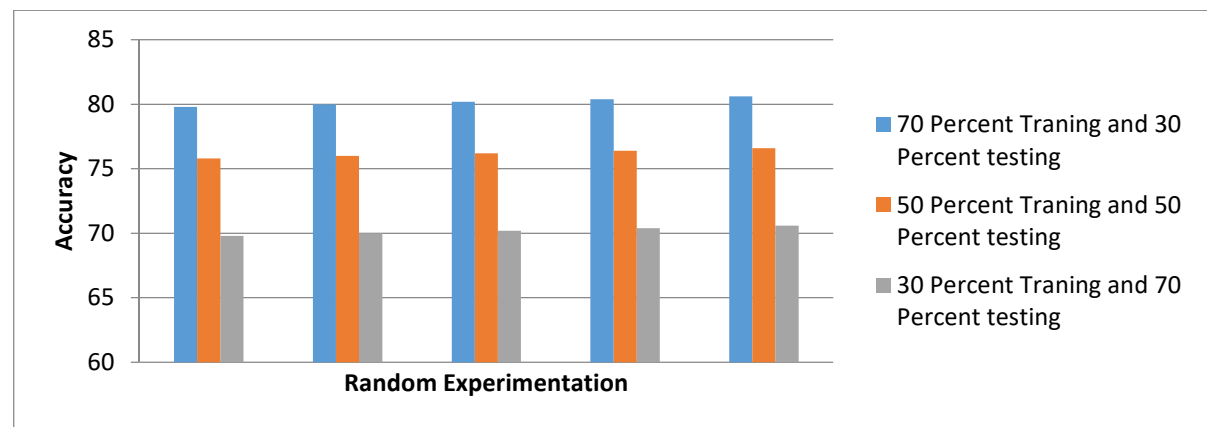
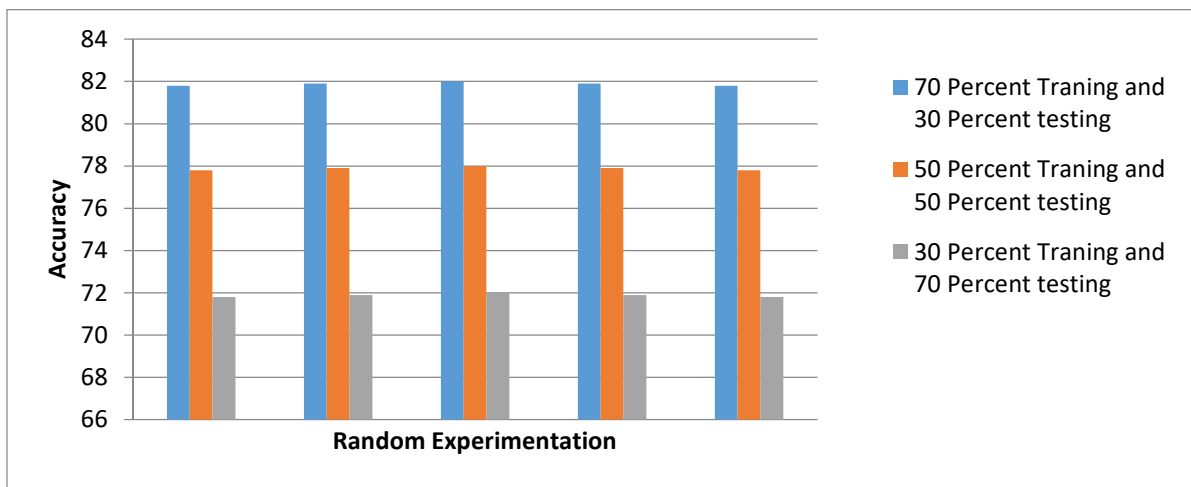
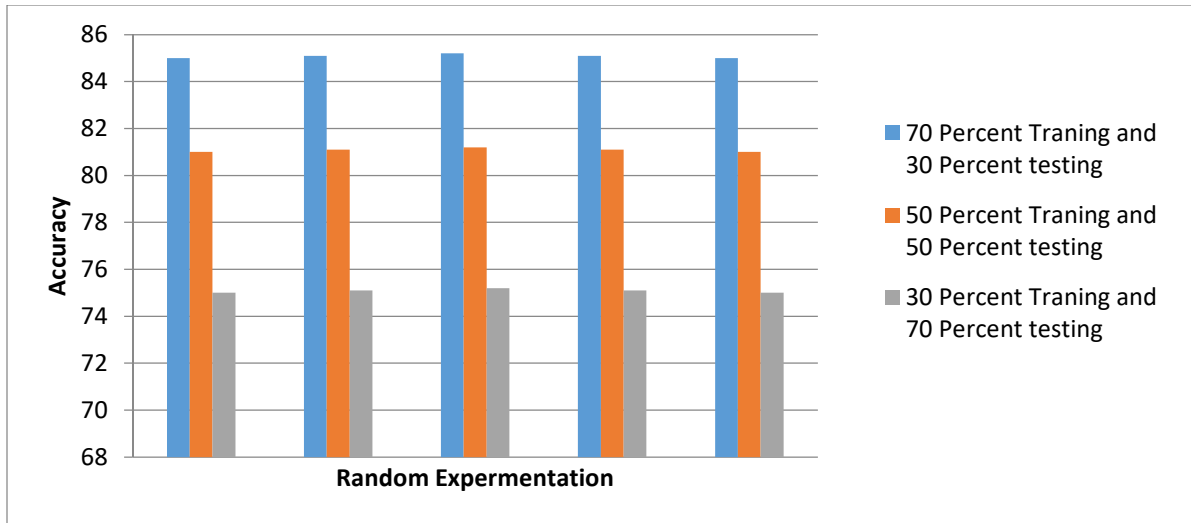
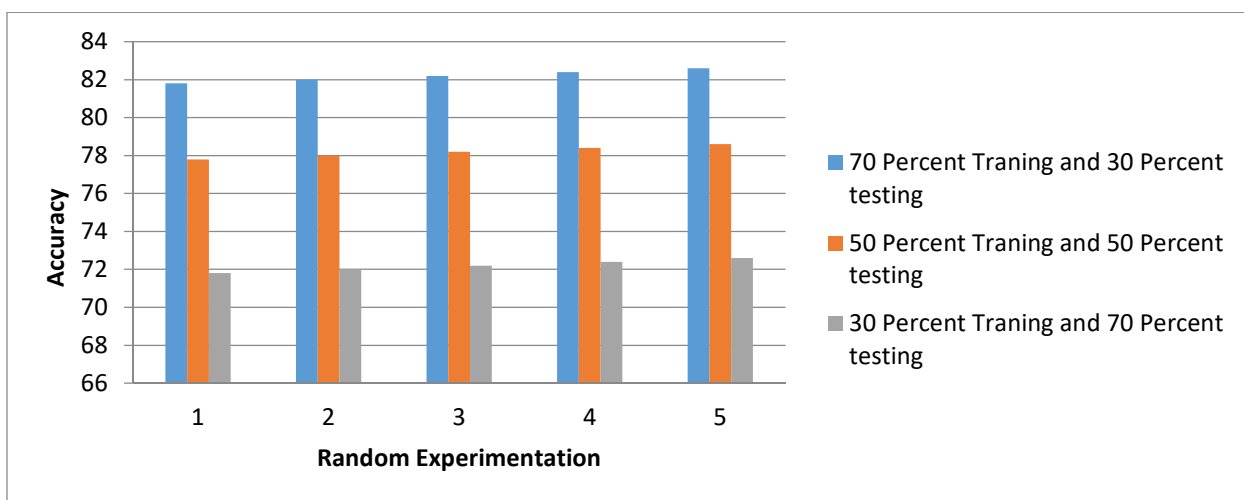
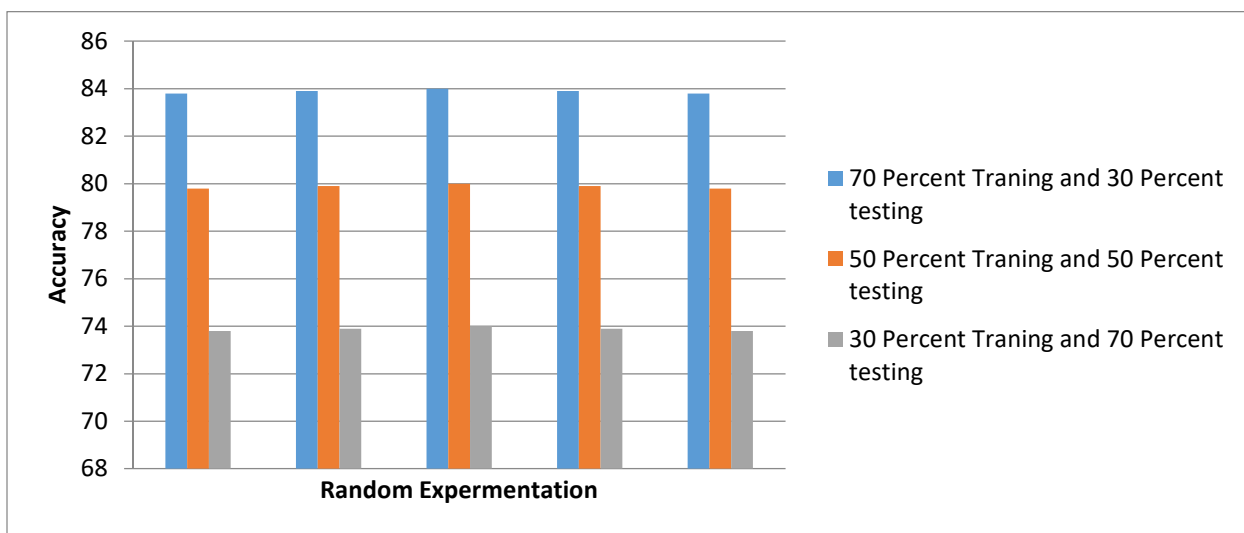
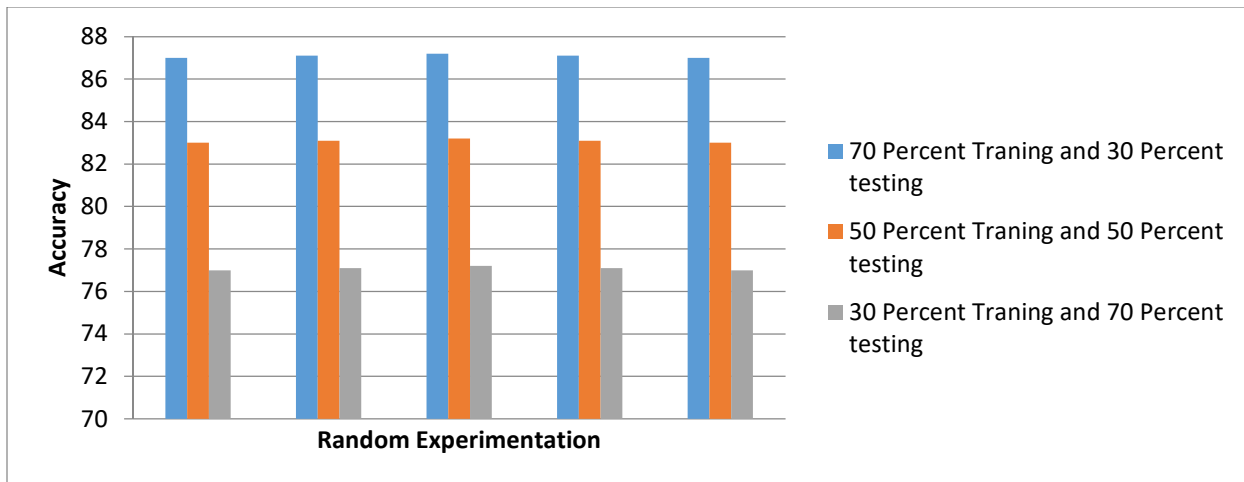
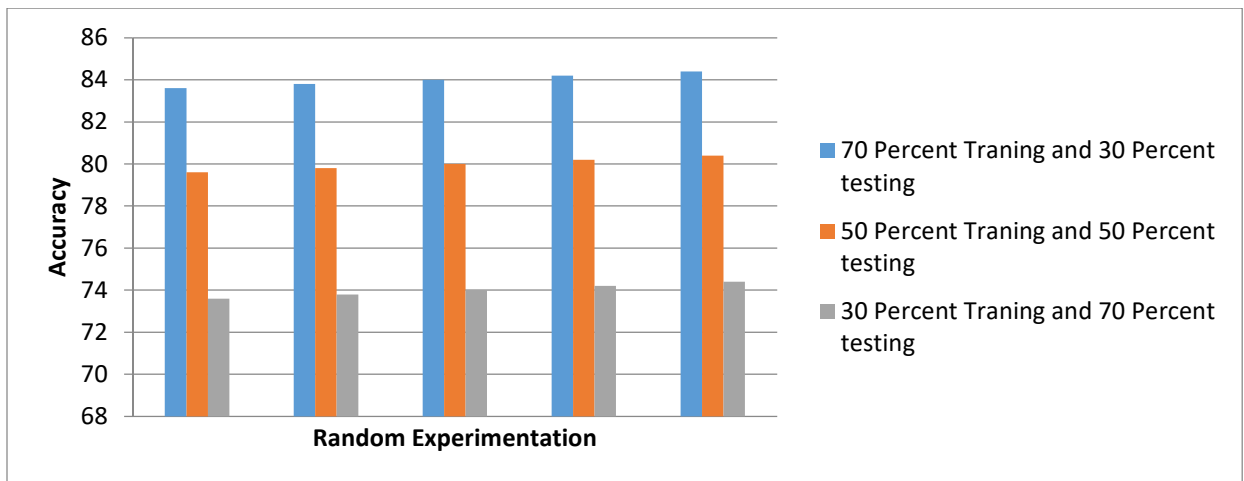
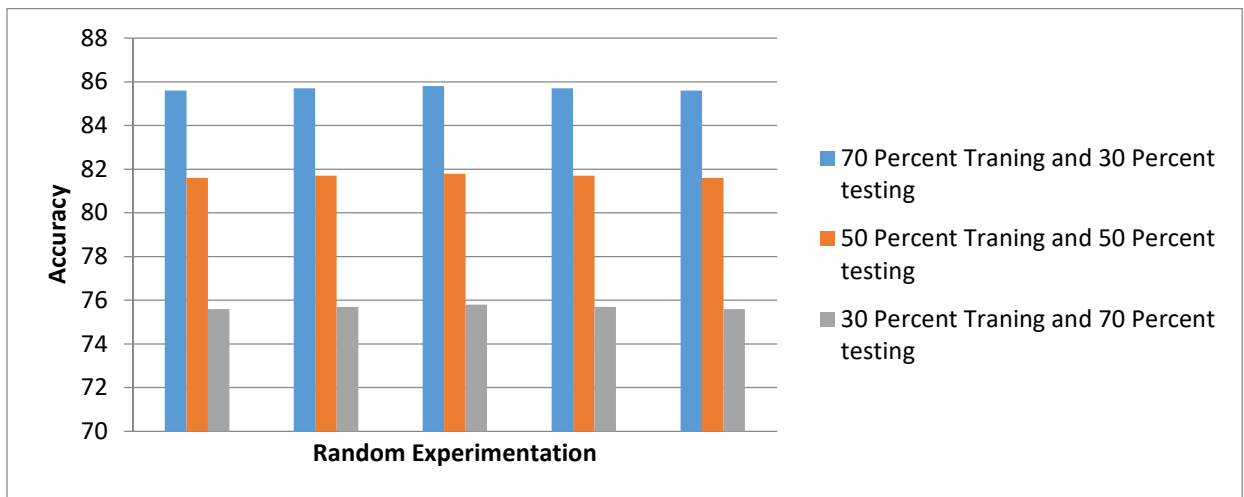
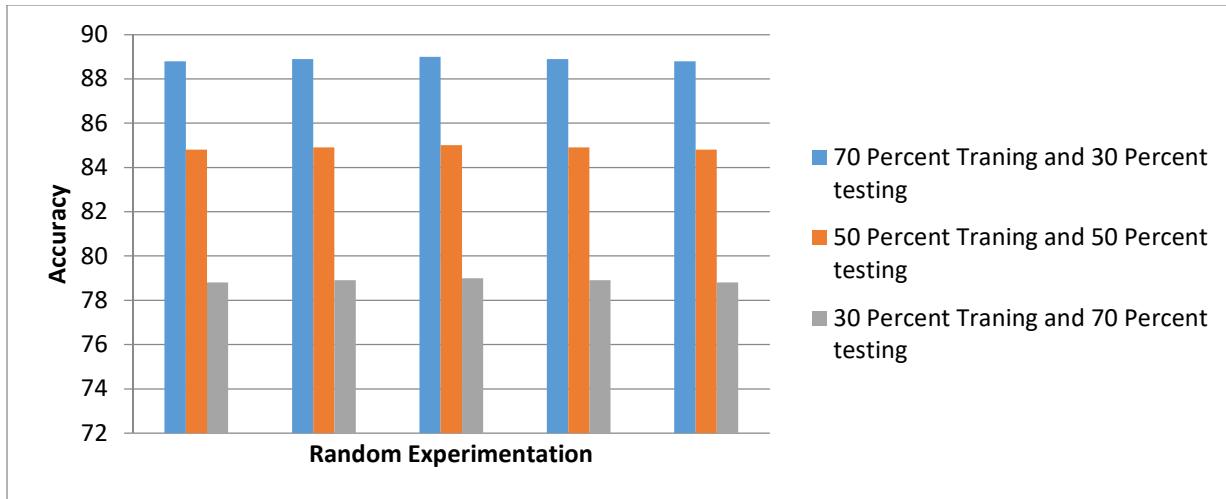


Figure 8: shows the Accuracy of HOG Features in Block 16 Partition



**Figure 9: shows the Accuracy of SIFT Features in Block 16 Partition**



**Figure 10: shows the Accuracy of Fusion Features in Block 16 Partition**

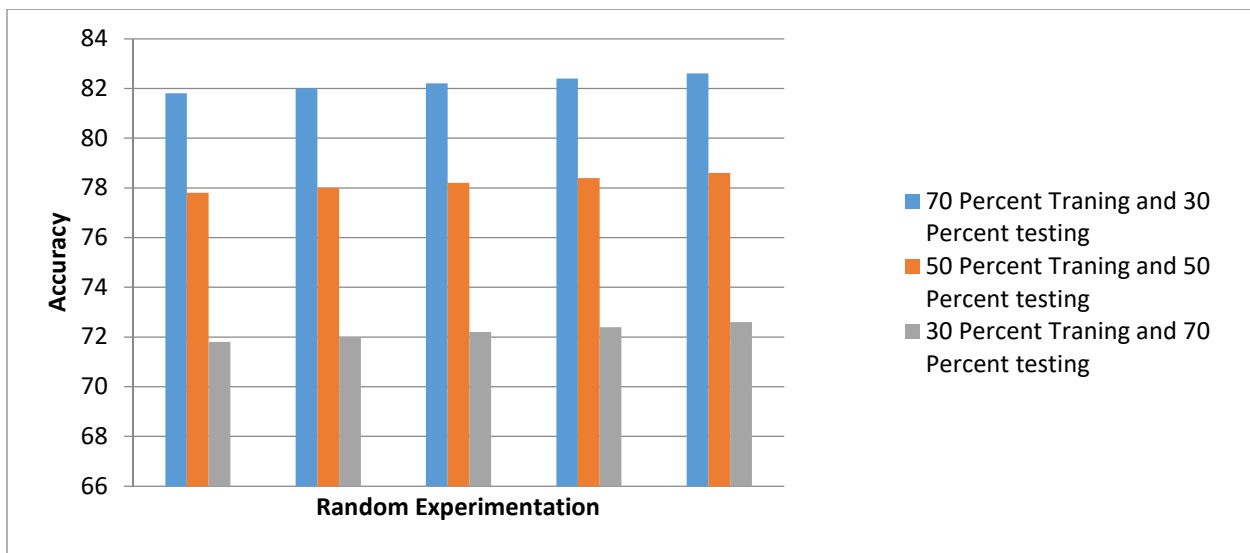
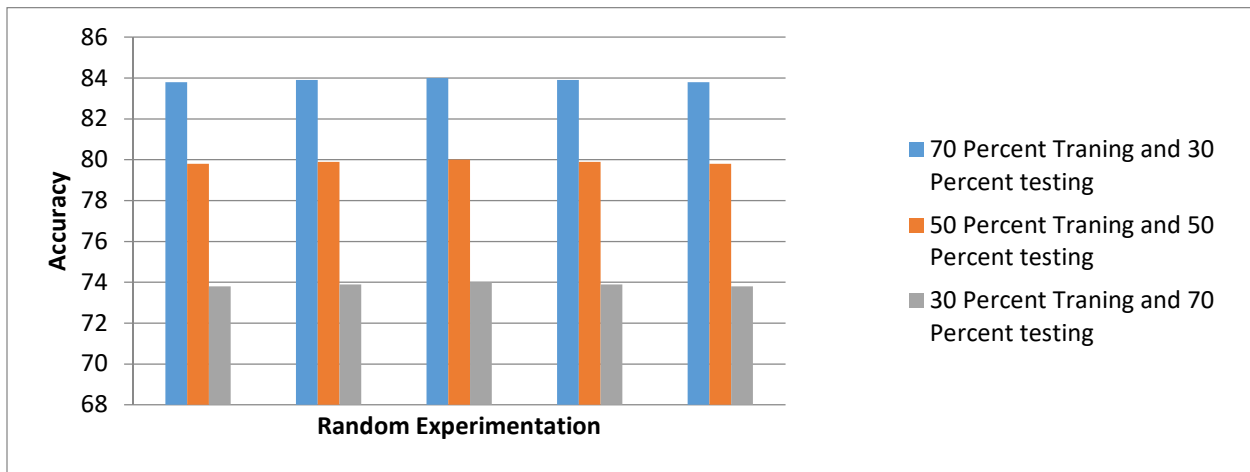
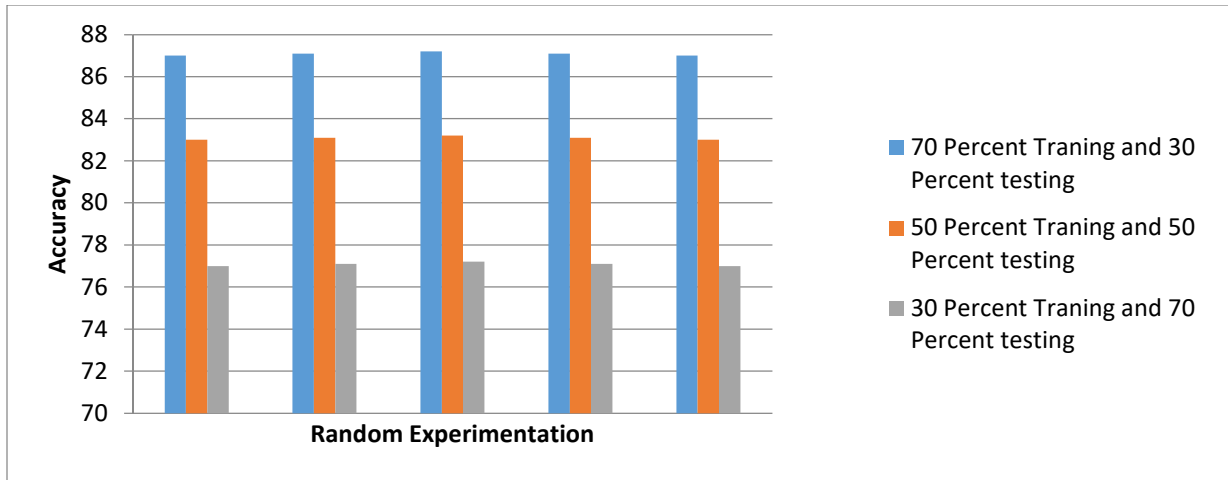


Figure 11: shows the Accuracy of HOG Features in Block 32 Partition



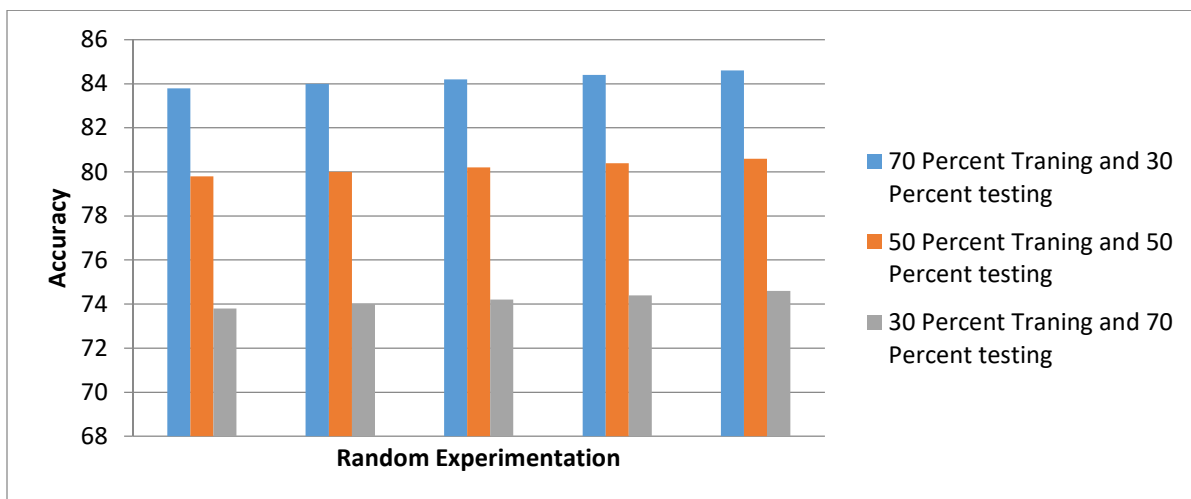
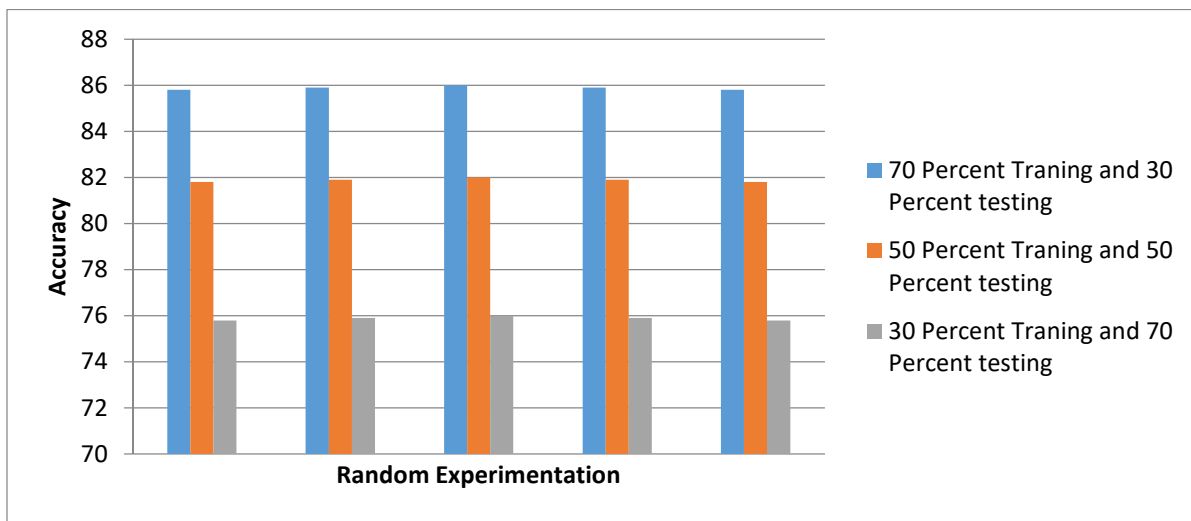
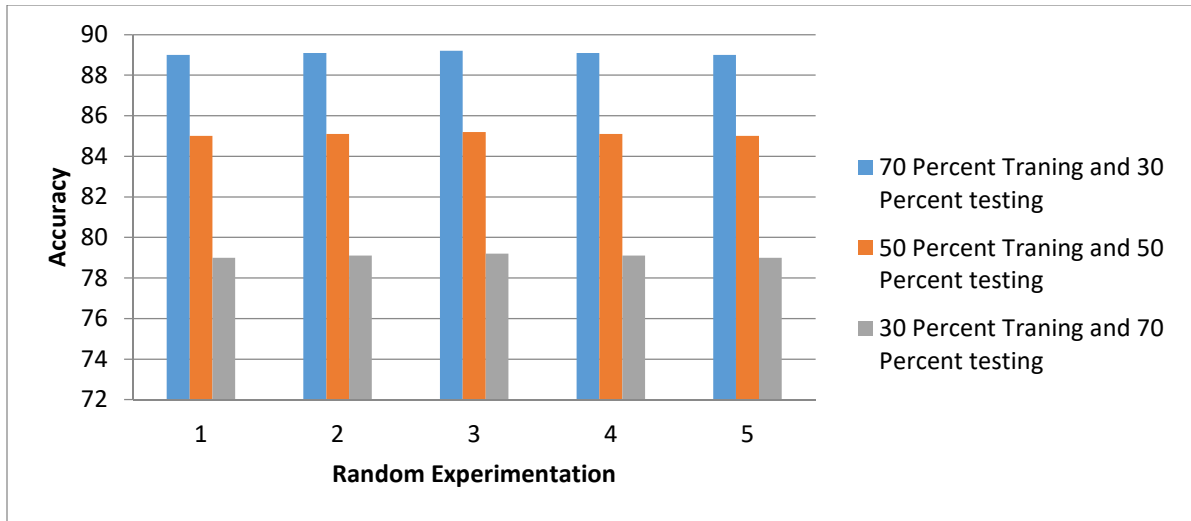
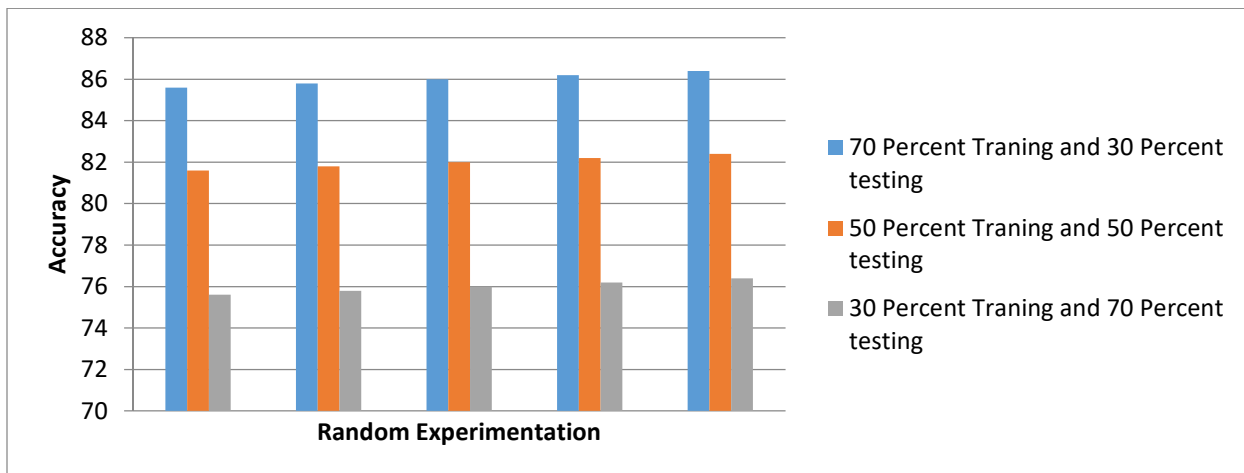
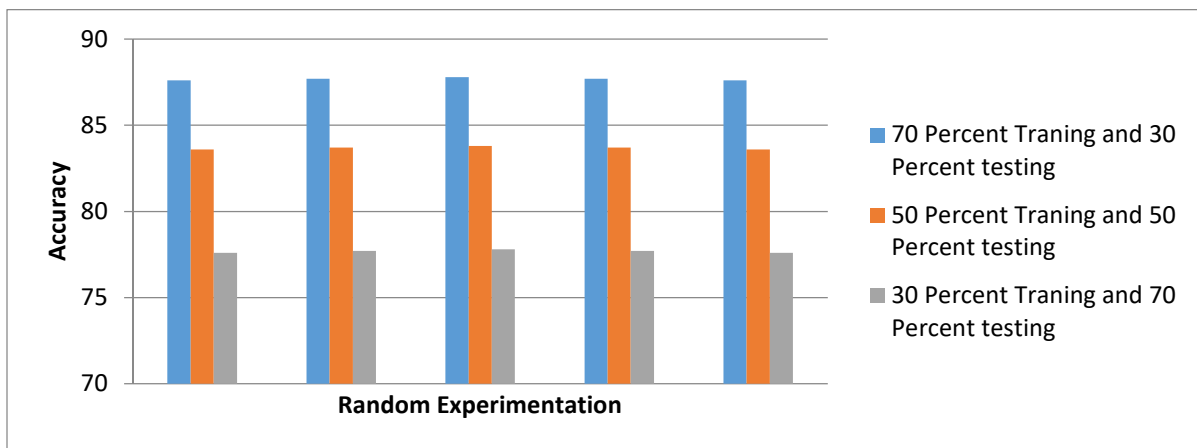
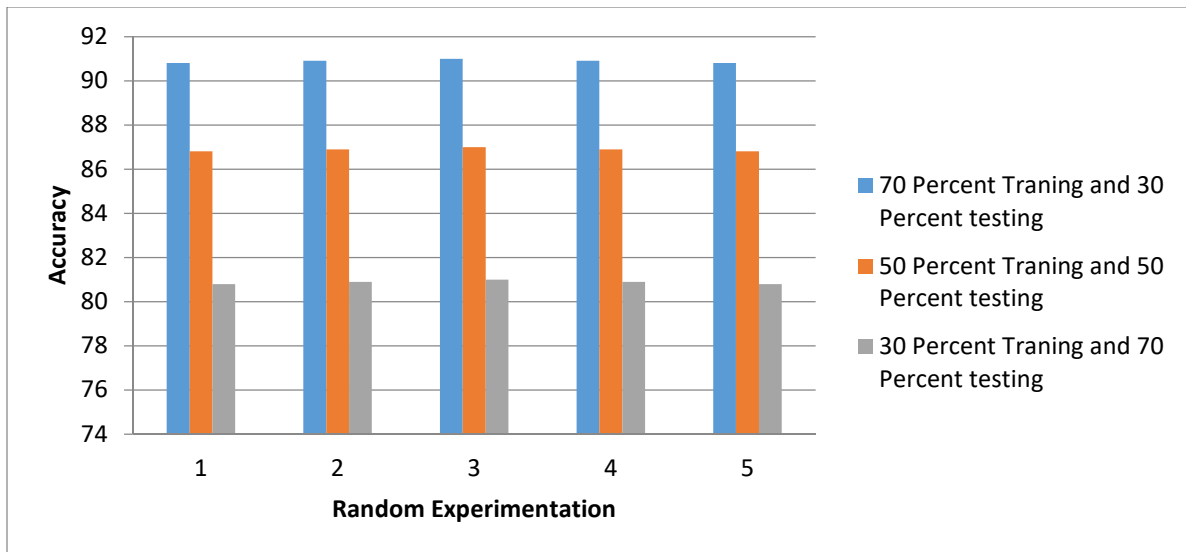


Figure 11: shows the Accuracy of SIFT Features in Block 32 Partition



**Figure 12: shows the Accuracy of Fusion Features in Block 32 Partition**

## 5. Conclusion

Investigating Historical document is not straight advance procedure because of low quality, differentiation, contrast and covering of characters. In this analysis, the authors propose a HOG and SIFT features with symbolic classifier to recognize Historical kannada characters. To begin with, the character is divided utilizing Connected Component Analysis and later the Different Features are detached. At long last, form a powerful Symbolic classifier with min-max and Mean-Std deviation representation to recognize the historical Kannada archives. Proposed tale schemes during the preprocessing stage to guarantee strong, precise and constant grouping. They assess their strategy all alone datasets their characterization results surpass 91% on all datasets, which are superior to the cutting edge in this space.

## References

1. Guru, Devanur & Nagendraswamy, H.' Symbolic representation of two-dimensional shapes'.Pattern Recognition Letters. 2007,pp .144-155.
2. Hassan, Ehtesham & Chaudhury, Santanu & Gopal, Madan & Dholakia, Jignesh. 'Use of MKLas symbol classifier for Gujarati character recognition'.2010,pp.255-262.
3. Guru Devanur & Harish, B S & Shantharamu, Manjunath.'Symbolic representation of text documents'.2010.pp.1-8.
4. Harish, B S & M B, Revanasiddappa & Shantharamu, Manjunath . 'Document Classification using Symbolic Classifiers'. Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014.
5. Harish, B S & M B, Revanasiddappa & Aruna kumar, S V.' Symbolic Representation of Text Documents Using Multiple Kernel FCM'. 2015.pp. 93–102.
6. D.S. Guru, K.S. Manjunatha, S. Manjunath, M.T. Somashekara, 'Interval valued symbolic representation of writer dependent features for online signature verification',Expert Systems with Applications,Volume 80,2017,pp.232-243.
7. Swarnalatha, K. and Guru, D. S. and Anami, B. S. and Suhil, M. 'Classwise Clustering for Classification of Imbalanced Text Data' In: Proceedings of International Conference Emerging Research in Electronics, Computer Science and Technology ICERECT 2018. 2019.
8. F. Alaei, N. Girard, S. Barrat and J. Ramel, 'A New One-Class Classification Method Based on Symbolic Representation: Application to Document Classification,' 2014 11th IAPR International Workshop on Document Analysis Systems, Tours, 2014, pp. 272-276.
9. T. N. Vikram, K. C. Gowda and S. R. Urs, 'Symbolic representation of Kannada characters for recognition' 2008 IEEE International Conference on Networking, Sensing and Control, Sanya, 2008, pp. 823-826.
10. M. Amrouch, Y. Es saady, A. Rachidi, M. El Yassa and D. Mammass, 'Printed amazigh character recognition by a hybrid approach based on Hidden Markov Models and the Hough transform' 2009 International Conference on Multimedia Computing and Systems, Ouarzazate, 2009, pp. 356-360.
11. C. Papaodysseus, P. Rousopoulos, D. Arabadjis, F. Panopoulou and M. Panagopoulos,'Handwriting automatic classification: Application to ancient Greek inscriptions'2010 International Conference on Autonomous and Intelligent Systems, AIS 2010, Povia de Varzim, 2010, pp. 1-6.
12. Singh, Rahul R., Chandra Shekhar Yadav, Prabhat Verma and Vibhash Yadav. 'Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network'. IJCSC, 2010,pp. 91-95.

13. P. Rousopoulos et al., 'A new approach for ancient inscriptions' writer identification,' 2011 17th International Conference on Digital Signal Processing (DSP), Corfu, 2011, pp. 1-6.
14. Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 'Recurrent convolutional neural networks for text classification'. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press, 2015, pp.2267–2273.
15. Kuttala, Radhika & K R, Bindu & Parameswaran, Latha. (2018). 'A text classification model using convolution neural network and recurrent neural network'. International Journal of Pure and Applied Mathematics. 2018, pp. 1549-1554.
16. A. Garz, M. Diem and R. Sablatnig, 'Detecting Text Areas and Decorative Elements in Ancient Manuscripts' 2010 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, 2010, pp. 176-181.
17. Garz, Angelika & Sablatnig, Robert & Diem, Markus. (2011). 'Using Local Features for Efficient Layout Analysis of Ancient Manuscripts'. European Signal Processing Conference. 2011, pp. 1259-63.
18. S. Choudhary, N. K. Singh and S. Chichadwani, 'Text Detection and Recognition from Scene Images using MSER and CNN' 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAIECC), Bangalore, 2018.