Feature Selection For An Automated Ancient Tamil Script Classification System Using Machine Learning Techniques

T S Suganya,
Research Scholar,
Sathyabama University Chennai,
Tamilnadu, India Email: tssuganya@yahoo.com

Dr. S Murugavalli,
Professor and Head,
CSE Department, Panimalar Engineering College, Chennai,
India

***Abstract*—Tamil is one of the oldest languages in the world, spoken in Tamil Nadu, South India, which is inherited from Brahmi Script. The main source of information about history are the stone inscriptions. OCR aids in digitizing Tamil scripts from the ancient and old era to the latest, making its access easy through Internet. Ancient Tamil character recognition from stone inscription is a challenge due to the large disparities of writing style. Efficient feature extraction and selection is essential for effective Ancient Tamil character recognition system. The aim of this paper is the use of Shape and Hough transform for feature extraction using Group Search Optimization and Firefly algorithm for feature selection to recognize the ancient Tamil script.**

I. Introduction

Tamil is one of the major language and its writing is in the combination of alphabetical and syllabic system. Ancient Tamil inscriptions are in Tamil-Brahmi form. The inscriptions date back to 3rd century. Epigraphs are translated forms of ancient inscriptions which are deciphered from inscription on stones, palm leaves etc. The epigraphs contain historic information regarding kingdoms, lifestyle, military strategy, business and also medicines. The present Tamil language is not based on the Tamil-Bramhi but on Pallava Granth. The ancient Tamil has been written in wide range of continuum of scripts.

With the advent of image processing, automated script recognition is gaining popularity. The approach of reclassifying characters is in incorporating structure analysis typographically which categorizes characters in the initial step and thus reducing the scope of character recognition. Printed and hand written characters can be recognized through Optical Character Recognition (OCR). The problem can be addressed through a number of approaches the features vary from the graphical representation of characters.

Image processing and pattern recognition is achieving utmost importance of late and script recognition is one of the most challenging and intriguing area of research in pattern recognition [1]. This contributes to the advancement of automation and the interface between man and machine is improved in many applications. In the case of handwritten character recognition, the characters are written in a sheet of paper by the individuals which is stored in a computer through a scanner and later on these characters are recognized; while in machine printed system the texts or characters are typed from input devices like keyboard or typewriter and so on. These are taken from books, journals, newspapers etc. Similar techniques may be applied to script recognition by scanning the inscriptions from stone, palm leaves etc. Stone inscriptions are rich source of information on various aspects of life during ancient time. In this work, an automated ancient tamil script recognition using machine learning techniques is presented.

OCR plays an important role in digitizing documents in Tamil from ancient era to the latest by sharing data through internet. In Indian language like Tamil due to its inter-class dependencies, character recognition is tough. Moreover in Tamil many characters look alike and so classification is a big challenge [2]. Several steps of preprocessing are involved in automatic script recognition which helps to reduce the variation in the texts as much as possible, simultaneously preserving the information that is relevant for recognition. Preprocessing of handwritten text line, does not have any general solution but it relies mainly on character size, slope and slant correction of characters. The handwritten word is rotated with slope correction such that the alignment of lower baseline is in the horizontal axis of the image.

The clockwise direction between vertical angle and direction of vertical slopes is slant. Word is converted to an upright position in slant correction. In a word image, the removal of slope and slant leads to a word image that is independent of every other factor. Ultimately, due to the normalization of size the system becomes invariant to the size of the character and the empty background areas are reduced caused by the ascenders and descenders of some letters.

Nonexistence of standard/benchmark databases is the greatest challenge for handwritten character recognition of Indian scripts. Two steps are involved in the character recognition procedure:

- Feature extraction – here ever character is depicted as a feature vector;
- The vectors are classified into a number of classes.

Feature extraction is the process of extraction of the most representative information from the data that is raw, where the within class pattern variability is minimized while the between class patter variability enhanced. A set of features are extracted for this purpose for each class which helps it to distinguish from other classes, within the class, it remains invariant to characteristic differences. In general features are of two kinds, structural and statistical features. Pixel density, mathematical transformation, and moment etc., are included statistical features while stroke, contour, number of circles and bifurcation points are included in structural features.

Style variations are taken care of up to a certain extent when a document is represented through statistical distribution. Structural features have its basis on topologic and geometric properties of the character. A number of global and local properties are represented through topological and geometrical features with variation in style and distortions from high tolerance. A combination of both these features can provide maximum benefit. The most distinguishing properties of pattern or character form can be provided through structural features but with less noise resistibility.

Though statistical features have the disadvantage of overlooking the specific information on the form of character, it assures a better guard against interference [3]. So a combination of structural and statistical feature works better by overcoming the misgiving of each other. There are certain common features such as the beginning and ending of character stroke, number of strokes, existence of loops and their number, time order of strokes, number of intersections and free ends, vertical and horizontal symmetry and so on aid the feature selection in handwritten recognition technology and these are reached from the characteristics of the sequence of the character.

Features of statistics that are hidden from human mind are brought out by mathematical observation and calculation. The most common and visible feature in statistical feature election approach is that of directional feature. Simultaneously one can reduce over-fitting of learning method and increase speed of prediction computationally, by reducing the number of features. From a set of available features, feature selection can be deemed as the process of selection of the subset allowing the highest discriminative power. In any classification problem it is vital to choose a good feature subset [4].

For a given feature set Y of cardinality N, the feature selection involves the growth exponentially of the search space of the order 2N. So, for detecting near-optimal solutions, heuristic algorithms are necessary. Filter and wrapper are the two types of feature selection methods. As the filter methods are not dedicated to any specific type of method of classification, they are grouped as classifier dependent; while on the other hand, wrappers basically rely on the performance of one type of classifier in evaluating the quality of a feature set.

The first and foremost step in ancient Tamil recognition system is preprocessing, which is followed by segmentation and feature extraction. In preoprocessing, the input image is shaped into a form that is suitable for segmentation. In segmentation, the image is segmented into individual character and for the process of training the network, each character is resized into m x n pixels [5]. The next step is feature selection based on recognition performance. Ancient Tamil characters can be recognized through various feature extraction methods. The widely used feature extraction methods are Template matching, Image transforms, Projection Histograms, Contour profiles, Zoning, Geometric moment invariants, Fourier descript ors, and Gradient feature and Gabor features. With respect to offline ancient Tamil character recognition system, neural networks form the fastest and the most reliable tool for classification for achieving the best recognition accuracy.

In this proposed study, feature extraction is done through Shape and Hough transformation and feature selection through Group Search Optimization and Firefly methods. Feature classification is done through Neural Network, J48, Naïve Bayes and k Nearest Neighbor classifiers. Section 2 of the paper discusses related works over hand writer character recognition; section 3 presents detailed overview of the proposed framework; and section 4 deals with comparative results with discussion. Section 5 concludes the report.

II. RELATED WORKS

Rajakumar and Bharathi [6] proposed an artificial immune algorithm for ancient Tamil character recognition which proved to be efficient and comparatively better than neural network in recognising the ancient Tamil characters.

Subashini [7] presented a technique for predicting the period from the Tamil script using SVM. Experiments proved that the proposed method is effective in distinguishing four centuries of Tamil character.

Soumya & Kumar [8] used Nearest Neighbor clustering algorithm for segmenting characters from epigraph images and classified the character. It was shown that the proposed system efficiently recognized ancient text and maps into equivalent modern character.

Azmi et al., [9] proposed a hybrid feature selection technique that is based on genetic and simulated annealing algorithms. A dataset of 100 samples were considered for each 33 hand-printed characters of Farsi hand printed characters and the proposed approach was evaluated using Bayesian classifier. With the consideration of two minimum and maximum thresholds, simulated annealing, aided in improving the acquired results.

Sridevi & Subashini [10] used Extreme Learning Machine to classify handwritten ancient Tamil scripts. Zenike moments and regional features train the Learning Machine. Performance of Extreme Learning Machine and Probabilistic Neural Networks are compared. Highest accuracy of about 95% was given by Extreme Learning Machine according to the results of this study.

An adaptive segmentation technique was proposed by Farkya et al., [11] which showed less and implementation of handwritten Devanagri character recognition system. Various morphological operation and thinning techniques were done for preprocessing. From 10 different individuals, training was done using self-created database consisting of 50 samples. Digitization of the characters were done using Unicode the database consisting of these characters were linked and speech signal was formed. A database that was self-created with all the phonemes was created.

Random Forest (RF) Machine Learning Technique was used by Pant & Bal [13] for printed Nepali text using a hybrid OCR system – for Holistic and Character level recognition. First, the system tried to recognize the word as a whole, if word-level recognition does not take place, then character level recognition is performed. Feature vector of a word or character is defined through Histogram Oriented Gradients (HOG) descriptors. Thus character level recognition method and Hybrid method reached recognition rates of approximately 78.87% and 94.80% respectively.

III. METHODOLOGY

The flowchart in figure 1 shows the proposed framework for the Tamil character recognition.

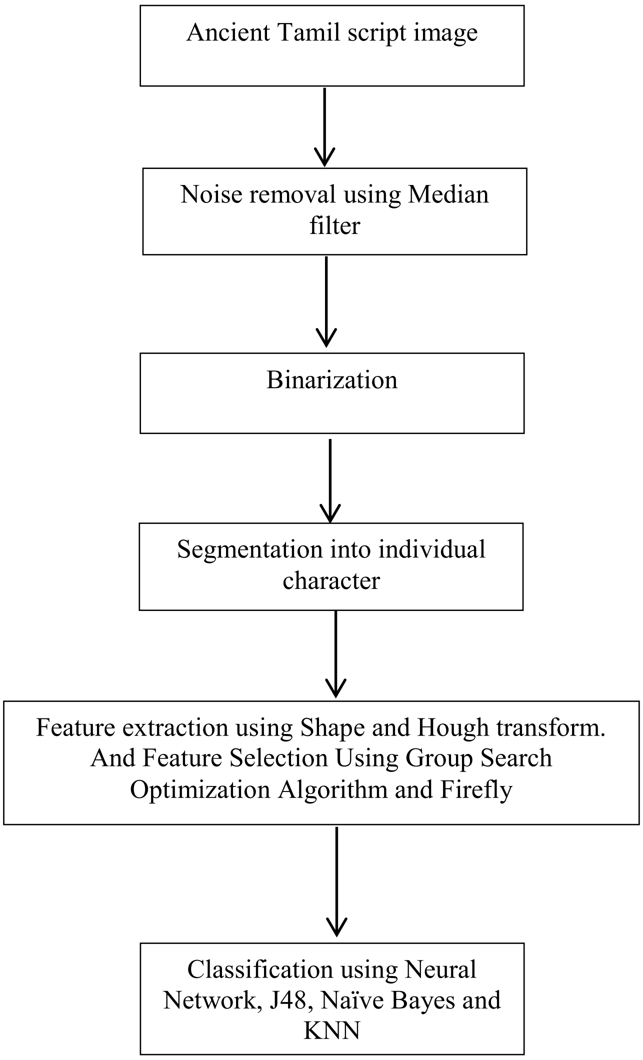

Fig. 1. Flowchart for the proposed framework

Noise Removal

During the process of image acquisition noise is introduced. A random variation of image intensity is produced and will be visible as grains in the image. The process of removing or mitigating noise from the image is termed as noise removal [21]. The various noises that exist in document images are Salt and Pepper noise, Gaussian noise, Gamma noise, Uniform noise, and so on. The types of filtering methods are Gaussian filtering method, Min-max filtering method and so on are used for noise removal. Salt and pepper noise is removed through median filter.

Median filter is a form of nonlinear filter based on smoothing technique such as linear Gaussian filtering. All filtering techniques in general can remove noise effectively in smooth patches or smooth areas of a signal but a lot of its edges are affected. In text extraction and recognition, often, noise should be removed from image and preserving the edge is vital. For visual appearance of images, edges are vital. So, median filtering is mostly used in digital image processing [22].

Binarization

The value of pixel varies from 0 to 255 for any grayscale captured image. Grayscale value is thresholded for any binarization process and it is being white ('0') for background pixel and black ('1') for foreground pixel. Thresholding is the process of binarization of any image in which conversion of gray scale image takes place by a definite threshold value. Local and global thresholding are two groups of thresholding. Different threshold values are utilized by local or adaptive thresholding for each pixel on the basis of local area information; whereas in global thresholding, one threshold is utilized for the entire image which has its estimation through intensity histogram. For images having varying levels of intensities local thresholding is used, for instance, images from satellite cameras [23]. While, global thresholding would be enough for images that are simple, including handwriting on white background.

Segmentation

One of the principal stages of OCR processes include segmentation. The accuracy of recognition will be increased with the use of very good segmentation method. In this phase, subdivision of the input images containing sequence of characters take place, which divides into sub images of isolated characters. In this proposed system, segmentation takes place through line and character segmentation. In line segmentation, for each row, the horizontal projection histogram of pixels is determined so as to differentiate between regions of high density (lines) and regions with low density (interspace among the lines.

Handwritten image is taken as input image for segmentation input and is resized at 512 * 512. The input image should be converted into grayscale from RGB. Canny edge detection, dilation and erosion is applied after binarization. Unwanted pixels are removed through these morphological operators before generating final segmentation output. The output is treated as segmentation results after morphological operations and each character image shows in bounding box [25].

Feature Extraction

In the proposed study, Shape and Hough transform methods of feature extraction are used.

Shape Transformation

The shape transformation algorithm has its resemblance to a certain extent dynamic programming technique which is used for lexicographical corrections, as post-processing step in OCR. The following steps are followed in the transformation of character image x into image y:

a) Some pixels are left intact (substitution);
b) Some pixels of x are moved to their pixels corresponding to the closest pixels in y irrespective of their locations;
c) Depending upon necessity, some pixels are deleted.

Due to the differences in context, insertion, deletion and substitution matrices are not available and we had to invent our own. As they are both common to images (characters), there are no cost to substitutions, as this is equal to leaving them in place. The cost of moving a pixel to yet another location is equal to that of the location to the Euclidean distance between and source and destination. As either the pixel in x is substituted or translated, the pixels that are corresponding in x and y are associated. The associated pixels do not get involved again [26].

We may be in a position to face two possible situations after all substitutions and translations:

a) Pixels might be remaining in x;
b) Pixels that are not associated might remain in y.

Pixels still remaining in x are deleted. The cost of deleting a pixel image x is equal to that of the cost of moving pixel outside (distance from) the character frame, in a straight-line path, depending on the available closest route.

Hough Transform Based Methods

Through Hough Transform, imperfect images are found with instances of objects within particular class of shapes using a particular procedure of voting. The procedure of voting takes place in a parameter space, where the objects are obtained as local maxima in a presumable accumulator space which is constructed effectively by the algorithm in computing Hough Transform. Straight lines such as ruled line can be found through an image. Line is found in every direction by the extraction of dominant features of an image; therefore, these group of methods are comparatively robust against nose and also cope better with broken lines.

Feature Selection

For the achievement of good recognition results, the design of suitable features is crucial. The features should be in such a way that they represent the maximum relevant information for classification purpose at hand. Here the determination takes place by enhancing the inter-class pattern variability and minimizing the intra-class pattern variability.

Firefly Algorithm (FA)

he inspiration for Firefly Algorithm (FA) comes from biochemical and social aspects of real fireflies. Real fireflies attract their mating partners through the rhythmic flash which also acts as a protective warning mechanism. The problem to be optimized is correlated with the flashing behavior of the firefly. For basic formulation of FA, the following rules are idealized [29]:

(1) As fireflies are unisex, they attract their partners, regardless of sex.

(2) Brightness decreases as the distance increases and so, attractiveness is proportional to their brightness. So, automatically, the less brighter one will move towards the brighter one. If in case it is unable to detect more brighter ones then they move arbitrarily.

(3) The landscape of the objective function is the basis for the brightness of a firefly

Group Search Optimization (GSO)

Animal searching behavior is the basis for GSO which is a nature-inspired algorithm. Based on the fitness of the individuals the hierarchy is defined. From a group of animals, producers, scroungers and rangers are selected. Producer is the fittest among the population who acquires the highest value of optimization; the remaining 80% of the population is chosen as scroungers and rest as rangers [31].

Neural Network

A data modeling tool that can capture and represent complex input/output relationships is Neural Network. The inspiration for neural networks stems from the "intelligent" tasks performed by brain. In the following manner, neural network resembles human brain [32]:

1. Learning is the route for knowledge in artificial neural network;
2. Synaptic weight is the interneuron connection strength where knowledge is stored.

Character recognition is done due to high noise intolerance in ANN, which can produce good results. The step which is of utmost importance is feature extraction step. In case of any ANN, set of features that are chosen poorly, yield poor rate of classification. When the current stage of development is considered, a software performs well in terms of speed or accuracy though not better [33].

J48

An improved version of the C4.5 algorithm is decision tree induction algorithm. A top-down recursive divide-and-conquer manner is used in this protocol and it is supposed to be a greedy algorithm. Sub tree replacement and sub tree raising are the two pruning methods employed by J48 [34].

In a decision tree, the sub tree replacement nodes are replaced with a leaf, to reduce the number of test in a certain path. The process starts from fully-formed leaf to root. In the case of sub tree raising, the node may move upwards towards the root of the tree where the nodes are replaced along the way. On decision tree models, sub tree raising has negligible effect.

For training instances

1. Construction of tree takes place in a top down recursive divide and conquer manner
2. Root node is the feature that is having the highest information gain.
3. The attribute that gives us the next highest information gain is selected.
4. Step 2 and 3 are repeated until reach from the root node to the leaf node

For test instances

1. The new instance are classified based upon this decision tree Stopping criteria
2. All sample for a given node belong to the same class
3. If there is no remaining attribute for further partitioning

K- Nearest Neighbor Classifier (KNN)

In a feature space, the objects are classified with regards to the closest training examples in the case of KNN. A same class label is assigned to the test sample as that of its K-nearest neighbors. Proper choice of K and a metric used to measure the neighbors distances are the basis for the performance of KNN classifier [35].

x_j is assigned as the most frequently represented class in the k nearest training samples. Distance metric is the basis for neighborhood. Scaling of data is done through zero mean and one standard deviation.

Naive Bayes classifier

Multiple applications of information processing including language processing and information retrieval are done through Naïve Bayes (NB) classifier. This has its basis on probabilistic Bayes' theorem (from Bayesian statistics) with strong assumptions without dependence. The effect of a variable value on a given class is not dependent on the values of other variables in this classifier. Conditional probabilities are computed for the classes given the instance and the class with highest posterior is picked. In a supervised learning setting, Bayes classifiers can be trained based on the nature of the probability model.

Feature concatenation

By connecting the aligned units, concatenation operations are executed. At the feature level, fusion is performed in the selection of features and are combined to eliminate redundant and irrelevant features. By an 8-dimensional feature N x N sub regions by an 8-dimensional feature is described, $d = N \times N \times 8$. Finally a variable transformation $Y=X^{0.4}$ is applied on each element of the extracted feature vector X.

A character feature list is generated from training samples which stores sample features occurring for each character. Suppose that the input word W has sequences of characters C_1, C_2, C_3... Cm, where m is the total number of characters making up the word, then, sample features of the word are generated as all combinations of sample features of each character, yielding w sample features of the word computed as [38]:

$$w = \prod_{i=1}^{m} n(C_i)$$

Wherein $n(C_i)$ signifies the total quantity of samples for character C_i. Once the transformation model is trained with sample features of the word, the model is stored into a master model file which will be used later during the recognition phase.

IV. RESULT AND DISCUSSION

For the experiments, 9 ancient characters with each character containing 35 samples each were used.

Table 1 Classification Accuracy

Classification Accuracy	**J48**	**KNN**	**NN**
GSO based feature selection	93.65	92.38	94.29
Firefly Based Feature Selection	93.33	91.75	93.97

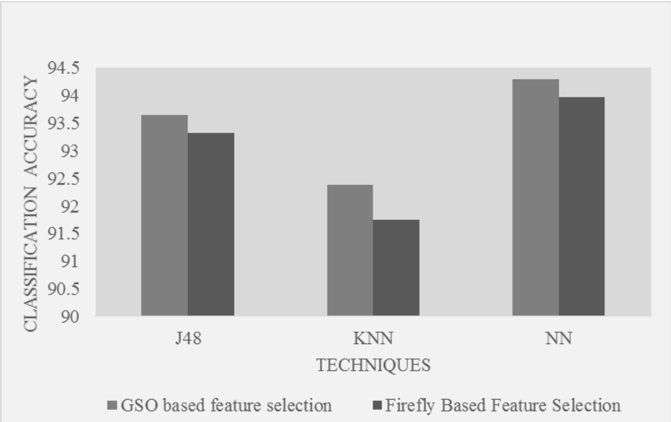

Fig. 2. Classification Accuracy

From the figure 2, it can be observed that the Feature selection with classifier improved accuracy. Neural Network classifier improved accuracy for concatenated features averagely by 2.05% and 2.39% than KNN classifier with GSO and Firefly based feature selection respectively

Table 2 Precision

Precision	**J48**	**KNN**	**NN**
GSO based feature selection	0.936533333	0.923822222	0.942888889
Firefly Based Feature Selection	0.933355556	0.917477778	0.939711111

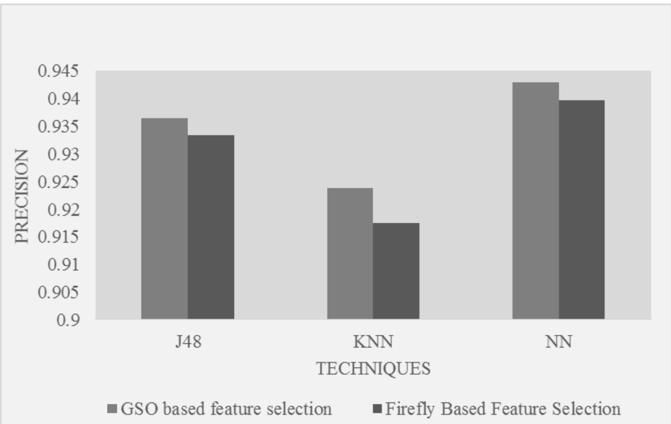

Fig. 3. Precision

From the figure 3, it is observed that the Neural Network classifier improved precision for concatenated features averagely by 2.04% and 2.39% than KNN classifier with GSO and Firefly based feature selection respectively.

Table 3 Recall

Recall	**J48**	**KNN**	**NN**
GSO based feature selection	0.936955556	0.924877778	0.943488889
Firefly Based Feature Selection	0.933855556	0.918711111	0.940477778

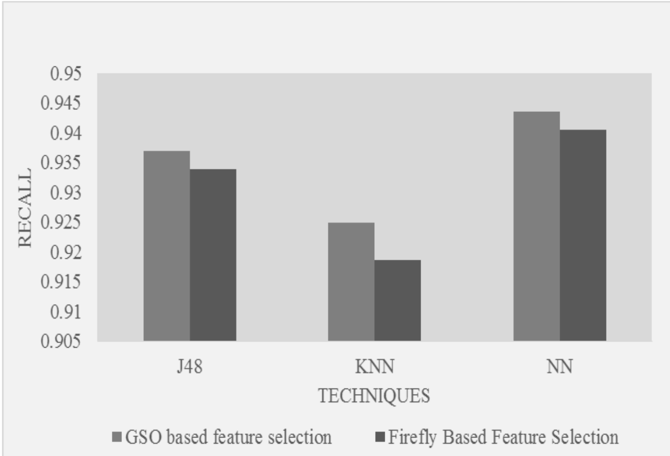

Fig. 4. Recall

From the figure 4, it is observed that the Neural Network classifier improved recall for concatenated features averagely by 1.99% and 2.34% than KNN classifier with GSO and Firefly based feature selection respectively.

Table 4 F Measure

F Measure	**J48**	**KNN**	**NN**
GSO based feature selection	0.936566667	0.924044444	0.943
Firefly Based Feature Selection	0.933377778	0.917788889	0.939855556

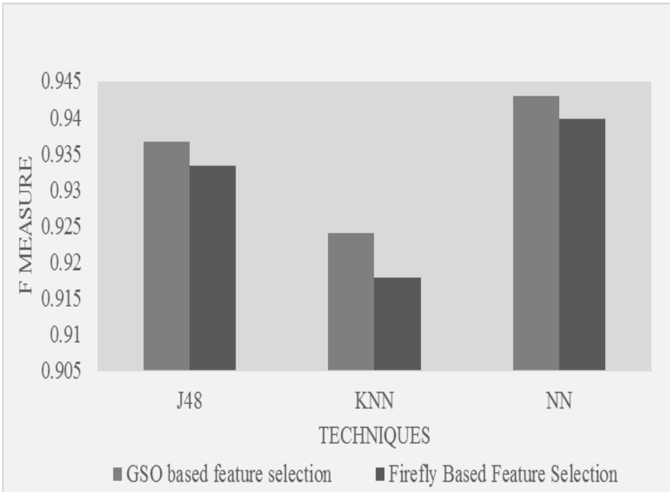

Fig. 5. F measure

From the figure 5, it is observed that the Neural Network classifier improved f measure for concatenated features averagely by 2.03% and 2.37% than KNN classifier with GSO and Firefly based feature selection respectively.

V. CONCLUSION

Tamil is one of the most ancient language in the world with rich heritage and literature. Identification of proper feature selection method is the major aspect in achieving suitable character recognition for ancient Tamil. In the proposed work, feature selection is done through Group search optimization and Firefly optimization and classifiers such as K-Nearest Neighbor (KNN), NN, J48 classifier. For feature extraction and concatenation shape and hough transformation is done. Validated results show that the outputs are compared with other methods and improvements are recorded. The results of the experiments proved that the proposed feature selection outperformed the other methods

REFERENCES

[1] Bansal, S., Garg, M., & Kumar, M. (2014). A technique for offline handwritten charachter recognition. IJCAT, 1(2), 2010-2015.

[2] Kanimozhi VM, & Muthumani I. (2014). Optical character recognition for English and Tamil script. IJCSIT, 5(2):1008–10.

[3] VenkateswaraRao, N., Srikrishna, A., RaveendraBabu, B., & Babu, R. M. (2011). An efficient feature extraction and classification of handwritten digits using neural networks. International Journal of Computer Science, Engineering and Applications, 1(5), 47-56.

[4] Simayi, W., Ibrayim, M., Tursun, D., & Hamdulla, A. (2015). Survey on the Features for Recognition of on-line Handwritten Uyghur Characters. International Journal of Signal Processing, Image Processing and Pattern Recognition, 8(9), 45-58.

[5] Singh, P., Verma, A., & Chaudhari, N. S. (2014). Devanagri Handwritten Numeral Recognition using Feature Selection Approach. International Journal of Intelligent Systems and Applications, 6(12), 40-47.

[6] RajaKumar, S., & Bharathi, V. S. (2012, April). Eighth century Tamil consonants recognition from stone inscriptions. In Recent Trends In Information Technology (ICRTIT), 2012 International Conference on (pp. 40-43). IEEE.

[7] Subashini, P., Krishnaveni, M., & Sridevi, N. (2011). Period prediction system for tamil epigraphical scripts based on support vector machine. In Information Systems for Indian Languages (pp. 23-30). Springer Berlin Heidelberg.

[8] Soumya, A., & Kumar, G. H. (2014, January). Classification of Ancient Epigraphs into different periods using Random Forests. In Signal and Image Processing (ICSIP), 2014 Fifth International Conference on (pp. 171-178). IEEE.

[9] Azmi, R., Pishgoo, B., Norozi, N., Koohzadi, M., & Baesi, F. (2010, October). A hybrid GA and SA algorithms for feature selection in recognition of hand-printed Farsi characters. In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on (Vol. 3, pp. 384-387). IEEE.

[10] Sridevi, N., & Subashini, P. (2013, March). Combining Zernike moments with regional features for classification of handwritten ancient tamil scripts using extreme learning machine. In Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on (pp. 158-162). IEEE.

[11] Farkya, S., Surampudi, G., & Kothari, A. (2015, April). Hindi speech synthesis by concatenation of recognized hand written devnagri script using support vector machines classifier. In Communications and Signal Processing (ICCSP), 2015 International Conference on (pp. 0893-0898). IEEE.

[12] Pant, N., & Bal, B. K. (2016, July). Improving Nepali OCR performance by using hybrid recognition approaches. In Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on (pp. 1-6). IEEE.

[13] Mohan, M., & Jyothi, R. L. (2015). Handwritten Character Recognition: A Comprehensive Review on Geometrical Analysis, 17(2), pp. 83-88.

[14] Gaikwad, M. R., & Pardeshi, N. G. (2016). Text Extraction and Recognition Using Median Filter, 3(1), 717-721.

[15] Ardeshana, M., Sharma, A.K., Adhyaru, D.M., & Zaveri, T.H. (2016). Handwritten Gujarati character recognition based on discrete cosine transform, pp. 23-26.

[16] El Bahi, H., Mahani, Z., & Zatni, A. An offline handwritten character recognition system for image obtained by camera phone. Mathematical Applications in Modern Science, ISBN, 978-1, pp. 180-189.

[17] Liolios, N., Kavallieratou, E., Fakotakis, N., & Kokkinakis, G. (2002). A new shape transformation approach to handwritten character recognition. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 1, pp. 584-587). IEEE.

[18] Farahmand, A., Sarrafzadeh, A., & Shanbehzadeh, J. (2013). Document image noises and removal methods, pp. 1-5.

[19] Banati, H., & Bajaj, M. (2011). Fire fly based feature selection approach. IJCSI International Journal of Computer Science Issues, 8(4), 473-480.

[20] Prasad, K., & Agrawal, S. (2015). Character Recognition Using Neural Network. Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March, 90-92.

[21] Tadesse, S. (2011). Feature Extraction and Classification Schemes for Enhancing Amharic Braille Recognition System (Doctoral dissertation, aau), pp. 1-126.

[22] Renjini, L., & Rubeena, B. A Comprehensive Study On Handwritten Character RecognitionSystem. IOSR Journals (IOSR Journal of Computer Engineering), 1(17), 1-7.

[23] Abandah, G. A., & Malas, T. M. (2011). Feature selection for recognizing handwritten Arabic letters. Dirasat: Engineering Sciences, 37(2), 1-21.

[24] Assabie, Y., & Bigun, J. (2011). Offline handwritten Amharic word recognition. Pattern Recognition Letters, 32(8), 1089-1099.