International Journal of Research in Advent Technology (IJRAT), Special Issue, March 2019
E-ISSN: 2321-9637
ICCCMIT 2019 organised by M.O.P. Vaishnav College for Women (Autonomous)
Chennai-34, India
Available online at www.ijrat.org

A Hybrid RBF Network with GA for Systematic Classification of Historical Tamil Palm Leaf Manuscript

Poornima Devi. M[1], Sornam. M[2]
Department of Computer Science [1, 2], University of Madras, Chennai.[1, 2]
Email:poornima160492@gmail.com[1],madasamy.sornam@gmail.com[2]

Abstract-Tamil character classification plays an important role in present scenario due to number of characters in Tamil language (247 characters), cursive, structural forms of characters and different styles of characters (ancient and modern Tamil characters). Classification of characters was still scientifically challengeable task especially for Tamil palm leaf manuscript because which was written in dried palm leaf and style of writing will differ from person to person. So the digitization is must in the field of palm leaf manuscript to well know the ancient secrets. This article proposed the method to classify characters which was segmented from Tamil palm leaf manuscript. The combination of Radial Basis Function (RBF) and Genetic Algorithm (GA) is used for classification of characters. Nguyen-Widrow weight initialization technique is used to initialize the weights in GA. This system performed better and achieved 97.7% of classification accuracy with the performance analysis: precision, recall and F1-score.
Index Terms-RBF; Nguyen-Widrow; Genetic Algorithm; Tamil palm leaf manuscript.

1. INTRODUCTION

Machine Learning (ML) and Deep Learning (DL) is a fastest growing research area in Artificial Intelligence which enables to work like human being with best accuracy and computational time. Some of the research work in the field of AI is character recognition, character classification, sentiment analysis, weather prediction, face recognition, speech recognition etc. Character recognition and classification plays a vital role, because it preserves the content from different documents in different languages. So far many researches has been done in the field of classification and recognition characters but still need more research in Tamil characters due to century or generation based characters.

Here the dataset considered is Tamil palm leaf manuscript. Mostly this type of manuscript was written by our ancestors in dried palm leaf to provide the information for future generation. But now, the modern world is unable to read the Tamil characters especially ancient characters. So to know the traditional and historical culture of Tamil people, digitalization is required for Tamil palm leaf manuscript. The sample of Tamil palm leaf manuscript was shown in Fig.1.

The major issues of palm leaf manuscript are:

- Most of the palm leaves are written 200 to 300 years ago.
- Degraded manuscripts due to natural disasters, lack of maintenance etc.
- Ancient characters in the manuscript.
- Writing style of individuals.

Tamil language is one of the oldest languages in the world, which is also an official language in Tamil Nadu and also classical language in India and world. It is spoken by Tamil people in all over the world. Tamil is declared as an official language of foreign countries like Sri Lanka and Singapore. It is also spoken by the countries like Malaysia, South Africa and Mauritius. The country following Tamil culture and using Tamil Language but not officially declared is Reunion and Seychelles in Africa. The countries like Malaysia and Myanmar offers Tamil language in schools. The month January was announced as "Tamil Heritage Month" in Canada. Tamil language contains 247 characters which was classified as Uyirezhuthu (12 characters), Meyezhuthu (18 characters), Uyirmeyezhuthu (12 Uyirezhuthu * 18 Meyezhuthu = 216 characters) and ayudhaezhuthu (1 character). It can also divide into Vallinam, Mellinam and Idayinam as shown in Fig.2.

Fig. 1. Tamil palm leaf manuscript collected from Tamil Nadu Archeological Department, Chennai.

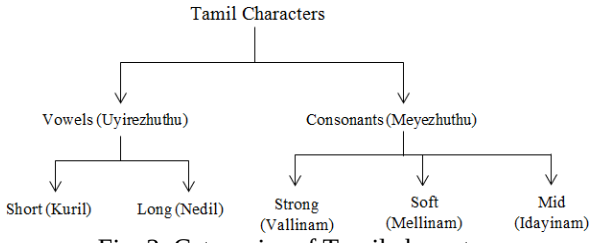

Fig. 2. Categories of Tamil characters.

Tamil language can be categorized as old Tamil (300 BCE – 700 CE), Medium Tamil (700 CE – 1200 CE) and Modern Tamil (1200 – Present). Modern Tamil characters are currently in use, which

will slightly differ from old and Medium Tamil and also difficult for today's generation to recognize the ancient characters. The illustrations for some characters are shown in Fig.3.

Fig. 3. Comparison of Tamil characters with respect to generation.

This article is organized as follows: Section 2 contains related literature review, Section 3 contains proposed work, Section 4 contains observed outcome and examination and Section 5 contains conclusion.

2. RELATED LITERATURE REVIEW

Ramanan et al. [13] proposed a novel approach to recognize Tamil characters using binary Support Vector Machine (SVM) for multiclass classification with Decision Tree. Here Binary rooted Directed Acyclic Graph (DAG) decision was used for Unbalanced Decision Trees (UDT). The implementation is based on One Versus One (OVO) with SVM and One Versus All (OVA) with SVM. The Histogram of Oriented Gradients (HOG) and transition was used for feature extraction. The dataset contains 12400 samples and achieved 98.80% of recognition accuracy.

Aladhahalli et al. [3] proposed a method to segment the characters from degraded historical documents using Watershed model. Sobel and Laplacian are used to eliminate the noise from the document image. Different datasets were used to test the model, where the proposed methods showed better accuracy in terms of precision, recall and f-measure.

Aarthi et al. [1] developed the application for people to help them technologically which converts text to speech in English and Tamil languages. In this system, Optical Character Recognition (OCR) has been used to convert the text present in image. Then the text obtained from OCR was converted to audio with generated voice. Here Gabor filter algorithm was used for feature extraction. For classification, SVM was implemented.

Dhole and Kale [5] developed the system to analyze the stress level of the human being by using speech signal. RBF network was used as a classifier and Mel Frequency Cepstral Coefficient was used as a feature extraction technique. The developed system is independent on real time speech, language and word. This work proposed to identify the stress with six basic emotions such as happiness, sadness, fear, anger, disgust and surprise.

Chaouki et al. [4] proposed supervised learning of Artificial Immune system to recognize the Arabic characters. This system includes preprocessing, feature extraction and recognition. The ten-fold cross-validation technique was used to train and test the system with IFN/ENIT dataset. The parameters used in this system were tuned using leave-one-out cross-validation with grid search algorithm which obtained 93.25% of accuracy.

Karun and Sharma [7] developed the method to recognize the online handwritten Gurumukhi characters. In this article, the writing zone identification algorithm was developed to identify upper, middle and lower zones of characters. The strokes of characters were recognized using Hidden Markov Model with 95.3% of accuracy and the zone identification achieved 88.4% of accuracy. The dataset contains 4280 characters of different users.

Shruti et al. [15] described the method to identify the defective region of cerebrum using neural and fuzzy classifiers. Here Dark Level Co-event Matrix strategy was used to extract the features from dim scale picture and GA was used to improve the system with alternative lessening measure.

Santhosh et al. [14] developed the system to classify and recognize the Tamil handwritten characters with two approaches such as Self Organizing Map (SOM) and RBF with SOM. The combination of RBF with SOM gives the better result with classification accuracy 89% and 96.9% of recognition accuracy.

Nilesh et al. [12] proposed a work for classification of brain tumor from MRI. To improve the tumor detection performance different segmentation methods were implemented such as Fuzzy Classifier Means (FCM), Discrete Cosine Transform (DCT), Watershed and Berkeley Wavelet Transform (BWT) to pick the best segmentation method. The classification with GA yields 92.03% of accuracy using BWT based segmentation for MR images.

Agnel and Merlin [2] developed a system to recognize Tamil script using HMM with fuzzy logic classifier. The unknown characters were identified using frames in feature distance and member function. It recognizes characters with accuracy vary from 89% to 93% with 750 characters.

Munish et al. [11] presented various feature extraction technique for offline handwritten Gurumukhi characters. The feature extraction technique includes Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi Squared Attribute (CSA), Independent Component Analysis (ICA), Latent Semantic Analysis (LSA), Principal Component Analysis (PCA) and Random Projection.

The classifiers used were Nearest Neighbors (NN) and SVM. In this work CSA feature extraction works better and classified the characters based on upper, middle and lower zone with accuracy of 88.3%, 95.2% and 91.3%.

Mukesh and Suman [10] presented the method to extract and identify the high-level strokes in the Gujarati Printed document characters. The dataset includes machine printed books, newspapers and laser-printed documents. The classifiers used was Naive Bayes and HMM. The Hidden Markov Model produced 93% to 96% of accuracy for recognizing Gujarati characters.

Lahcen et al. [9] proposed a method to recognize the Tifinagh handwritten characters. This system includes image scanning, preprocessing, feature extraction and character recognition. Feedforward neural network was used for classification with GA and the system for OCR produced 89.5% of accuracy.

3. PROPOSED WORK

This work is proposed to classify the palm leaf characters which using hybrid Radial Basis Function (RBF) network with Genetic Algorithm (GA) in which the characters were segmented from Tamil palm leaf manuscript. GA is used to generate the optimized weights which were fed into RBF network. Normally weights were generated randomly, rather than random initialization Nguyen-Widrow weight initialization technique [6] was implemented to generate the population (weights) for GA. The flow of the proposed work was shown in Fig.4.

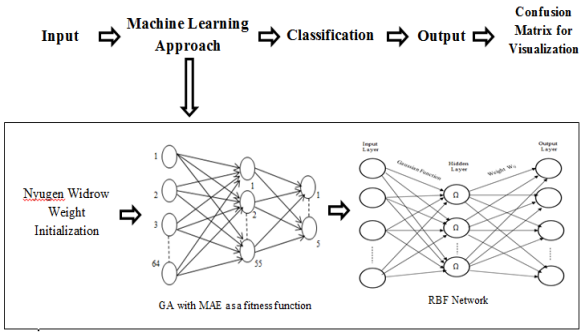

Fig. 4. Flow of the proposed work.

3.1. Radial Basis Function (RBF) Network

Radial Basis Function Network is one of the feedforward neural networks, which contains three layers: input layer, hidden layer and output layer. Input layer to hidden layer transformation is represented as non-linear and hidden layer to output layer transformation is represented as linear. Here the network architecture contains 64 input neurons, 55 hidden neurons and 5 output neurons. Gaussian function is used to find the hidden layer response and gradient descent learning technique is used to adjust the weights and centers.

In RBF, weights were used from hidden layer to output layer. Mostly these weights were initialized randomly; instead of using random weights, GA was implemented to generate the weights. In GA, weights were initialized using Nguyen-Widrow weight initialization technique refer Fig.5.

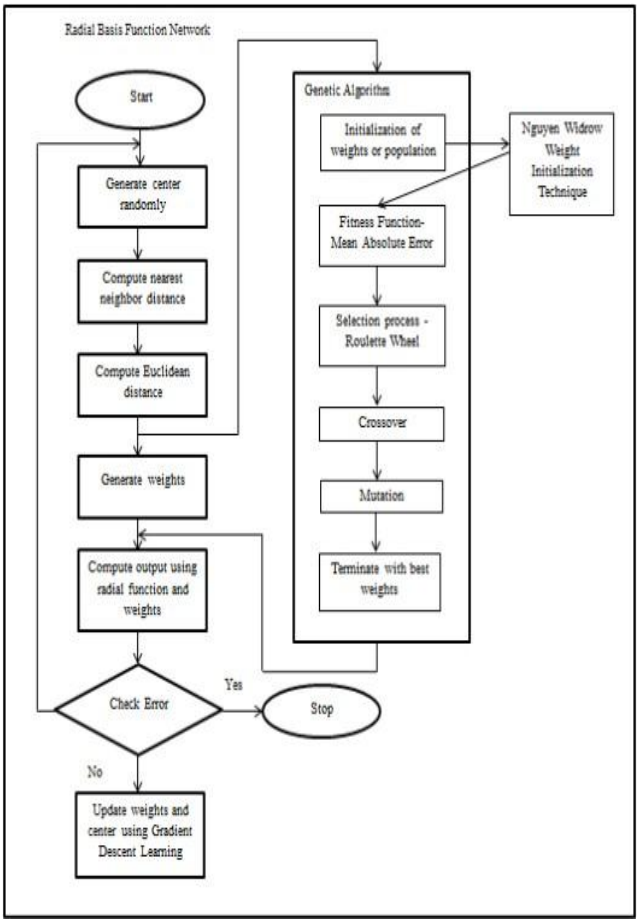

Fig. 5. Flow of algorithm.

3.2. Genetic Algorithm (GA)

Genetic algorithm is one of the optimization techniques. GA starts with initialization of weights or population. Then parents were selected for mating and then crossover and mutation were applied to compute the new weights. Finally the generated new weights were replaced to the existing weights refer Fig.5.

For this proposed work, Nguyen-Widrow weight initialization is used to initialize the weights. Fitness function used is Mean Absolute Error (MAE). To select the parent weights Roulette wheel selection is used. One point crossover and uniform mutation was implemented to generate the new off spring. Table 1 shows the parameters and their corresponding values used for this proposed system.

Table 1. Parameters and their corresponding values of the proposed system

Parameters		Values	
No. of input neurons		64	
No. of hidden neurons		55	
No. of output neurons		5	
Nguyen-Widrow Weight Initialization	β	0.7	
	Initialization	range between -0.5 to 0.5	
Genetic Algorithm	Crossover rate	0.9	
	Mutation rate	0.5	
	Fitness function	MAE	
RBF Network		Technique	Gradient descent learning
		MSE	<0.0003

4. OBSERVED OUTCOME AND EXAMINATION

The proposed system was implemented using 64 bit Windows 7 OS in Anaconda Navigator Python 3.7 (Spyder) with Intel Pentium Processor CPU @2127U and 4GB RAM.

Tamil palm leaf dataset used is lagnam scripts written by SubtharishiNaash [16] which were collected from Tamil Nadu Archaeological Department, Chennai, India. Here the preprocessed and segmented palm leaf manuscript characters were used to develop the network. 130 samples of Tamil palm leaf segmented characters were used. Thesample of segmented characters with class labels was shown in Fig.6. From this 91 characters were used to train the network and 39 characters were used for test the network.

The overall classification accuracy with precision, recall and f1 score with corresponding class labels was reported in Table 2. The confusion matrix was drawn for the proposed hybrid RBF with GA for better understanding of classification as shown in Fig .6.

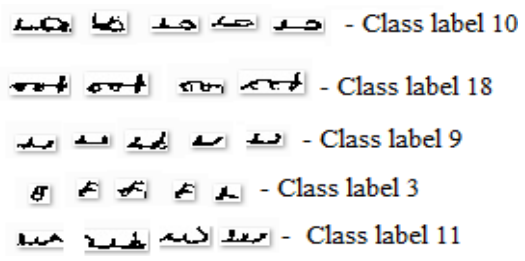

Fig. 6. Sample of segmented characters with class labels.

Table 2. Evaluation of propose work

Class labels	Precision	Recall	F1-score	No. of Characters
1	1.00	0.93	0.97	15
2	0.50	1.00	0.67	1
3	0.92	1.00	0.96	12
4	1.00	1.00	1.00	1
5	1.00	1.00	1.00	6
6	1.00	1.00	1.00	1
7	1.00	1.00	1.00	1
8	0.00	0.00	0.00	1
9	0.96	1.00	0.98	24
10	1.00	1.00	1.00	24
11	1.00	1.00	1.00	10
12	1.00	1.00	1.00	2
13	1.00	1.00	1.00	2
14	1.00	0.92	0.96	13
15	1.00	1.00	1.00	1
16	1.00	1.00	1.00	9
17	1.00	1.00	1.00	1
18	1.00	1.00	1.00	7
Average/ Total	0.97	0.98	0.97	130
Classification Accuracy	97.7 %			

The confusion matrix illustrates that 3 characters were shown in Fig.7 misclassified from the total of 130 characters. Class label 0 has totally 15 characters from this 1 character was misclassified as class label 2. Then class label 7 has only one character, which was misclassified as class label 1. Finally, class label 13 has totally 12 characters from which one character was misclassified as class label 8.

The comparison of existing work with the proposed RBF with GA network was shown in Table 3.

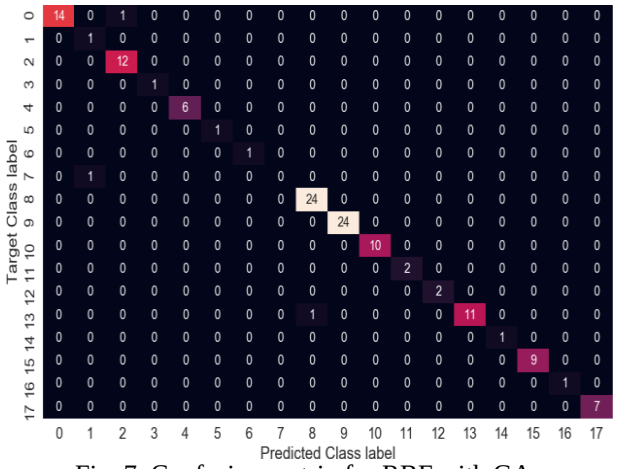

Fig. 7. Confusion matrix for RBF with GA.

Table 3. Comparison of existing work with proposed work

Article and Year	Area/ Network	Script type	Accuracy
Kiruba et al. [8], 2017	Image Processing	Tamil palm leaf script	87%
Agnel and Merlin [2], 2017	HMM and Fuzzy logic	Tamil Handwritten document	89% - 93%
Mukesh and Suman [10], 2018	HMM	Gujarati printed characters	93% – 96%
Proposed	Hybrid RBF with GA	Tamil palm leaf manuscript	97.7%

5. CONCLUSION

Tamil palm leaf manuscript character classification is proposed in this article using the combination of hybrid Radial Basis Function (RBF) with Genetic Algorithm (GA). RBF network is executed to classify the characters and in RBF network, to get the optimized weights GA was implemented. In GA, to initialize the weights Nguyen-Widrow technique was applied. The combination of hybrid RBF with GA performed better and produced 97.7% of accuracy with 0.0003 MSE.

REFERENCES

[1] Aarthi, R; Kruthika Devi, N. R.; Prithi, S.(2018): Advanced character based recognition and phone handling for Blind peoples, International Research Journal of Engineering and Technology (IRJET), Vol. 05, Issue. 03, pp. 1751– 1755.

[2] Agnel Livingston, L. G. X.; Merlin Livingston, L. M.(2017): Handwritten recognition of Tamil scripts using HMM with fuzzy logic, International Journal of Scientific Research (IJSR), Vol.6, Issue.4, pp.738–740.

[3] AladhahalliShivegowdaKavitha;PalaiahnakoteShivakumar; GovindarajHemantha Kumar; Tong Lu.(2017): A new watershed based system for character segmentation in degraded text lines, International Journal of Electronics and Communications (AEU) 71, pp.45–52.

[4] Chaouki Boufenar; Mohamed Batouche; Marc Schoenauer. (2018): An artificial immune system for offline isolated handwriten arabic character recognition, Evolving Systems 9, pp.25–41.

[5] Dhole,N. P.; Kale,S. N.(2018): An application using Radial Basis Function classification in Stress speech identification, International Journal of Computer Science and Engineering, Vol.6, Issue.4, pp.448–451.

[6] Dino Neinhold; Kilian Schwab; Rolf Dornberger; Thomas Hanne.(2015): Effects of weight initialization in a Feedforward Neural Network for classification using modified Genetic Algorithm, IEEE, 3rd International Symposium on Computational and Business Intelligence, pp.6– 12.

[7] Karun Verma; Sharma,R. K.(2018): An efficient writing-zone identification technique for Online Handwritten Gurumukhi character recognition, Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. 88 (2), pp.297–307.

[8] Kiruba, B.; Nivethitha, A.; Vimaladevi, M.(2017): Segmentation of handwritten Tamil characters from palm script using Histogram approach, International Journal of Informative and Futuristic Research, Vol. 4, Issue. 5, pp.6418–6424.

[9] Lahecen Niharmine; Benaceur Outtaj; Ahmed Azouaoui.(2018): Tifinagh handwritten character recognition using Genetic Algorithm, IEEE, International Conference on Advanced Communication Technologies and Networking (CommNet), DOI: http://10.1109/COMMNET.2018.8360267

[10] Mukesh, M.; Goswami; Mitra, Suman K..(2018):Printed Gujarati character classification using High-level strokes, Proceedings of 2nd International Conference on Computing Vision and Image Processing, Advances in Intelligent Systems and Computing 704, pp. 197–209.

[11] Munish, Kumar; Jindal, M. K.; Sharma,R. K.; Simpel, Rani Jindal.(2018): Performance comparison of several feature selection techniques for offline handwritten character recognition, IEEE, 3rd IEEE International Conference on Research in Intelligent and Computing in Engineering, pp.1–6.

[12] Nilesh Bhaskarrao; Arun Kumar Ray;Har Pal Thethi.(2018): Comparative approach of MRI-Based Brain Tumor segmentation and classification using Genetic Algorithm, Journal of Digital Imaging

31, 2018, pp.477–489.

[13] Ramanan,M.; Ramanan, A.; CharlesE. Y. A.(2015): A Hybrid Decision tree for printed Tamil character recognition using SVMs, IEEE, 2015 International Conference on Advances in ICT for Emerging Regions, pp.176–181.

[14] Santhosh Baboo, S.; Subashini, P.; Krishnaveni,M.(2009): Combining Self-Organizing Maps and Radial Basis Function networks for Tamil handwritten character recognition, ICGST-GVIP Journal, Vol. 9, Issue.4, pp.1–7.

[15] Shruti Bhargava; SPV SubbaRao; Abhishek Choubey.(2018): Categorization of MRI Brain Images using Genetic Algorithm, International Journal of Advance and Innovative Research, Vol.5, Issue.1(IV), pp.60–64.

[16] Sornam, M.;Poornima Devi, M.(2018): Tamil Palm leaf manuscript character segmentation using GLCM feature extraction, International Journal of Computer Science and Engineering, Vol.6, Issue.6, pp.167–173.