Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 1439–1447
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5790

Deep learning based hand written character recognition for manuscript documents

T. Jerry Alexander[a,*], S. Suresh Kumar[b], N.R. Krishnamoorthy[c]

[a]*Faculty of Electronics Engineering, Sathyabama Institute of Science & Technology, IT Highway, Chennai, India*
[b]*Swarnandhra College of Engineering & Technology, Narasapur, India*
[c]*School of Electrical & Electronics, Sathyabama Institute of Science & Technology, Chennai, India*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Handwritten manuscripts contain much ancient information related to astrology, medicines, grammar etc. They are of various forms such as palm leaves, paper, stones etc. These manuscripts are preserved by the method of digitization with noise introduced. By using proper filtering as well as denoising methods these noises are eliminated and the images are restored. It is finally required to recognize the handwritten characters automatically from the restored image enabling the researchers and enthusiasts for going through the document very easily. This proposed work deals with the creation of a handwritten characters dataset for all the characters within a specific dimensional area and the recognition of handwritten characters using the deep learning method. First, the handwritten dataset is created from different human handwritings in a specific format, scanned and each character with suitable dimension is obtained by labelling them as per the sequence. Then various forms of convolution network are applied for the character recognition and the results are compared to obtain the suitable net for the Tamil character recognition from the handwritten document.

Keywords: Character recognition, Convolution Neural Network, Historical Manuscripts; handwritten character recognition, dataset creation, tamil character recognition, deep learning.
2010 MSC: Please write mathematics subject classification of your paper here.

*Corresponding author
Email addresses: tjerryalexander@gmail.com (T. Jerry Alexander), nice.ssk@gmail.com (S. Suresh Kumar), moorthy26.82@gmail.com (N.R. Krishnamoorthy)

Received: August 2021 *Accepted:* November 2021

1. Introduction

From ancient period onwards it is a usual practice for recording all the important information. Majority of the tasks carried out are handwritten as it can be done easily without any difficulty compared to the data stored in a digital format like word document, image etc. The handwritten documents range from palm leaf manuscript in the ancient period to the paper document in this present period. These documents are stored and retrieved whenever required. From this document the character recognition still remains a challenging task. Character recognition is one of the important application in the image processing field and still researches are going on to attain the best algorithm or network to attain the maximum accuracy. Many techniques have been employed. This proposed work performs one of the basic steps in the hand written character recognition of the palm leaf manuscript [23] or paper document [13]. As an initial step, the tamil handwritten character database is created where a certain portion of the image is allocated for testing purpose and the remaining portion for the purpose of training.

2. Related Works

Literature speaks a lot of research work related to the hand written character recognition from the traditional method to the advanced method, but till now not even a method is applicable for classifying all the different aspects more accurately. Again it depends on the dataset and its type also. Hence there must be a compromise among what type of dataset to be used and on the testing, training and development network.

The dataset for this work is obtained from the University of Jaffna [7], also, prepared from obtaining the various handwritings manually with original and reduced image sizes and the printed characters from the web resources. There are dataset available in languages like devanagiri [1, 9], Arabic [25], Hiragana [14], Chineese [5], tamil [8, 24], other indic scripts [22] etc. This proposed work uses the tamil dataset for the character recognition purpose.

In this proposed work, the concept of deep learning is used for classification and recognition of the tamil characters. Normally deep learning based works finds it application in the areas of computer vision based image processing. They are used widely in the biomedical field for classification of medical dataset [16, 2, 21] and other general purpose applications [15].

Normally the dataset are created [23] by various methods, preprocessed [19] before its utilization. Then they are trained, developed and tested using methods employed [18, 10, 6] different convolution neural networks [20] like LeNet, AlexNet, ResNet, GoogleNet etc and finally to recognize the characters.

The preparation of the data set is obtained by taking the all set of tamil language characters from 100 different human handwriting. The wrongly or improper characters were deleted and due to this the database can't have equal no of characters. The dataset taken here were allotted 70% for training and 30% for testing. Again from internet sources, few printed dataset, handwritten dataset, were taken for internet printed data, hand written characters as large and small.

3. CNN Networks

The character recognition falls under the class of object recognition and this problem still exists a prolonged time, many robust models has to be built for maintaining the accuracy and reliability. The CNN network finds its application in the image processing particularly for image classification. There were various CNN networks [3, 17] like GoogleNet, AlexNet, ResNet, VGGNet since the inception of LeNet from the year 1988. The AlexNet came into existence from 2012, gave a good

promising results for classification and recognition of images.
The networks emerged from shallow at the inception perod to the deep network at recent years. Followed by the AlexNet were the GoogleNet, ResNet which had some additional computational features and better performance than AlexNet. The CNN architecture, generally comprises of three major groups namely the input layer, feature extraction or learning layer and finally the classification layer. The input layer is the one where the raw input image data of suitable size (width and height) along with the depth (normally 3 for RGB images) is loaded as input for the further processing by the next stage of convolution layer where learning or feature extraction takes place. The final stage of the network performs the image classification based on the features learned by the second stage. Though there are many networks available, it is concluded that each one is having its own advantages and limitations with respect to the dtaset size, speed of operation and its network complexity. Hence there should be made a compromise based on the selection of these network models.

3.1. AlexNet

AlexNet [11, 12] is an interesting architecture algorithm mainly used in the computer vision and image classification process. It has a combination of 5 convolution layers each having number of filters with varying sizes 11×11, 5×5 and 3×3 and 3 fully connected (FC) layers totaling to a 8 layers. AlexNet has 10 multiple times of convolution layers more than the LeNet architecture. Here the sigmoid activation function is converted to a simple Rectified Linear Unit(ReLU) function. It is one of the deep Convolutional Neural Network.

3.2. ResNet

It is one of the best networks introduced by Microsoft in the year 2015 having the deep learning features. ResNet [10] performs the residual learning and hence called as Residual Network. This network solves the problem of degrading accuracy and the training made under this network is simpler compared to other deep learning network. The ResNet34 is the basic network while ResNet50, ResNet101 and ResNet152 remains the other variants of the ResNet network. The suffix numbering represents the no of learning layers for the residual representation. In this proposed work the ResNet50 and ResNet101 are utilized for the comparison. The ResNet50 is a 50 layered network as derived from its name suffix numerals while the ResNet101 network has 101 learning layers for its residual representation.

4. Proposed Methodology

The proposed methodology involves various stages where the first stage is the preparation of the unique dataset obtained from the different handwritings in a specific format, scanned and then preprocessed. The second stage involves the training stage trained with AlexNet(), ResNet50() and ResNet101(). The final stage was the testing stage where the unique dataset was tested with the same dataset of different image sizes, dataset available from the repositories and images obtained from the internet.

A. Preparing Datasets The datasets are prepared by obtaining the total tamil characters written on a single page marked as shown in Fig.1 with an equally spaced square dimension drawn with the help of pencil and then the horizontal and vertical lines are erased after obtaining handwritten characters. Similarly, the character sample images are obtained from hundred different people for creating the dataset. The dataset are scanned with the scanner or a good mobile phone and images are stored in a folder with proper labeling for each character and hence the dataset folder is ready for its usage. The Fig.2 shows few samples of the obtained

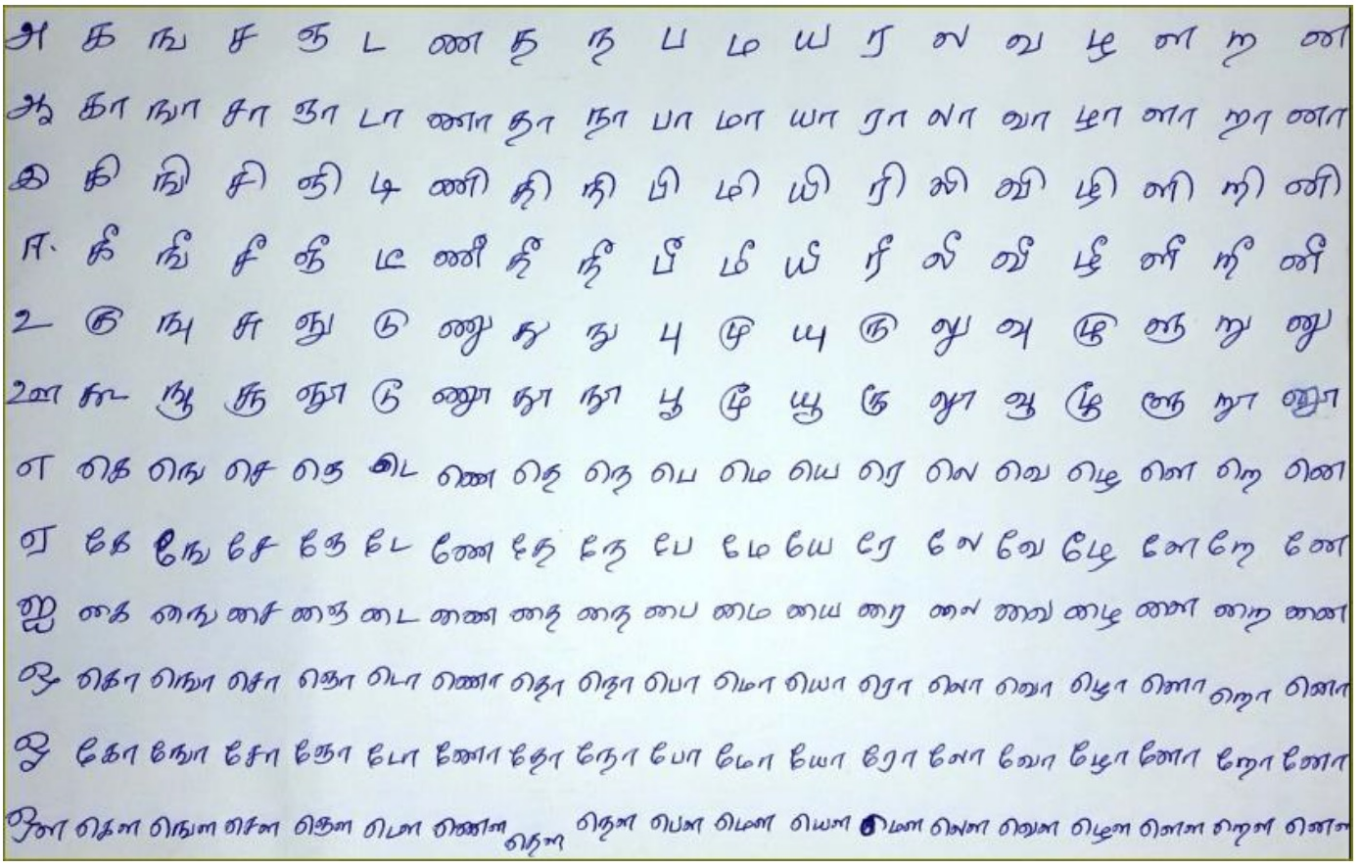

Figure 1: A sample original image obtained for dataset creation

characters from the Fig.1. The processing method for obtaining the dataset is given below with the following steps.

Algorithm

1. Give the scanned input image of handwritten characters written in an equally dimensioned square box that is erased.
2. Convert the input image to grayscale image.
3. Compute the length of the characters in both the row and column wise.
4. Obtain each resized characters and store this in the corresponding labeled file
5. Similarly repeat this for the remaining images containing all the tamil handwritten characters.

Hence the dataset is created using the above steps.

B. Training and Testing Datasets Using Three Networks and Its Comparison The algorithm for training and testing the handwritten characters using ResNet50, ResNet101, AlexNet is given below where the dataset created is tested with the same dataset of different image sizes and from the web repositories.

ALGORITHM

1. Read the labeled characters from the dataset folder.

Figure 2: Sample images obtained for the dataset creation from Fig.1

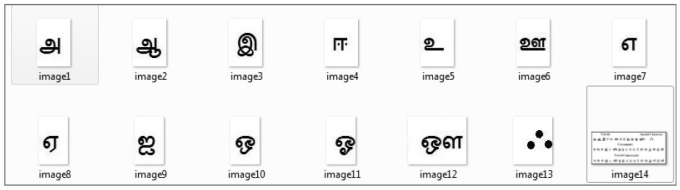

Figure 3: Printed dataset images obtained from the web resources for testing purpose.

2. Count the labels and split the labels containing the characters.
3. Obtain the characters from each label.
4. Set the proper type of network for both training and testing.
5. Set the layers.
6. Split the images with proper ratio for testing and training purpose.
7. Train the network using the network like AlexNet, ResNet50, ResNet101 etc from step (iv)
8. Test the network by giving any character as input.
9. Obtain the confusion matrix, each column (across the top) to the actual characters to the each row (downside) for the predicted characters.
10. If both the expected and predicted are same, then a maximum value of 1 would be obtained to show that it is 100% correct in the prediction.
11. Display the results and the confusion matrix.

5. Result and Discussion

The above fig. 4 details some of the sample tamil characters and their corresponding labeling numbers. In a similar fashion the labeling is done for all the total 247 tamil characters where there are 12 vowels, 1 special characters, 18 consonants and 216 compound characters formed by the combination of vowels and consonants. The created dataset size will account to the product of charcters

1	2	3	4	5	6	7
அ	ஆ	இ	ஈ	உ	ஊ	எ
8	9	10	11	12	13	
ஏ	ஐ	ஒ	ஓ	ஔ	ஃ	

Figure 4: List of sample characters and their corresponding labeling.

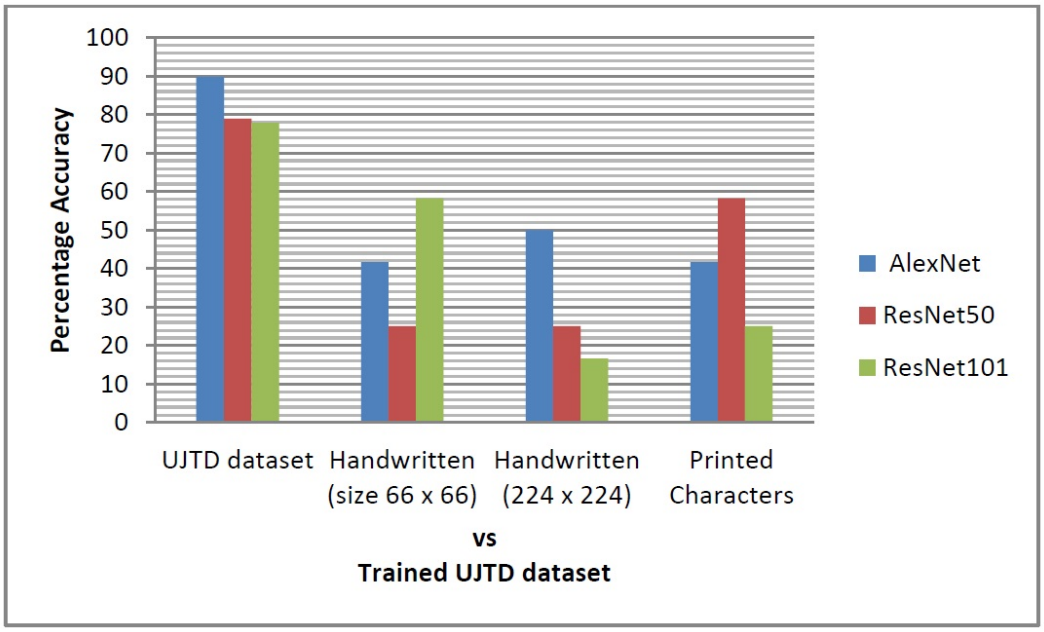

Figure 5: Accuracy graph using trained UJTD dataset vs. tested UJTD dataset, Handwritten character images of two different sizes and printed character dataset.

and the 100 different characters leading to 24700 images.
For simulation, four different dataset are used, first dataset is downloaded from the university web resource [7] and labeled as UJTD character. The second and third dataset are unique handwritten dataset collected from different persons. To study the effect of resolution of handwritten images, the second dataset is obtained by scanning and cropping the image into 224 x 224 pixel as the pre-trained network accepts this format. Also the same handwritten images are scanned and cropped into the image of resolution 66 x 66 pixels. The purpose of this dataset is to save the memory occupied during the training phase of the network.
Finally the fourth dataset is a printed one, downloaded from the internet source. This dataset is used only for the validation. The first three dataset are trained using ResNet50, ResNet101, AlexNet and the validation was done with the three dataset allocated for testing along and also with the printed characters. Fig. 5, 6, 7 shows the accuracy graph obtained as a result of training using the three networks. The fig.5 depicts the accuracy graph obtained using trained UJTD dataset vs. tested UJTD dataset, handwritten character images of two different sizes and printed character dataset. Fig, 6 shows the graph between trained handwritten character dataset size (66 x 66) pixel vs. tested UJTD dataset, handwritten character images of two different sizes and printed character dataset while fig. 7 exhibits the accuracy graph using trained handwritten character dataset size (224 x 224) pixel vs. tested UJTD dataset, Handwritten character images of two different sizes and printed character dataset

Confusion Matrix:
The purpose of confusion matrix is to show the response summary of the classification method implemented. Here the classification method is confused while making the predictions. The confusion matrix is used to throw light not only on the classifier error but also used in the classifications of error types. There methods to calculate the confusion matrix are depicted here [4]

From the above three graphs as shown in fig.5, fig.6 and fig.7 respectively, it is obtained that the

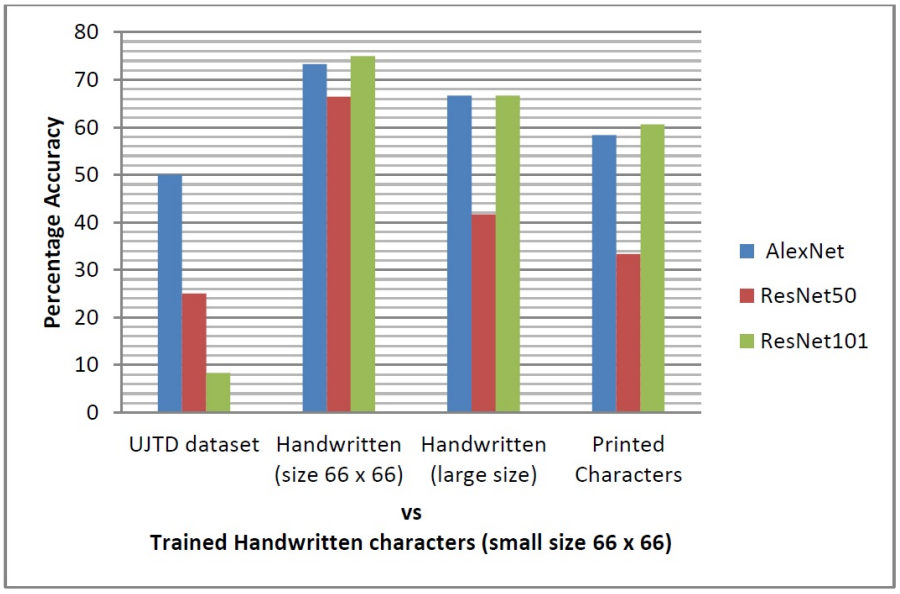

Figure 6: Accuracy graph using trained handwritten character dataset size (66 x 66) vs. tested UJTD dataset, Handwritten character images of two different sizes and printed character dataset.

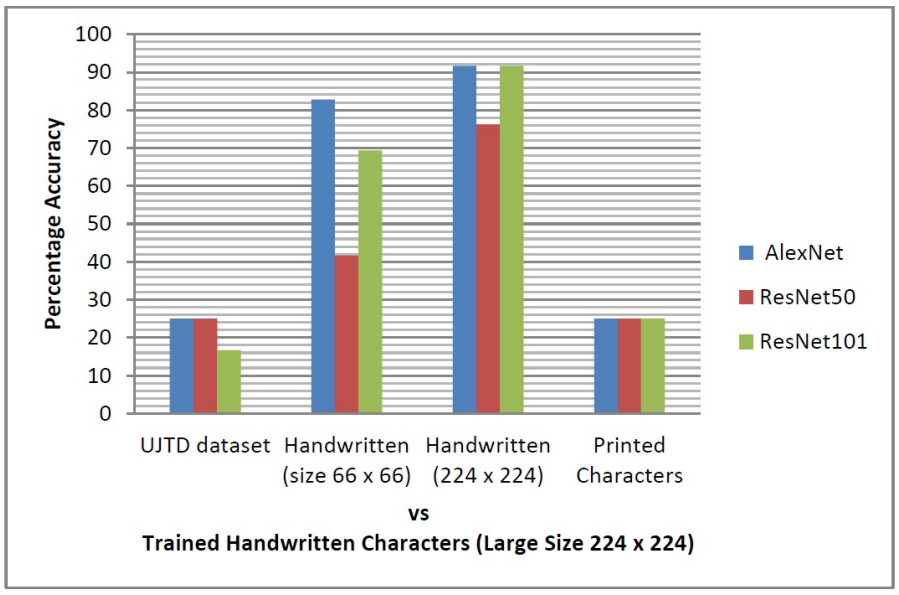

Figure 7: Accuracy graph using trained handwritten character dataset size (224 x 224) vs. tested UJTD dataset, Handwritten character images of two different sizes and printed character dataset.

Figure 8: Few samples of a typical character written in different styles from the dataset

AlexNet outperformed the other two network variants with respect to the different dataset trained and tested. Here each character label has 100 samples of which 70% were allotted to training and 30% for testing purpose. The accuracy can be increased by adding more samples to the database.

The metrics term used here are classification accuracy, classification error or misclassification rate or error rate which can be found from the total and correct predictions obtained. The accuracy rate depends not only on the number of classes but also with size of the classes varying unequally. The accuracy rate depends on the type of classes of data which is given for testing and training purpose as shown in fig 8. Normally the unequal sized classes would show with less accuracy compared to the equally sized classes. Here in this proposed work, each class of characters is unequal and varying in the quantities. This may fluctuate in the accuracy to show less accurate results.

6. Conclusion and Future Expansion

Hence the proposed work paved the way as the initial step in recognizing the correct training network for the recognition of tamil handwritten characters. Thus the dataset has been prepared and created using the algorithm provided in the dataset creation section of this work and they are trained and tested along with the created dataset of small and big size along with the existing dataset and the characters available from the web resources. This proposed method is applied for the character recognition of the historical palm leaf manuscript, paper documents etc to obtain the results in an accurate manner. The various training network was used and it was found that from among the three networks used the AlexNet performed well for the handwritten characters. The accuracy can be increased by allocating more samples to the dataset allocated for training purpose. This proposed work can be implemented with palm leaf dataset images as well as other handwritten document manuscripts for the purpose of character recognition.

References

[1] N. Aneja and S. Aneja, *Transfer learning using CNN for handwritten devanagari character recognition*, 1st Int. Conf. Adv. Inf. Technol. Chikmagalur, India, (2019) 293–296.

[2] S. Anto, *Supervised machine learning approaches for medical data set classification-A review*, IJCST, 2(4) (2011) 234–240.

[3] A. Anwar, *Difference between AlexNet, VGGNet, ResNet and Inception*, Medium, (2020).

[4] J. Brownlee, *What is a confusion matrix in machine learning*, Machine Learning Mastery, (2016).

[5] D. Ciresan and U. Meier, *Multi-column deep neural networks for offline handwritten Chinese character classification*, International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, (2015) 1–6.

[6] P. Chansung, *CIFAR-10 image classification in TensorFlow*, Medium, (2018).

[7] *Dataset – department of computer science*, http://www.csc.jfn.ac.lk/index.php/dataset/, 2020.

[8] *Data sets of printed Tamil characters and printed documents*, University of Jaffna, http://www.jfn.ac.lk/index.php/data-sets-printed-tamil-characters-printed-documents/, Apr. 19, 2016.

[9] S.P. Deore and A. Pravin, *Devanagari handwritten character recognition using fine-tuned deep convolutional neural network on trivial dataset*, Sādhanā, 45(1) 243.

[10] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, (2016) 770–778.

[11] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, University of Toronto, 2009.
[12] A. Krizhevsky, I. Sutskever and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*, Commun. ACM 60(6) (2017) 84–90.
[13] L. Likforman-Sulem and E. Kavallieratou, *Document image processing*, J. Imag. 4(7) (2018) 84.
[14] A.H. Mawaddah, C.A. Sari, D.R.I. Moses Setiadi and E.H. Rachmawanto, *Handwriting recognition of hiragana characters using convolutional neural network*, International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, (2020) 79–82.
[15] M.A. Muthiah, E. Logashanmugam, N.M. Nandhitha, C.K. Kumar and D. Hariteja, *Performance evaluation of conventional CNN architectures and modified ALEXNET for the classification of potatoes by thermal imaging*, Russ. J. Nondestruct Test 56(9) (2020) 718–726.
[16] K. Muthuvel and S.L. Padma, *Hybrid features and classifier for classification of ECG signal*, RJASET, 9(12) (2015) 1034–1050.
[17] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*, O'Reilly Media, Inc., 2017.
[18] M.A. Pragathi, K. Priyadarshini, S. Saveetha, A.S. Banu and K.O.M. Aarif, *Handwritten Tamil character recognition using deep learning*, International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, (2019) 1–5.
[19] M. Ramanan, A. Ramanan and E.Y.A. Charles, *A preprocessing method for printed Tamil documents: Skew correction and textual classification*, IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Abbassia, Egypt, (2015) 495–500.
[20] J. Schmidhuber, *Deep learning in neural networks: An overview*, Neural Networks 61 (2015) 85–117.
[21] B. Sen and V. Vedanarayanan, *Efficient classification of breast lesion based on deep learning technique*, BIJAIP 6(1) (2016) 01–06.
[22] R. Sharma and B. Kaushik, *Offline recognition of handwritten Indic scripts: A state-of-the-art survey and future perspectives*, Comput. Sci. Rev. 38 (2020) 100302.
[23] K. Subramani and M. Subramaniam, *Creation of original Tamil character dataset through segregation of ancient palm leaf manuscripts in medicine*, Expert Syst. 38(1) (2020).
[24] *Tamil Character Classification*, https://kaggle.com/dhuruvapriyan/tamil-character-classification, 2020.
[25] K. Younis and A. Khateeb, *Arabic hand-written character recognition based on deep convolutional neural networks*, JJCIT 3(3) (2017) 186.