Image Classification using a Hybrid LSTM-CNN Deep Neural Network

Aditi, Mayank Kumar Nagda, Poovammal E

Abstract: ***This work elaborates on the integration of the rudimentary Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM), resulting in a new paradigm in the well-explored field of image classification. LSTM is one kind of Recurrent Neural Network (RNN) which has the potential to memorize long-term dependencies. It was observed that LSTMs are able to complement the feature extraction ability of CNN when used in a layered order. LSTMs have the capacity to selectively remember patterns for a long duration of time and CNNs are able to extract the important features out of it. This LSTM-CNN layered structure, when used for image classification, has an edge over conventional CNN classifier. The model which has been proposed is based on the sets of Artificial Neural Network like Recurrent and Convolutional neural network; hence this model is robust and suitable to a wide spectrum of classification tasks. To validate these results, we have tested our model on two standard datasets. The results have been compared with other classifiers to establish the significance of our proposed model.***

Keywords**: *Artificial Intelligence, Computer Vision, Deep Learning, Neural Networks.***

I. INTRODUCTION

Computer Vision is a topic which has observed wide attention of researchers in the past few decades. It is an interdisciplinary field that aims at gaining a high-level understanding from digital images and videos. It aims at developing methods that can reproduce the capability of human vision. Computer Vision aims at developing different methods to understand the content of digital images. In order to understand the images better, computer vision automatically extracts information from input images and videos. Some of the major applications of computer vision are navigation, assisting humans in identification tasks, event detection and organizing information [26][27]. Image recognition and classification continues to be a predominant area in the field of computer vision. It has recently gained a lot of fame due to its widespread applications. Image recognition is used to perform a large number of visual tasks which are based on machines. A few of these tasks include the labelling of the images with meta-tags, performance of image content search, guidance to autonomous robots and self-driving cars [1][2][3][4]. Image classification has an integral role to play in the field of computer-aided-diagnosis as well. Medical image classification which is a part of computer-aided-diagnosis aims at achieving a high accuracy along with the identification of the parts of the human body which are infected by the disease [7].

These wide range applications of image classification and recognition necessitate the need for good learning algorithms and models which can perform these tasks with high accuracy. In the last few decades rapid advancement has been done in this field using different machine learning algorithms and models. Machine learning is an important application of artificial intelligence (AI). Systems trained with the help of machine learning do not require explicit programming. They can learn automatically based on the previous experience [9]. In recent years researchers have discovered another breakthrough technology in this field which is popularly known as Deep Learning [28]. Unlike machine learning which employs shallow architectures, deep learning resembles the pattern of our brain which is quite similar to a deep architecture. Due to these deep architectures, the information undergoes through multiple transformations before it is finally represented. It passes the input through various layers of simulated neural connection to achieve improved accuracy.

A novel hybrid deep neural network consisting of LSTM and CNN layer has been proposed in this paper. In order to perform a comprehensive evaluation of the proposed model of deep convolutional neural networks, we have applied it on two of the classic image classification datasets i.e. MNIST [10] and IDC Breast Cancer [25]. To establish the efficiency of the proposed model, a comparison has been made with the other state-of-the-art classifiers.

II. RELATED WORK

Image classification task has recently gained fame amongst researchers due to its colossal contribution in the computer vision field. It finds its application in a variety of automation tasks such as self-driving cars, scene detection and computer-aided diagnosis [4][5][6].

Due to a large number of applications of image classification, different methods have been adopted to achieve higher accuracy in this field. Out of all these methods, Deep learning models are the most preferred ones since they possess deep architectures. They have been employed to achieve reliable results on simple image classification tasks such as handwritten digits, face recognition, textures and objects [11]. It has been observed that deep architectures perform better than shallow ones like SVM, and hence this justifies their increased usage in this field [12].

Revised Manuscript Received on August 05, 2019

***** Correspondence Author

Aditi, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India.

Mayank Kumar Nagda, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India.

E. Poovammal*, Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India.

One of the most common types of deep neural networks applied in the domain of visual imagery is Convolutional Neural Network (CNN) [13]. LeCun et al. applied the supervised deep back-propagation convolutional network for recognition of digits [10]. A state-of-the-art layered HCCR-GoogLeNet has been proposed for the Chinese handwritten characters recognition by Zhuoyao Zhong et al. They found out that such architectures perform better even with a lesser number of parameters [16]. Apart from this, CNNs have proved to be efficient in a variety of other tasks. Hokuto Kagaya applied CNN to recognize food images. They concluded that the results obtained through CNN are better than those of conventional methods [14]. These Deep Convolutional Networks (DNN) have been found to perform remarkably good in the task of face recognition as well [15]. CNN has also been employed to visualize the performance on several scenic datasets [17]. Activation functions play a vital role in the functioning of Neural Networks. They make the task of backpropagation possible by supplying gradients along with the errors to update weight and bias. Kaiming He [18] investigated the rectifier properties of Neural Networks. They proposed a novel Parametric Rectified Linear Unit (PReLU) as a generalized version of the traditional rectified unit which gave remarkable results on ImageNet 2012 dataset [19]. Rectified units help to alleviate the problem of vanishing gradient. However, this problem can be avoided by using special gated recurrent units having tanh as an activation function. These special units are known as Long Short-Term Memory [20]. Wonmin Byeon et al. performed the pixel-level segmentation and classification of scene images using LSTM which outperformed the state-of-the-art methods in the same field [21]. Soo Hyun Bae et al. proposed a parallel combination of CNN and LSTM layer to efficiently improve the classification accuracy of acoustic scenes [5]. The datasets used in our work have been extensively used to discover and unveil new breakthrough technologies in this field. Ciresan et al. attained an accuracy of 99.65% on MNIST dataset with the help of 6-layer Neural Network [8]. Dan Cires et al. achieved the near-human accuracy on the same dataset with an error rate of only 0.23% [29]. Cruz-Roa, on the other hand applied a deep learning approach on the IDC breast cancer dataset. They attained F-measure and balanced accuracy of 71.80% and 84.23% respectively [25].

III. PROPOSED MODEL

The proposed image classification model is a layered Deep Neural Network consisting of Long short-term memory (LSTM) and a Convolutional Neural Network (CNN). LSTM is a kind of RNN which unlike the traditional feedforward neural networks has feedback connections. This feature of possessing feedback connections make LSTM a type of "general purpose computer" enabling it to compute everything a Turing machine can. Section III (A) and III (B) gives a detailed explanation of LSTMs and CNNs.

A. *Long short-term memory (LSTM)*

As represented in Fig.1, we can define a unit of LSTM at each time step t as a collection of vectors in R^d consisting of forget gate f_t, input gate i_t, memory cell c_t, output gate o_t and a hidden state h_t, where d is the magnitude of the memory dimension [30][31]. The LSTM equations are numbered from (1) through (6).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (1)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \quad (2)$$

$$\tilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \quad (5)$$

$$h_t = o_t * \tanh\left(C_t\right) \quad (6)$$

In the equations, b and W denote the bias vector and weight matrices for the input gate, output gate, forget gate, memory cell, tanh layer and the hidden layer. While σ denotes logistic sigmoid function.

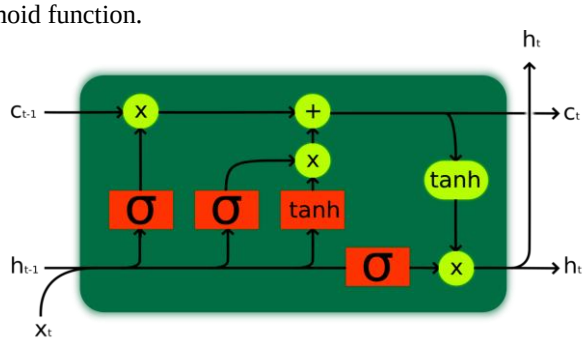

Figure 1 A LSTM Cell

In an LSTM unit, the input gate is fed with a new stream of data at every time step t and is responsible for making the decision on remembering the information it processes. Forget gate, on the other hand, is responsible for regulating the amount of information that should be removed from the memory cell.

B. *Convolutional Neural Network*

CNNs are influenced by the biological neural networks and are defined as a regularized version of multilayer perceptron [22][23][24]. They are structured as a fully connected layer, which means that each neuron present in one layer is connected to other neurons in the succeeding layer. CNNs are known to use little bit of pre-processing in comparison to other traditional image classification methods. A CNN is made up of different layers, one of them is input layer, one is output layer and the others include multiple hidden layers. The hidden layers are composed of multiple convolutional layers that convolve with multiplication or dot product. Fig. 2 represents a simple architecture of a CNN being used to predict handwritten digits. CNNs can easily capture important features from an image by taking advantage of local spatial coherence with which they give good results on image classification problems. Also, possessing the quality to extract important features makes CNNs a very good option for a completely new task.

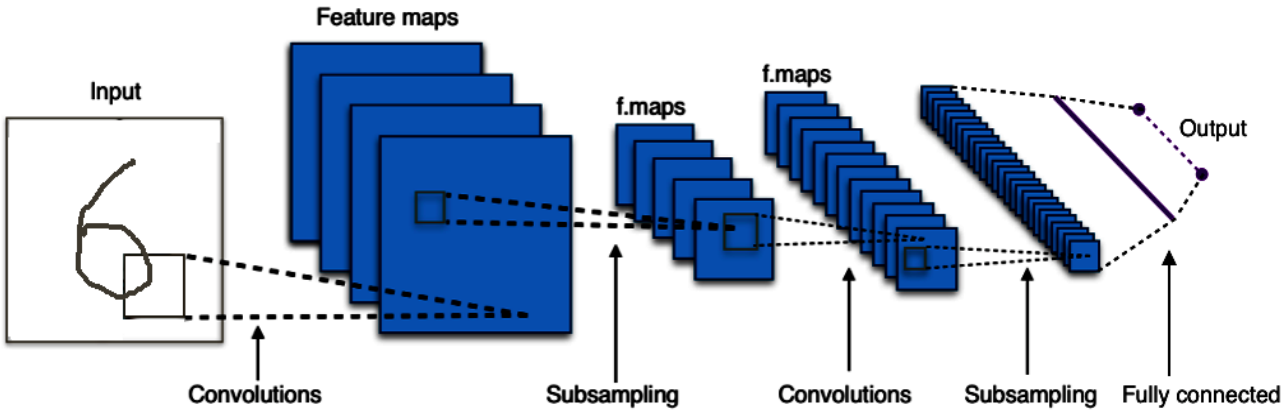

Figure 2 Convolutional Neural Network Architecture

C. *The LSTM-CNN Neural Network*

Figure 3 represents the architecture of our proposed LSTM-CNN model. The images are passed through the input layer. This input layer is then fed into the Batch Normalization layer. Batch Normalization layer applies transformation to the preceding layer which maintains the standard deviation of activation function close to 1 and mean activation close to 0, thus normalizing it. We apply normalization to each feature such that every feature map of the input is separately normalized. "Axis" argument specifies the axis on which the normalization has to be performed. we have applied statistics to every batch in order to normalize the data during training, and during testing, we use running averages computed during the training phase.

The output shape from Batch Normalization layer is the same as that of the input shape, which makes it unusable for LSTM cell. To change the shape to the desired dimension, a reshape layer can be used before the LSTM layer. After the dimensions of the input layer are reshaped it is passed through the LSTM cell. Tanh i.e. the Hyperbolic tangent is used as an activation function of the LSTM cell. The LSTM cell also has a dropout rate to help prevent overfitting of data.

Because of these characteristics of LSTM, it will particularly remember the long-term dependencies and shape of the input image in a particular pattern. The output from the LSTM layer is directly provided to the convolutional layer. A convolution kernel is created by the convolutional layer which produces a tensor of outputs by convolving with the layer input over a single spatial (temporal) dimension. The convolutional layer will extract the local important features. Rectified Linear Unit (ReLU) has been used as an activation function in this convolutional layer. A dropout layer can be applied after the convolutional layer to prevent the overfitting present due to "fully-connectedness" of the neurons in the CNN. For complex classification problems, a committee of LSTM-CNN networks can be used.

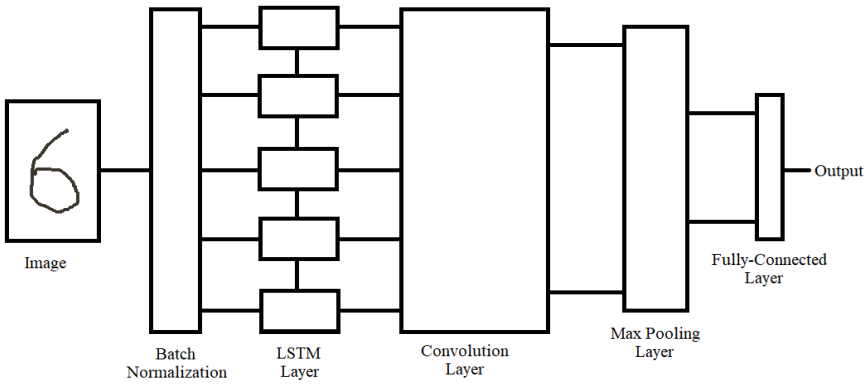

Figure 3 LSTM-CNN Neural Network Architecture

IV. EXPERIMENT AND RESULTS

The hybrid model proposed in Section III (C) is theoretically supposed to have a better performance than its conventional counterparts in the task of image classification. In theory, this model has an edge over other models in tasks such as handwriting recognition, sentimental analysis, and other such problems which can benefit from the LSTMs capability to remember long term dependencies.

To validate the proposed model, we have used the benchmark datasets and compared the final results. We have chosen two different datasets to benchmark our results. The highly competitive MNIST and Breast Cancer IDC datasets are chosen for testing the proposed model.

A. *MNIST Dataset*

The available MNIST dataset is composed of a large number of images of handwritten digits. This MNIST dataset is a subset of the large NIST dataset [10]. By default, to get a fixed-size image the digits have been size-normalized and centered already.

To evaluate our proposed model, we compare it against CNN, LSTM, and CNN-LSTM as these models have attained excellent and near perfect accuracy on the same dataset. The parameters of these models are kept identical to perform a fair comparison.

The average training accuracy and validation accuracy are used to perform comparisons between different models. As represented in Fig. 4 and Fig. 5, it was observed that the proposed LSTM-CNN model performed significantly better than the other classifiers. The overall training, as well as validation accuracy of the LSTM-CNN model, was found to be greater than those of other models it was compared against.

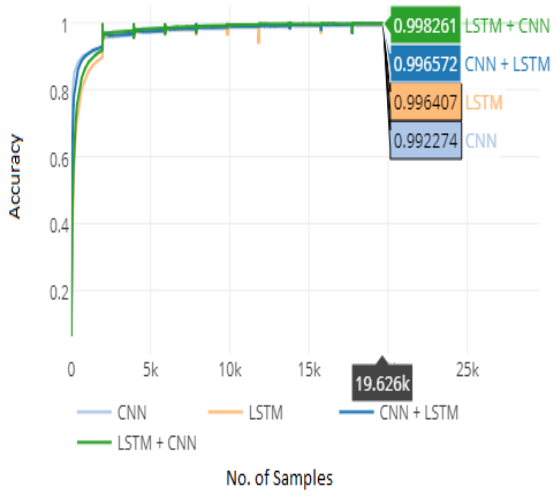

Figure 4 Training Accuracies for MNIST dataset

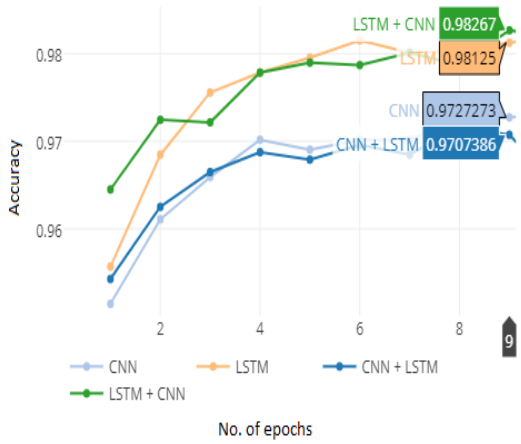

Figure 5 Validation Accuracy for MNIST dataset

Table I displays the accuracy achieved by different classifiers. On analyzing the results, it can easily be comprehended that the LSTM-CNN model outperforms the other models as a significant difference in percentage accuracies can be observed.

Table I Accuracies comparison over MNIST dataset.

Model	Training Accuracy (%)	Validation Accuracy (%)
LSTM + CNN	**99.8261**	**98.267**
CNN + LSTM	99.6572	97.074
LSTM	99.6407	98.125
CNN	99.2274	97.273

The proposed LSTM-CNN model was given a full run with an addition of 2 more LSTM-CNN layers as mentioned in the model. This particular model gives an accuracy of more than 99% on both validation and test set. It gives near-perfect accuracy of 99.7% on training set and 99.29% on the validation set as represented in Fig. 6 and Fig. 7 respectively.

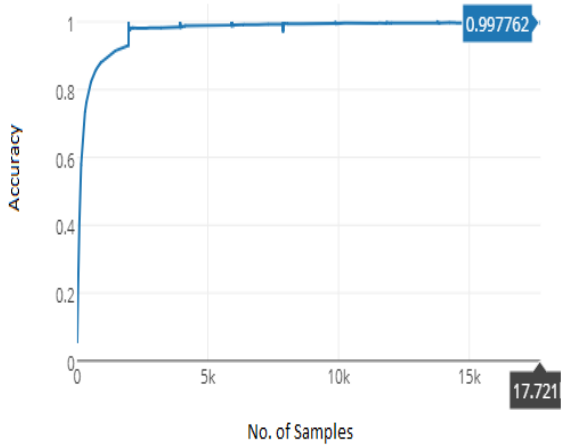

Figure 6 Training accuracy of LSTM-CNN

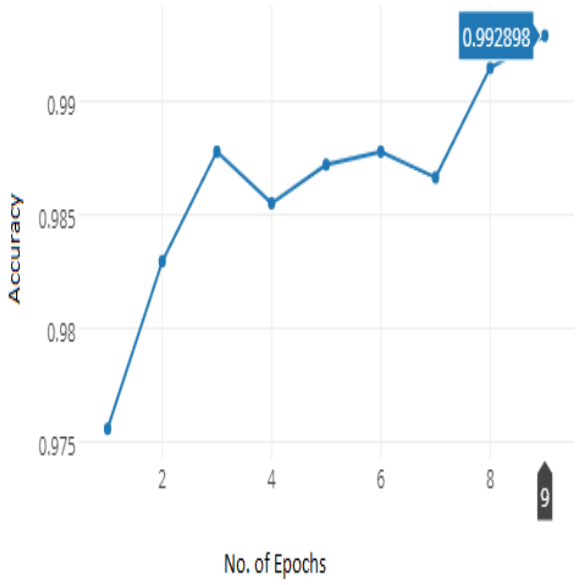

Figure 7 Validation Accuracy of LSTM-CNN

B. *Breast Cancer IDC Dataset*

Breast cancer is one amongst the common forms of cancer in females. The manual evaluation of the presence of invasive ductal carcinoma (IDC) tissue regions present in the whole slide images (WSI) is a critical task, which can be assisted with the help of computerized evaluation. Since Invasive Ductal Carcinoma (IDC) is one of the most common type of a breast cancers, we use the IDC dataset produced by Cruz-Roa A et al. for evaluation of our proposed model [25]. This dataset consists of digital image patches that were derived from 162 patients. These images are small patches that were extracted from digital images of breast tissue samples. We utilize these images to detect the presence of IDC tissue regions in WSI.

Fig. 8 and Fig. 9 represents the training and validation accuracies of different models on the IDC dataset.

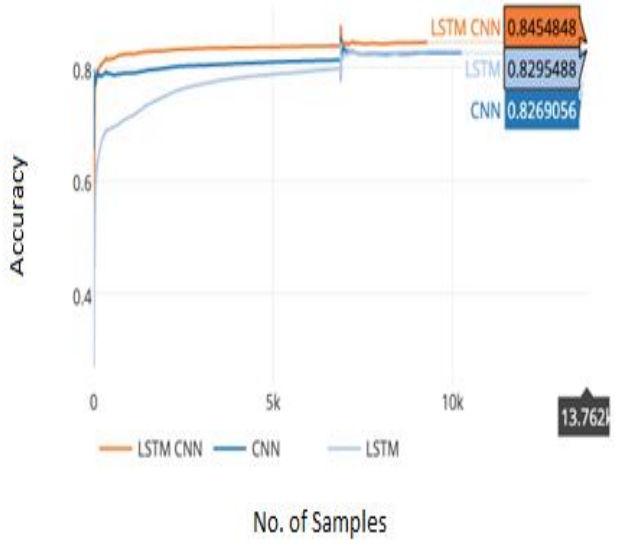

Figure 8 Cancer IDC dataset training accuracies

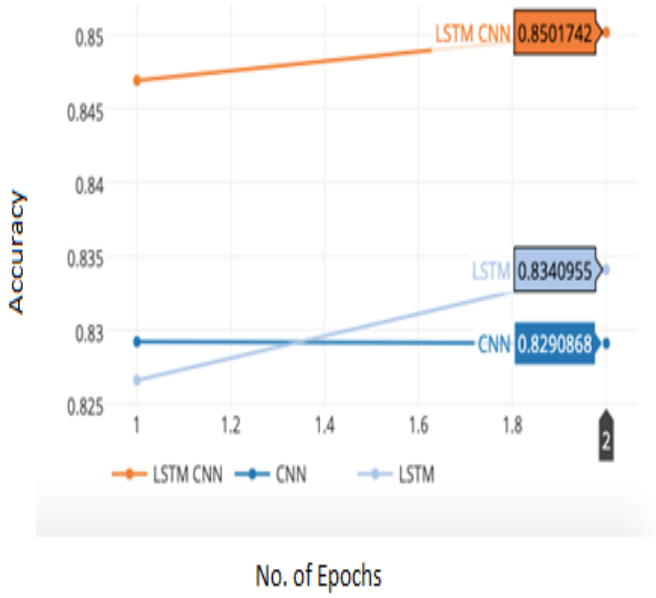

Figure 9 Validation accuracies of IDC dataset

Table II shows a comparison of our model with other classifiers. Our hybrid model achieves a training accuracy of 84.5% and a validation accuracy of 85% which is significantly better than the two other classifiers that it was compared against. These results hence set a new benchmark in this field.

Table II. Accuracy comparison over IDC Breast Cancer Dataset

Model	Training Accuracy (%)	Validation Accuracy (%)
LSTM + CNN	**84.548**	**85.017**
LSTM	82.955	83.410
CNN	82.691	82.909

V. CONCLUSION

In this work, we have proposed a novel LSTM-CNN hybrid model for improving the accuracy of the image classification task. In comparison with other state-of-the-art classifiers like CNN, LSTM and hybrid CNN-LSTM, we found out that our proposed model significantly outperforms them. To establish the significance of our model, we tested it against two benchmark datasets i.e.

MNIST handwritten digit dataset and IDC Breast Cancer dataset. On both the datasets, our model gave remarkable accuracy. On MNIST dataset, the proposed LSTM-CNN hybrid model attained a training accuracy of 99.8% and a validation accuracy of 98.2%. On using the multiple LSTM-CNN layers it further gave us an improved validation accuracy of 99.29%. Similarly, benchmark results were obtained on

IDC Breast Cancer dataset with the attained training and validation accuracies of 84.5% and 85% respectively. Having attained a high accuracy with a single layer of LSTM-CNN, the model lays the foundation for further improvements by utilizing its multiple layers and controlling the overfitting in the presence of powerful GPUs and optimized drop out layers.

REFERENCES

1. Sharma, A., Hua, G., Liu, Z. and Zhang, Z., 2008, June. Meta-tag propagation by co-training an ensemble classifier for improving image search relevance. In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-6). IEEE.
2. Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y. and Li, J., 2014, November. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 157-166). ACM.
3. Diallo, A.D., Gobee, S. and Durairajah, V., 2015. Autonomous tour guide robot using embedded system control. Procedia Computer Science, 76, pp.126-133.
4. Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R. and Mujica, F., 2015. An empirical evaluation of deep learning on highway driving. arXiv preprint arXiv:1504.01716.
5. Bae, S.H., Choi, I. and Kim, N.S., 2016, September. Acoustic scene classification using parallel combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016) (pp. 11-15).
6. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H. and Chen, Y.J., 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy, 8.
7. Miranda, E., Aryuni, M. and Irwansyah, E., 2016, November. A survey of medical image classification techniques. In 2016 International Conference on Information Management and Technology (ICIMTech) (pp. 56-61). IEEE.
8. Claudiu Ciresan, D., Meier, U., Gambardella, L.M. and Schmidhuber, J., 2010. Deep big simple neural nets excel on handwritten digit recognition. arXiv preprint arXiv:1003.0358.
9. Michie, D., Spiegelhalter, D.J. and Taylor, C.C., 1994. Machine learning. Neural and Statistical Classification, 13.
10. LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.
11. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z. and Ma, Y., 2015. PCANet: A simple deep learning baseline for image classification?. IEEE transactions on image processing, 24(12), pp.5017-5032.
12. Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H., 2007. Greedy layer-wise training of deep networks. In Advances in neural information processing systems (pp. 153-160).
13. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
14. Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M. and Cagnoni, S., 2016, October. Food image recognition using very deep convolutional networks. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management (pp. 41-49). ACM.
15. Chen, J.C., Patel, V.M. and Chellappa, R., 2016, March. Unconstrained face verification using deep cnn features. In 2016 IEEE winter conference on applications of computer vision (WACV) (pp. 1-9). IEEE.
16. Zhong, Z., Jin, L. and Xie, Z., 2015, August. High performance offline handwritten chinese character recognition using googlenet and directional feature maps. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR) (pp. 846-850). IEEE.
17. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A., 2014. Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).
18. He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), pp.211-252.
20. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.
21. Byeon, W., Breuel, T.M., Raue, F. and Liwicki, M., 2015. Scene labeling with lstm recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3547-3555).
22. Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), pp.193-202.
23. Hubel, D.H. and Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology, 195(1), pp.215-243.
24. Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y., 2003. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Networks, 16(5-6), pp.555-559.
25. Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J. and Madabhushi, A., 2014, March. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In Medical Imaging 2014: Digital Pathology(Vol. 9041, p. 904103). International Society for Optics and Photonics.
26. Sinha, P., Balas, B., Ostrovsky, Y. and Russell, R., 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. Proceedings of the IEEE, 94(11), pp.1948-1962.
27. Vu, A., Ramanandan, A., Chen, A., Farrell, J.A. and Barth, M., 2012. Real-time computer vision/DGPS-aided inertial navigation system for lane-level vehicle navigation. IEEE Transactions on Intelligent Transportation Systems, 13(2), pp.899-913.
28. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), p.436.
29. Cireşan, D., Meier, U. and Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. arXiv preprint arXiv:1202.2745.
30. Tai, K.S., Socher, R. and Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.
31. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

AUTHORS PROFILE

Aditi Arora is an undergrad student at SRM Institute of Science and Technology where she is currently pursuing her bachelor's degree in Computer Science and Engineering. Her research interests include big data, Machine Learning, Data Analytics and Natural Language Processing. Aditi has been actively involved in various research endeavors with an aim of holistic development and advancement of society.

Mayank Kumar Nagda is an alumnus of SRM Institute of Science and Technology where he earned a bachelor's degree in the field of Computer Science and Engineering. Artificial Intelligence, Natural Language Processing, and Data Analytics are some of his areas of interests and expertise. Mayank is also fascinated by the idea of introducing automation in the regular lives of human beings so that it can assist and can create a new firm base for the next human evolution.

Dr. E. Poovammal is a Professor in the Department of Computer Science and Engineering at SRM Institute of Science and Technology. She joined in SRM in the year 1996. Before joining SRM, she worked in industry for five years. She obtained her B.E. Degree in Electrical and Electronics Engineering from Madurai Kamaraj University in the year 1990, M.E degree in Computer Science and Engineering from Madras University in the year 2002 and Ph.D. degree in Computer Science and Engineering from SRM University in 2011. Her research interests include data Big Data Analytics and machine learning. She is certified as Adjunct Faculty by Institute of software Research, Carnegie Mellon University, Pittsburgh, USA. She has published more than 40 referred journals and presented various international and national conferences.