Impact of Automatic Feature Extraction in Deep Learning Architecture

Fatma Shaheen, Brijesh Verma and Md Asafuddoula
Centre for Intelligent Systems
Central Queensland University, Brisbane, Australia
{f.shaheen, b.verma, a.asafuddoula}@cqu.edu.au

Abstract— **This paper presents the impact of automatic feature extraction used in a deep learning architecture such as Convolutional Neural Network (CNN). Recently CNN has become a very popular tool for image classification which can automatically extract features, learn and classify them. It is a common belief that CNN can always perform better than other well-known classifiers. However, there is no systematic study which shows that automatic feature extraction in CNN is any better than other simple feature extraction techniques, and there is no study which shows that other simple neural network architectures cannot achieve same accuracy as CNN. In this paper, a systematic study to investigate CNN's feature extraction is presented. CNN with automatic feature extraction is firstly evaluated on a number of benchmark datasets and then a simple traditional Multi-Layer Perceptron (MLP) with full image, and manual feature extraction are evaluated on the same benchmark datasets. The purpose is to see whether feature extraction in CNN performs any better than a simple feature with MLP and full image with MLP. Many experiments were systematically conducted by varying number of epochs and hidden neurons. The experimental results revealed that traditional MLP with suitable parameters can perform as good as CNN or better in certain cases.**

Index Terms— **Image Classification; Feature Extraction; Deep -Learning; Convolutional Neural Network; Multi-Layer Perceptron;**

I. INTRODUCTION

Deep learning architecture such as convolution neural network (CNN) has recently gain popularity in real-world applications. The main reason for this popularity is that it can automatically extract features and classify them so that there is no need for manual feature extraction and selection. However, there has been very little research to systematically evaluate automatic feature extraction and classification abilities of deep learning architecture.

Classification is one of the most important and essential process of feature identification in many real world applications [1]. A small error in classification process can have significant impact on information processing in different fields like disease detection in medical science [2, 3], customer identification for online banking [4], forecasting in environmental science [5] and many more. Therefore, it is significantly important to have an accurate classifier with high and consistent accuracy which can be applied in real-world applications. A lot of researches has been done in developing new classifiers in particular classifiers which can learn and adapt to new conditions with minimal parameters/model changes [2]. However, the performance of a classifier with a set of parameters can perform better in one application but may perform extremely poor in other real-world applications which leads the researchers to move and develop new methods which can perform better across different datasets. They also have the benefit of extracting features automatically. However, it is unclear whether feature extraction incorporated

in deep learning architecture is any better than the manual feature extraction techniques. Therefore it is significantly important to conduct a systematic study to answer the above mentioned research question.

The remainder of the paper is organized as follows. Section II presents relevant background. Section III describes the proposed research methodology. Section IV presents the experiments, and finally Section V concludes the paper.

II. BACKGROUND

Convolution Neural Network (CNN) is one of the successful machine learning techniques for image classification. CNN involves multiple processing layers, therefore it is known as deep structured learning [6]. CNN is also considered as a biologically-inspired variants of MLPs. Deep learning in CNN involves multiple processing layers, composed of multiple linear and non-linear transformations. The method is motivated by the animal's visual cortex, i.e., based on the arrangement of cells and its learning process. On the other hand, MLP is a popular form of artificial neural network which can be used for classification embedded with a manual feature extraction or without a feature extraction. MLP doesn't have automatic feature extraction as in CNN. Over the past few years, CNN research trends have grown (since 1972) [1, 7]. It can be seen from Figure 1 that the research articles in the field of CNN keeps increasing due to its popularity.
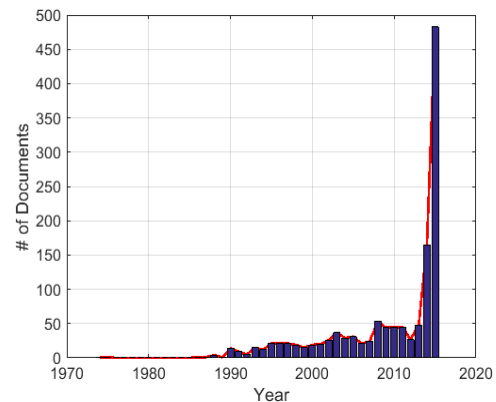

Fig. 1: CNN research trend since 1972 to 2015

CNN is popular due to its automatic feature extraction for image classification involving large datasets. A number of deep learning architectures have been proposed, which can successfully extract the features and classify them. Table I presents some of the top CNN architectures and its reported application. CNN divides the tasks into a number of layers. For a simple case of CIFAR-10 classification, the layers include [INPUT - CONV - RELU - POOL – FC] i.e., Input layer is fed

Table I: A brief review of CNN

Author	CNN Type	Brief Description of Architecture	Application
LeCun et al.,	LeNet [7]	First application of CNN using [INPUT-CONV-SUBSAMPLING-CONV-SUBSAMPLING-FC]	Zip codes, Handwritten digits.
Krizhevsky et al.,	AlexNet [8]	The approach popularizes the use of CNN for computer vision around 2012. It utilizes [CONV-5xMAX POOLING- RELU-FC]	Handwritten digits and ILSVRC-2010 image datasets.
Zeiler et al.,	ZF Net [9]	The approach is slightly similar with AlexNet and [UNPOOLED MAPS-RECTIFIED-RECONSTRUCTION-POOLING- RECTIFIED-FC]	ImageNet 2012, Caltech-101, Caltech-256, PASCAL 2012.
Szegedy et al.,	GoogLeNet[10]	The main contribution of the method is to introduce an *Inception Module* after pooling which reduces the parameters. [INPUT-CONV-MAX POOLING- INCEPTION-RELU-SOFTMAX]	ILSVRC 2012–2014
Simonyan et al.,	VGGNet[11]	This method utilizes almost similar configuration of GoogleNet but without *Inception*, [INPUT-3xMAXPOOL-3xFC-SOFTMAX]	ILSVRC 2012–2014
Kaiming et al.,	ResNet[12]	The method is winner of ILSVRC 2015. It skips connection but uses heavy batch normalization. It does not have FC at the end.	CIFAR-10, ILSVRC 2012

to Convolution Layer (CONV) where a set of learnable filters are used. These filters are then fed to the Rectified Linear Units (RELU) to increase the non-linearity of the decision function, after RELU layer, pooling layer is used where non-linear down sampling is done, and finally a Fully Connected (FC) layer is used for classification [13].

Apart from the previously described CNN, Ba et al., [14] proposed a new version of a Deep Recurrent of Visual Attention Model (DRVAM) using deep recurrent neural network and trained with reinforcement learning by attending to the most relevant regions of the input image. The method firstly applied to the Mixed National Institute of Standards and Technology (MNIST) dataset and then a real-world multi-digit Street View House Number (SVHN) dataset. It was found that multi-digit house number recognition was more successful when compared with the current state-of-the-art Convolutional Neural Networks (ConvNets).

Donahue et al., [15] proposed a recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable. This model demonstrated its value on benchmark video recognition tasks. The dataset used in the task was over 12,000 videos categorized into 101 human action classes. Dundar et al.,[16] found the need to label the data for training deep neural networks and proposed a clustering algorithm to reduce the number of correlated parameters and to increase test categorization accuracy. A new input patch extraction method for feature extraction was used to reduce the redundancy between the filters at neighbouring locations. An

accuracy of 74.1% was obtained on an image recognition dataset i.e., STL-10 and a test error of 0.5% on MNIST dataset.

Krizhevsky et al., [6] proposed ImageNet deep convolutional neural network to classify over 1.2 million high-resolution images. The neural network they used had 60 million parameters and 650,000 neurons, consisting of five convolutional layers, some of which were followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. Non-saturated neurons were used to make the training faster along with efficient GPU implementation.

Lenz et al., [17] proposed deep learning for robotic hand grasp detection. A two-step cascade system with two deep networks was used, where the top detections from the first network were re-evaluated by the second network. Deep network has also been applied for black box image classification problem with additional 130 thousand of extra unlabelled samples [18]. In [2] a robust 4-layer Convolutional Neural Network (CNN) architecture has been proposed for face recognition problem. The proposed method can handle facial images which may have occlusions, poses, facial expressions and varying illumination.

Although CNN has recently been applied to different computer vision tasks, it is important to understand the learning process of CNN over other techniques. The complexity of CNN makes it difficult to use it for some handy and small scale image processing tasks. In this paper, we have compared CNN with traditional MLP for 3-different image classification tasks. In this paper, we conducted a systematic experiments on CNN and

compared with MLP to answer the following research questions, e.g., (i) is it always better to use CNN with automatic feature extraction for image classification? (ii) How CNN performs on different real-world datasets in comparison with traditional MLP? iii) How the performance of CNN can further be improved?

III. PROPOSED RESEARCH METHODOLOGY

An overview of the proposed methodology is presented below in Figure 2. In the proposed methodology, a systematic approach is presented to conduct the appropriate experiments to answer the research questions. The input image is fed to three different models (1) a convolution neural network (2) image-based MLP (i.e., input of the MLP is an image), and (3) feature-based MLP (i.e., input of the MLP is manually extracted features from an image).
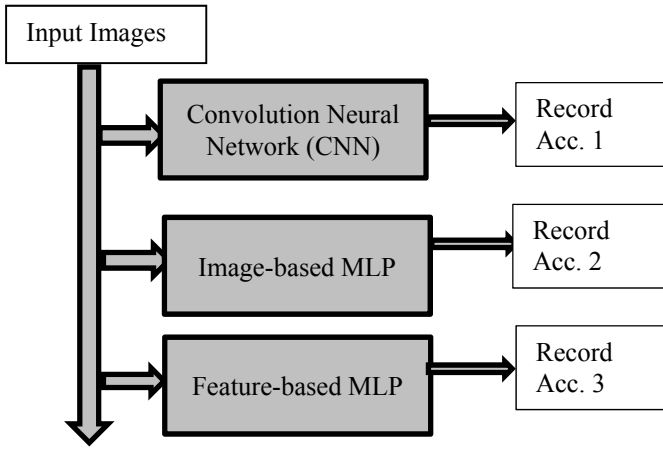

Fig. 2: Proposed research methodology for image classification with and without deep learning

As shown in Figure 2, the proposed method takes the input images directly and then apply these images to individual neural network model for classification. Similar parameter settings are used for training and testing the accuracies, where highest accuracy among different neural network models is used to verify the result, e.g., CNN. The individual component is described in the following subsection.

(1) **Automatic Feature Extraction based CNN Classifier**
In this proposed method, a similar architecture of LeCun with a slightly modified version is used. The architecture is composed of [INPUT-CONV-POOLING-CONV-POOLING-FC]. In the convolution layer, a set of learnable filters are used. Every filter is small spatially (span along width and height), but extends through full depth of input volume. An image having width W, height H and depth D color channels (i.e., $W \times H \times D$), the learnable filters divides the image width as $W_1 = (W-F+2P)/S+1$, where F refers to the spatially extends neuron size; P is the amount of zero padding, and S is the size of stride. Similarly, the height is divided by $H_1 = (H-F+2P)/S+1$, depth D_1 is the size of number of filters K. For example, an image having 28x28x3 (3 is for the color channels), if the receptive field (or filter) has a size of 5x5x3 (in total 75 neurons + 1 bias), a 5x5 window with depth three

moves along the width and height and produces a 2-D activation map.

The Pooling Layer operates independently on every depth slice of the input and resizes it spatially using the MAX operation. It accepts a volume size of $W \times H \times D$, and divides the image into $W_1 = (W-F)/S+1$ for width and $H_1 = (H-F)/S+1$ for height and depth D_1 is same as the input D. After the computation from all colour channels a max operation is done. Therefore, the feature matrix is then reduced in POOLING layer. In the last layer, a fully connected network is used. Here, in this CNN, a MLP based fully connected network has been performed for final classification. Figure 3 shows the architecture of a CNN with automatic feature extraction.
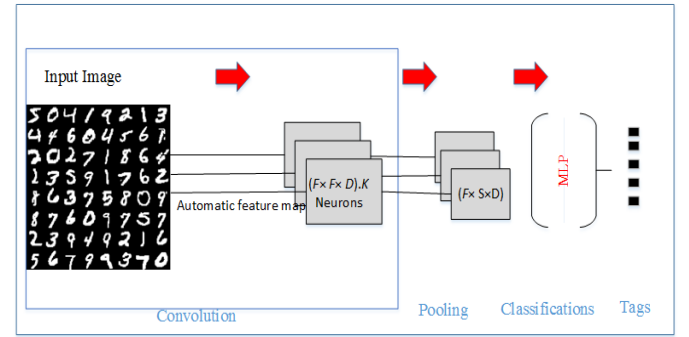

Fig. 3: Automatic Feature Extraction based CNN

(2) **Image-based MLP– Classifier**
In this proposed method, the full image is fed to a Multi-Layer Perceptron (MLP) based neural network. Firstly, the image data is normalized and then the whole image data is presented to the MLP. A conjugate gradient descent based backpropagation is used for the training. The number of hidden neurons and the training epochs are changed iteratively. An overview of this method is presented in Figure 4.
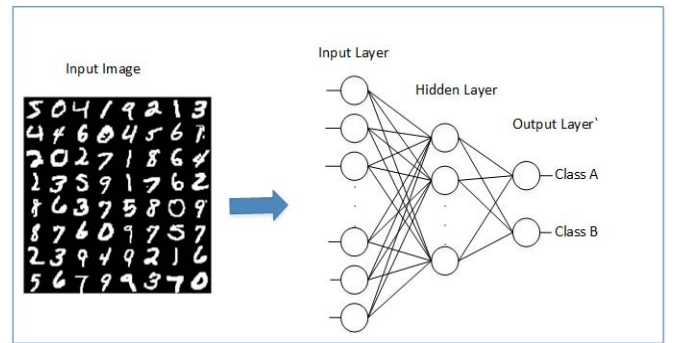

Fig. 4: Image-based MLP

(3) **Feature-based MLP–Classifier**
In this proposed method, a feature vector extracted from the image is fed to MLP base neural network. The feature-based MLP operates on the feature matrix. A formula based human

generated features/automated feature extraction is done before feeding the input image to MLP. The conjugate gradient descent based backpropagation algorithm is used for the training. In comparison to image-based MLP, feature-based MLP operates on relatively small input space due to the small number of features extracted from the image. An overview of this method is presented in Figure 5.
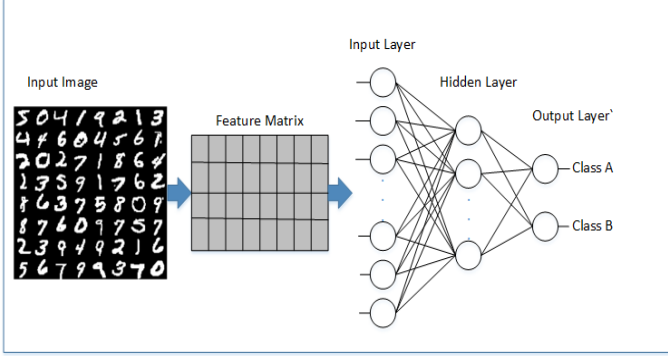

Fig. 5: Feature-based MLP

IV. DATASETS AND EXPERIMENTS

As mentioned earlier that, we have conducted the experiments on three different datasets. The datasets used in this study are the standard MNIST [19] dataset, *Cow-heat-sensor dataset* and *Roadside-vegetation* [20] dataset. MNSIT is a standard dataset used in many pattern recognition algorithms for performance analysis. The dataset contains 70,000 handwritten patterns. Since the dataset is already divided into 60,000 training and 10,000 test samples, we have used the same number of samples for consistency. For *Cow-heat-sensor* data, around 100 images have been collected from the cow firm and labelled those into two classes. The third dataset used in this study has 600 roadside images. These images are labelled with seven classes, i.e., grass-brown, grass-green, road, sky, soil, tree-leaf, and tree-stem. Since the data has seven different classes, it makes the problem relatively difficult to classify. In this experiment, we have used 75% data for training and 25% data for testing for *Cow-heat-sensor* and *Roadside-vegetation* data.

Experiments are conducted with the proposed method (i.e., CNN, image-based MLP and feature-based MLP) on each of the datasets. All the algorithms are developed and executed in MATLAB 2015b. For image-based MLP and feature-based MLP default parameters are used and trained with conjugate gradient descent based backpropagation algorithm.

A. Experiments on standard MNIST dataset

Firstly, the performance of the proposed methodology is evaluated on MNIST (Mixed National Institute of Standards and Technology) database of handwritten digits classification. The database has a training set of 60,000 examples and a test set of 10,000 examples. MNIST database is a good example for evaluating various learning techniques as it has been used by many researchers. Table II shows the classification accuracy

Table II: Classification accuracy (%) obtained from CNN on MNIST dataset

#Epochs	Accuracy on Training Set	Accuracy on Test Set
1	27.41	88.87
5	71.90	95.44
50	92.00	98,71
100	94.30	98.92
1000	98.19	98.99

obtained from CNN, which is almost 99% accurate on this dataset.

Table III shows the results obtained from image-based MLP with same parameter settings. The results obtained by image-based MLP shows almost 93.3% accuracy in comparison with CNN. The other results obtained from feature-based MLP in Table IV suggest inappropriate feature selection resulting lower accuracy.

Table III: Classification accuracy (%) on MNIST data using image-based MLP

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	82.90	66.70
	5	94.30	86.70
	50	100.00	80.00
	100	100.00	60.00
	1000	100.00	86.70
12	1	62.90	73.30
	5	100.00	86.70
	50	100.00	93.30
	100	100.00	86.70
	1000	100.00	86.70
16	1	65.70	60.00
	5	100.00	80.00
	50	100.00	86.70
	100	100.00	86.70
	1000	100.00	80.00
24	1	74.30	40.00
	5	100.00	86.70
	50	100.00	73.30
	100	100.00	86.70
	1000	100.00	73.30
120	1	57.10	60.00
	5	100.00	80.00
	50	100.00	80.00
	100	100.00	53.30
	1000	100.00	80.00

Table IV: Classification accuracy (%) on MNIST data using feature-based MLP

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	8.86	8.78
	5	9.59	9.87
	50	10.69	10.85
	100	9.23	8.95
	1000	11.55	11.32
12	1	10.47	10.76
	5	14.29	14.15
	50	11.37	11.57
	100	10.05	9.92
	1000	9.01	9.22
16	1	19.14	19.26
	5	38.73	38.25
	50	35.14	34.88
	100	33.80	34.71
	1000	40.80	41.53
24	1	30.05	30.13
	5	31.65	31.96
	50	42.36	43.02
	100	53.93	54.27
	1000	54.88	55.15
120	1	44.67	44.85
	5	65.83	65.39
	50	70.18	70.17
	100	71.66	71.66
	1000	71.74	72.00

B. Experiments on Cow-heat-sensor data

We have conducted the similar experiments on a real-world dataset to detect the change of body temperature in cows. The image data is divided into two categories (a) changed color due to the body temperature and (b) unchanged color sensor. Figure 6 shows two sample images showing device with colour change (class 1) and no colour change (class 2).

| Class-1 | Class-2 |

Fig. 6: Sample image classes of *Cow-heat-sensor* data

Table V: Classification accuracy (%) obtained from CNN on *Cow-heat-sensor* data

#Epochs	Accuracy on Training Set	Accuracy on Test Set
1	65.71	80.00
5	68.57	80.00
50	62.86	80.00
100	94.29	86.67
1000	100.00	100.00

Table V shows the classification accuracy obtained from CNN. Although the number of images in each class is very few (around 50), CNN can still successfully detect and classify the images with a good accuracy.

Table VI shows the results obtained by image-based MLP. It can be seen from the table that image-based MLP confirms similar test accuracy in comparison with the results obtained by CNN (for 1000 Epochs) for same parameter settings.

Table VI: Classification accuracy (%) obtained from image-based MLP on *Cow-heat-sensor* data.

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	54.05	30.77
	5	83.78	76.92
	50	100.00	84.62
	100	100.00	100.00
	1000	100.00	92.31
12	1	24.32	46.15
	5	97.30	84.62
	50	100.00	84.62
	100	100.00	92.31
	1000	100.00	84.62
16	1	78.38	46.15
	5	100.00	84.62
	50	100.00	84.62
	100	100.00	100.00
	1000	100.00	69.23
24	1	37.84	46.15
	5	100.00	76.92
	50	100.00	84.62
	100	100.00	76.92
	1000	100.00	76.92
120	1	72.97	46.15
	5	94.59	53.85
	50	100.00	84.62
	100	100.00	76.92
	1000	100.00	61.54

Table VII: Classification accuracy (%) obtained from feature-based MLP on *Cow-heat-sensor* data.

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	71.43	93.33
	5	74.29	66.67
	50	100.00	93.33
	100	100.00	100.00
	1000	100.00	93.33
12	1	74.29	80.00
	5	82.86	86.67
	50	100.00	93.33
	100	100.00	93.33
	1000	100.00	93.33
16	1	51.43	33.33
	5	85.71	80.00
	50	100.00	86.67
	100	100.00	86.67
	1000	100.00	93.33
24	1	60.00	53.33
	5	91.43	86.67
	50	100.00	93.33
	100	100.00	93.33
	1000	100.00	93.33
120	1	45.71	53.33
	5	91.43	66.67
	50	100.00	100.00
	100	100.00	86.67
	1000	100.00	73.33

Table VII shows the results obtained by feature-based MLP. It can be seen from the table that feature-based MLP again confirms the similar performance as obtained by image-based MLP. It is also noticeable that, feature-based MLP is also able to achieve high accuracy.

C. Experiments on Roadside-vegetation data

Similar experiments have been conducted on a third real-world dataset, where the main purpose is to identify areas of fire risk based on roadside vegetation [20], where the brown grasses are prone to bushfire. The dataset contains 600 images of 7 different classes (i.e., grass-brown, grass-green, road, sky, soil, tree-leaf, and tree-stem). Figure 7 shows the individual class representation.

The proposed method using CNN, image-based MLP, and feature-based MLP are applied to answer our research question. Firstly, CNN is applied to the dataset, and the performance is recorded. Table VIII shows the results obtained by CNN.

The results obtained from CNN shows relatively low classification accuracy than the results obtained in the previous
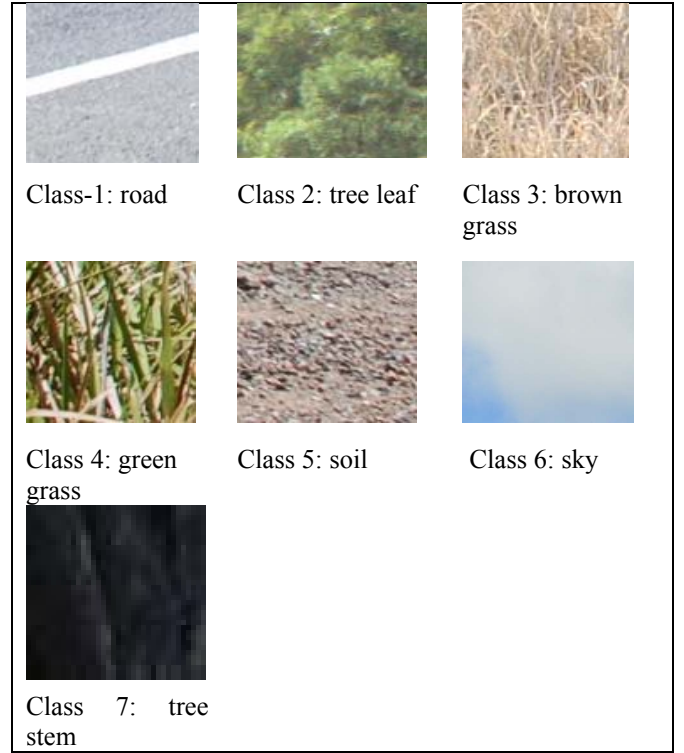

Fig. 7: Sample image classes of *Roadside-vegetation* data.

datasets. The experiments using proposed image-based and feature-based MLPs on the same *Roadside-vegetation* image dataset have been conducted. The results from image-based MLP and feature-based MLP are presented in Table IX and Table X. It is interesting to see that, image-based MLP and feature-based MLP has got slightly higher accuracy and confirm the results of CNN for some parameter settings.

This dataset is a good example showing the importance of improved and accurate feature extraction. Since the feature-based MLP utilizes manual feature extraction, therefore it shows slightly better accuracy in comparison with automatic feature extraction based CNN and image-based MLP.

D. Comparison and results analysis

We have conducted systematic experiments on three different datasets. The first dataset is a standard MNIST

Table VIII: Classification accuracy (%) obtained from CNN on roadside vegetation data

#Epochs	Accuracy on Training Set	Accuracy on Test Set
1	32.71	23.00
5	38.57	28.00
50	52.86	42.10
100	72.29	66.17
1000	93.20	72.71

Table IX: Classification accuracy (%) obtained from image-based MLP on *Roadside-vegetation* data

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	25.88	17.65
	5	19.35	22.94
	50	61.06	50.00
	100	72.11	57.65
	1000	98.99	56.47
12	1	5.78	5.88
	5	22.11	18.24
	50	71.36	58.82
	100	94.47	68.24
	1000	100.00	61.76
16	1	14.57	14.71
	5	27.39	22.94
	50	89.95	72.35
	100	95.48	70.00
	1000	100.00	63.53
24	1	0.75	1.18
	5	25.13	20.59
	50	86.43	68.82
	100	94.97	66.47
	1000	100.00	67.06
120	1	15.83	10.59
	5	30.90	32.35
	50	95.48	68.24
	100	98.74	70.59
	1000	100.00	64.12

Table X: Classification accuracy (%) obtained from feature- based MLP on *Roadside-vegetation* data

#Hidden Neurons	#Epochs	Accuracy on Training Set	Accuracy on Test Set
6	1	18.34	15.29
	5	20.35	15.88
	50	63.07	58.24
	100	84.42	69.41
	1000	97.99	64.12
12	1	16.83	17.06
	5	28.64	29.41
	50	67.09	61.18
	100	82.66	64.12
	1000	99.75	65.29
16	1	15.08	15.88
	5	20.60	23.53
	50	75.38	62.35
	100	87.94	72.35
	1000	99.75	68.82
24	1	13.82	14.71
	5	26.88	32.94
	50	70.10	63.53
	100	91.96	72.94
	1000	99.75	62.35
120	1	21.36	27.06
	5	30.40	41.18
	50	80.90	64.71
	100	93.22	72.94
	1000	100.00	67.06

handwritten digits classification data, which has a long history of CNN being successful. The proposed methodology is further evaluated on a slightly different and challenging real-world dataset (i.e., *Roadside-vegetation* data). Although CNN has performed well, it is worth to note that CNN is not always better feature extractor and classifier. Traditional image and feature based MLPs have performed as good as or better than CNN in 2 of three datasets. CNN with automatic feature extraction doesn't always perform better, so it is not recommended to be used for any classification task with small dataset.

V. CONCLUSION

This paper presented a research methodology to identify the impact of automatic feature extraction and classification used in deep learning such as CNN. An approach has been proposed to systematically analyze the classification accuracy
of CNN, image, and feature based traditional MLPs. CNN with automatic feature extraction was firstly evaluated on a well-established MNIST dataset and then a traditional Multi-Layer Perceptron (MLP) with full image, and a manual feature extraction were evaluated on the same benchmark dataset. Two

other real-world datasets such as *Cow-heat-sensor* and *Roadside-vegetation* were also used. The image data is firstly fed to CNN, MLP with full image and MLP with manual feature extraction. Similar experimental conditions were used for the training of each of the models. The research methods with exhaustive systematic experiments suggest that CNN with automatic feature extraction based image classification can perform well but does not substantiate as a robust technique for all types of image classification. It has been found that, for real-world dataset a simple traditional MLP may serve equivalent or better than CNN under certain experimental conditions. This research also suggests a number of things to improve the CNN performance i.e., (i) robust feature extraction method is needed to improve the convolution (ii) better classification technique is needed in conjunction with other well-established classifiers. The future study will further analyze the performance of CNN with ensemble classifiers.

VI. REFERENCES

1. J. Schmidhuber, "Deep learning in neural networks: An overview", Neural Networks, 2015. **61**: pp. 85-117.
2. R.S. Ahmad, K.H. Mohamad, S.S. Liew and R. Bakhteri, "Convolutional neural network for face recognition with pose and illumination variation",

International Journal of Engineering and Technology (IJET), 2014. **6**(1): pp. 44-57.

3. B. Sahiner, H.P. Chan, N. Petrick, D. Wei, M.A. Helvie, D.D. Adler and M.M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images", IEEE Transactions on Medical Imaging, 1996, **15**(5): pp. 598-610.

4. J.L. Marzo i Lázaro, "Enhanced Convolution Approach for CAC in ATM Networks, An analytical study and implementation", 1997: Universitat de Girona.

5. B. Klein , L. Wolf and Y. Afek, "A dynamic convolutional layer for short range weather prediction", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4840-4848.

6. Y. LeCun, "Convolutional Neural Networks (LeNet) – DeepLearning 0.1", 2016. [Online], Available: http://deeplearning.net/tutorial/lenet. html, [Accrssed: 08 September 2016].

7. Y. LeCun, B. Léon, B.Yoshua and H. Patrick, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, 1998, **86**(11): pp. 2278-2324.

8. A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet classification with deep convolutional neural networks". Advances in Neural Information Processing Systems, 2012, pp. 1097-1105.

9. M.D. Zeiler and R. Fergus,"Visualizing and understanding convolutional network", European Conference on Computer Vision, 2014, pp. 818-833.

10. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

11. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

12. K. He, X. Zhang, S. Ren and J. Sun, "Identity mappings in deep residual networks", arXiv preprint arXiv:1603.05027, 2016.

13. S.D. Learning, "CS231n: Convolutional Neural Networks for Visual Recognition", 2016, [Online], Available: http://cs231n.stanford.edu/, [Accessed: 08 September 2016]

14. J. Ba, V. Mnih and K. Kavukcuoglu, "Multiple object recognition with visual attention", arXiv preprint arXiv:1412.7755, 2014.

15. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625-2634.

16. A. Dundar, J. Jin and E. Culurciello,"Convolutional Clustering for Unsupervised Learning", arXiv preprint arXiv:1511.06241, 2015.

17. I. Lenz, H. Lee and A. Saxena,"Deep learning for detecting robotic grasps". The International Journal of Robotics Research, 2015, **34**(4-5): pp. 705-724.

18. L. Romaszko, "A deep learning approach with an ensemble-based neural network classifier for black box ICML 2013 contest". Workshop on Challenges in Representation Learning, ICML, 2013, pp. 1-3.

19. Y. LeCun, C. Cortes and C. J.C. Burges,"The MNIST database of handwritten digits", 2016. [Online], Available: http://yann.lecun.com/exdb/mnist/, [Accessed: 08 September 2016]

20. L. Zhang, B. Verma and D. Stockwell, "Class-Semantic Color-Texture Textons for Vegetation Classification". International Conference on Neural Information Processing, Springer, 2015, pp. 354-362.