Improvement in OCR Technologies in Postal Industry Using CNN-RNN Architecture: Literature Review

P. Verma and G. M. Foomani

Abstract—**Convolutional Recurrent Neural Network (CRNN) based architecture is an attractive branch of Optical Character Recognition (OCR) studies. OCR is the process for transforming the image or the text obtained by scanning documents into machine-modifiable or editable format. It belongs to the domain of automatic identification of algorithms, modeled loosely after the animal brains, which are designed for pattern and character recognition. Hence, it falls under the *neural networks* category. An OCR system relies for the most part on the pre-processing, character/ image segmentation and feature extraction. This technology is an essential segment for automation processed in postal industry to read mail addresses and process mails. The first objective of this paper is to summarize the research that has been conducted in the field and further to present best in practice examples in this regard. Secondly, this research will also discuss about some gaps in the area and try to identify opportunities for future studies.**

Index Terms—**Automatic mail sorting, convolutional recurrent neural network (CRNN), neural networks, optical character recognition, postal industry.**

I. Introduction

Optical character recognition (OCR), is a task of converting scanned images of handwritten text (digits, letters and symbols) or into machine stamped, into a format readable by machines i.e. character streams [1].

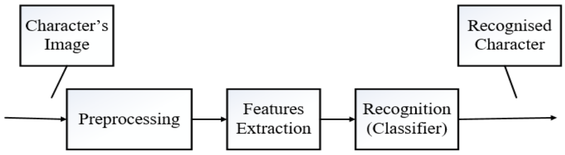


Fig. 1. Structure of OCR system [2].

The section below describes each of the steps, in detail, shown in Fig. 1.

a. *Image Acquisition*. To apprehend the image from an external source like a camera or a scanner.
b. *Preprocessing*. After the capturing of the image, preprocessing steps like noise removal using thresholding can be done to enhance the quality of the image [3].
c. *Feature Extraction*. This step helps in extracting geometrical features like loops, contours, corners, points and statistical features like moments. Based on these features, the characters are recognised.
d. *Character Segmentation*. In this process, the characters in the image are segregated into its constituent characters.
e. *Character Classification*. This step links the features of the segmented images to different categories or classes. [3], [4].

Trends in OCR

OCR is a widely used in variety of projects in various ranges from document scanning to label learning to creation of digital data archives [1] . The major trends being

- *Handwriting recognition*: This field has been extremely vigorous and susceptible of undergoing changes.
- Multi-language recognition
- Hand- written texts in forms or envelopes
- Postal envelopes and address readers on parcels [5]
- *Image Enhancement* OCR assists in automatically selecting and applying suitable filters to the document image to help better identify words and characters. [1]
- *Intelligent Post-Processing* OCR is of prime importance in creating robust information retrieval systems [1]
- The OCR system is also useful in *PIN Code recognition* and address validation. [1]
- OCR have been widely used for various applications like mail sorting, signature verification and bank cheque reading. Apart from that OCR has been used in passport validation, number plate recognition [4].

II. OCR in Postal Industry

Simplified depiction of how an OCR system fits into postal services is shown in the Fig. 2.

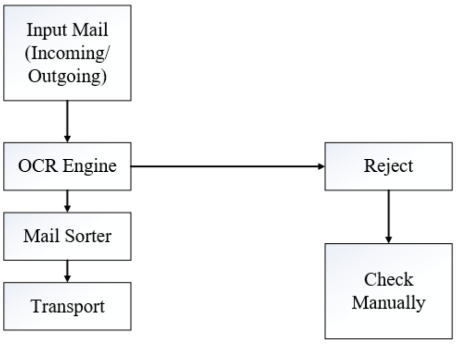


Fig. 2. OCR in Postal Services [3].

An OCR system is developed for the application of postal services in sequence to enhance the accuracy of mail sorting by recognising addresses from mail letters [3]. The OCR system is also useful in PIN Code recognition and address

Manuscript received April 20, 2021; revised March 4, 2022.

P. Verma is with the Canada Post Corporation, Ottawa, Ontario, Canada. She is with Global Innovative Campus, Canada (e-mail: palak.verma@canadapost.postescanada.ca, pverma0402@gmail.com)

G. M. Foomani is with Canada Post Corporation, Ottawa, Ontario,Canada. He is now with the Department of Civil Engineering, Concordia University, Montreal, Quebec, Canada (e-mail: matin.giahifoomani@mail.concordia.ca, Matin.Foomani@canadapost.postescanada.ca).

doi: 10.18178/ijmlc.2022.12.5.1095

validation [1].

Address Readers

The address readers in Postal industry helps to automatically sort out the mail i.e. the mail sorter pinpoints the destination address and reads the pin code in the address block. The resulting pin code is used to generate a barcode which is placed on the envelope. [6] And, one of the most significant example of an address reader, also used in United States Postal Service (USPS), is Multiline Optical Character Reader (MLOCR). The system can process up to 45,000 mail pieces per hour [1], [6].

III. Neural Networks

Datasets

1) *ImageNet*

ImageNet dataset, as shown in Fig. 3, has over 15 Million human labeled High-resolution images associated with 22,000 categories [7]. Such a large scale database of images is a crucial asset for developing advanced, large-scale content-based image search and understanding algorithms. Along with this, come up with various training and bench-marking data for such algorithms. ImageNet is built with the hierarchical structure provided by WordNet [8].

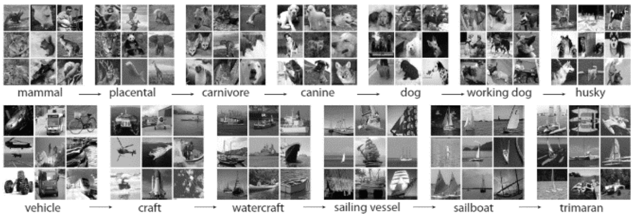


Fig. 3. ImageNet dataset [8].

2) *MNIST*

MNIST dataset, as shown in Fig. 4, consists of collection of handwritten digits of 0-9, widely used in optical character recognition and research in the field of machine learning, pattern recognition. [9]. The database consists of 70,000 handwritten digits out of which there are 60,000 training images and 10,000 test images within the dataset. [10]. The images are gray-scale images with 28×28 pixels [11] (width × height).

Fig. 4. Example of MNIST dataset [10].

3) *CIFAR-10/100*

CIFAR-10 Dataset, is an organised subset of the 80 million tiny images dataset [12] which, consists of 60,000 colored images of 32×32 pixels in 10 classes. Out of which, there are, 50,000 training images and 10,000 testing images. [11]. An example of CIFAR-10 Dataset is shown in Fig. 5.

On the other hand, CIFAR-100 has 100 classes of images in the same pattern as CIFAR-10 [11].

Fig. 5. Example of CIFAR-10 dataset [13].

IV. Convolutional Neural Network (CNN)

CNN is a hierarchical, multi-layer neural network trained with the back-propagation algorithm [14] . It has emerged as a powerful tool for object detection, face recognition, natural language processing, pattern recognition, spam detection, speech recognition, image classification, EEG signal classification. Table I below, depicts the existing research work on CNN.

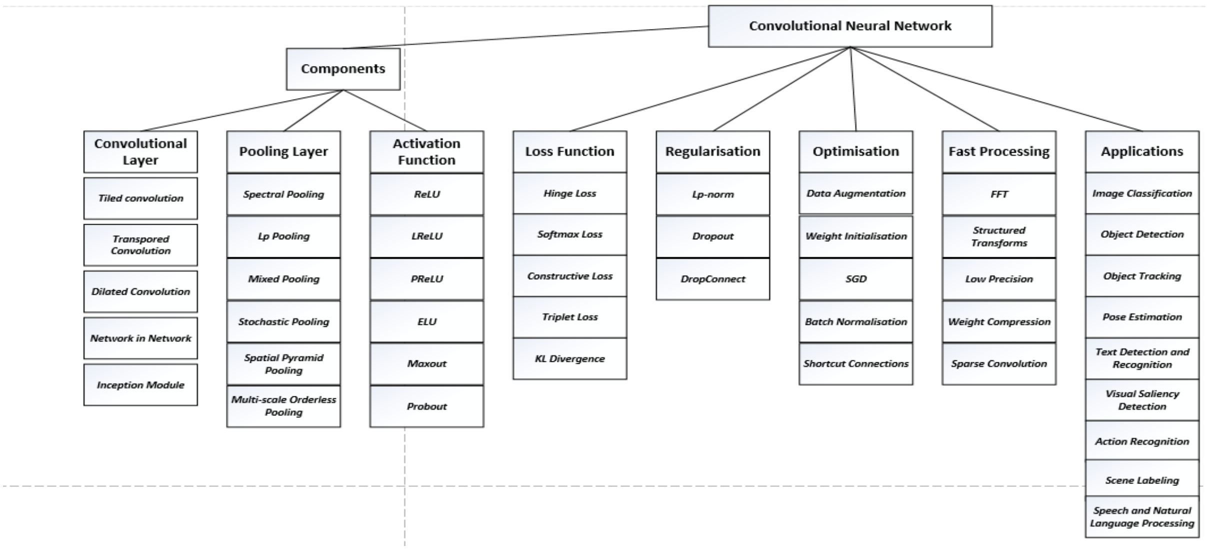


Fig. 6. Hierarchy of CNN [15].

There are many databases that are used to train CNN. And, the selection of databases is on the basis of what type of character classification is required. The hierarchy of the CNN is shown in Fig. 6.

For instance, *Modified Institute of Standards and Technology (MNIST) and CIFAR-100* are used in handwritten digit recognition problems. And, deep convolutional Neural Networks are used to classify high-resolution images in the *ImageNet* dataset.

Architecture

The CNN is composed of multiple layers namely, *Convolutional Layers, Pooling Layers, Fully-connected layers* as shown in Fig. 7. The model takes an input image and passes it to the convolutional layer. Each CONV layer is constituted of high-dimensional convolutions kernels as shown in Fig. 8. And, produces a consecutively higher-level extraction of the input data called as the *feature map*. The feature map stores important and unique information.

Altogether, each neuron of a feature map is affixed to a domain of adjacent neurons in the previous layer. The feature maps are computed by initially convolving the input with the learned kernel and then on the convolved results perform an element-wise nonlinear activation function.

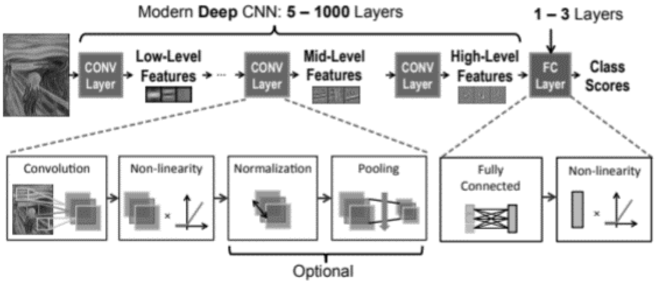


Fig. 7. Convolutional neural network [16]

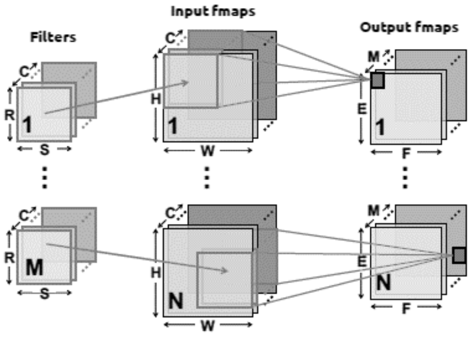


Fig. 8. Convolutions in CNN [16].

Mathematically, as stated in [15]

Feature value at (i,j) in the k^{th} feature map of the l^{th} layer $z^l_{i,j,k}$ is calculated as-

$$z^l_{i,j,k} = W_k^{l^T} \times V^l_{i,j} + b^l_k$$

Where, $w^l_k\ and\ b^l_k$ are the weight and the bias of the k^{th} filter and the l^{th} layer. And, $V^l_{i,j}$ is the input vector at location (i,j) of the l^{th} layer.

The thing to be considered here, the most, is that the weights that generate the feature maps are shared. This helps in reducing the model complexity and easily trainable model.

A classic model of the CNN architecture is shown in Fig. 9.

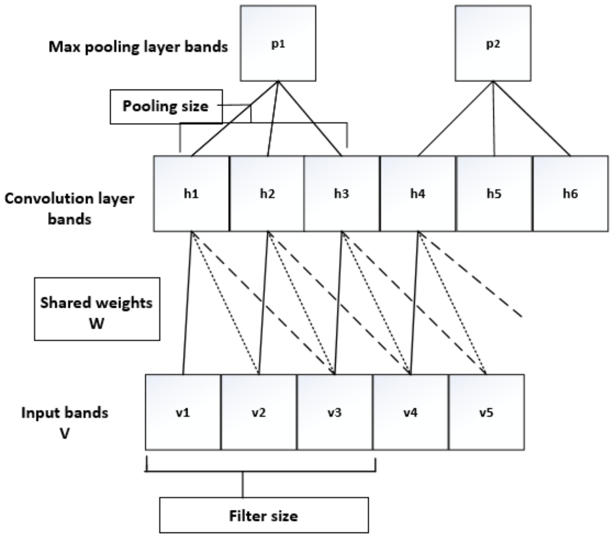


Fig. 9. Architecture of convolutional neural network [17].

It has a hidden nonlinear activation function, $h(\cdot)$ which helps in introducing some nonlinearities to CNN. This helps to detect nonlinear features. The activation value can be computed as –

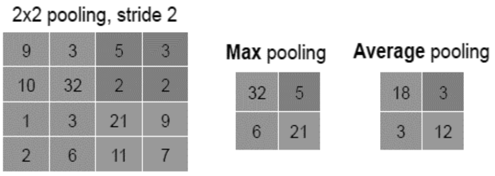
$$h^l_{i,j,k} = h(z^l_{i,j,k})$$

Typical activation functions that are used are sigmoid, Hyperbolic Tangent (tanh) and Rectified Linear Unit (ReLU).

The next layer is the pooling layer which has an objective to attain the shift-variance. It does so by, reducing the resolution of the feature maps. It is used to remove the variability in the hidden units existing due to distortions, noise etc [17] . Each feature map of a pooling layer is attached to its adjacent feature map of the succeeding convolutional layer. The pooling function is denoted by $pool(\cdot)$. For each feature,

$$y^l_{i,j,k} = pool\big(h^l_{m,n,k}\big), \forall(m,n) \in R_{ij}$$

where, R_{ij} is the local neighborhood around the (i,j) location.The Pools are a set of values in its receptive fields into smaller number of values. There are two types of pooling operations: average pooling and max-pooling. As shown in Fig. 10.


2x2 pooling, stride 2

9	3	5	3
10	32	2	2
1	3	21	9
2	6	11	7

Max pooling

32	5
6	21

Average pooling

18	3
3	12

Fig. 10. Forms of pooling [16].

After, several convolutional and pooling layers we have one or more than one fully-connected layers which help in performing high-level reasoning. In order to generate global semantic information all the neurons from the previous layer are connected to neurons in the current layer.

And, the final layer is the output layer. The commonly used methods are softmax operator and SVM for solving the different classification tasks. Supposedly, we have N input-output relations and $(x^{(n)}, y^{(n)}); n \in [1 \dots. N]$, where, $x^{(n)}$ *and* $y^{(n)}$ are the corresponding n –*th* input data and its label. And, $o^{(n)}$ is the output of CNN.

Therefore, the loss can be calculated as,

$$L = \frac{1}{N}\sum_{n=1}^{N} l(\theta;\ y^{(n)}, o^{(n)})$$

And, θ denotes all the parameters of CNN like weight and bias. By minimising the loss function we can attain the best set of parameters.

Stochastic Gradient Descent (SGD) is the most widely used solution for optimising the CNN network. [15]

The use of CNNs to achieve exceptional improvement in object recognition for ImageNet is the most crucial result.

TABLE I: EXISTING RESEARCH WORK - CNN

Author/ Year	Reference	Theoretical/ Conceptual Framework	Conclusions
F. Siddique, S. Sakib and M. A. B. Siddique (2019)	[18]	This paper observes the disparities between the accuracies of CNN to classify handwritten digits using various numbers of hidden layers and epochs and to then, make the comparison between the accuracies. MNIST dataset is used to perform the experiment. And, further, the network is trained using a backpropagation algorithm and a stochastic gradient descent.	The accuracies for handwritten digits were observed for 15 epochs by differing the hidden layers. The maximum and minimum accuracies were observed with a batch size of 100. In one of the cases, the accuracy obtained was pretty high. Such a high accuracy will contribute in increasing the performance speed. And, in another case, total lowest test loss was found. And, this loss helps to achieve better image resolution and noise processing.
M. C. G. Neri, J. H. S. Azuela, V. G. C. S ánchez, O. O. V. Villegas and M. Nandayapa (2020)	[19]	With a variety of handwriting style, handwritten character recognition has become a difficult challenge. In this paper, 2 handwritten digits dataset from MNIST dataset is presented for testing classification and, they are used to train a CNN to classify. And, thereafter, a comparison of the performance of the 2 datasets is presented.	The results depicted that it is feasible to achieve higher accuracies with both datasets. The research also depicted that by using the basic image preprocessing techniques to one of the data set helps in increasing the recognition accuracy as compared to the non-preprocessed dataset.
E. Rusakov, S. Sudholt, F. Wolf and G. A. Fink (2018)	[20]	CNNs have taken over the word spotting to recover the portions of the images in the document relevant to user-defined query. This paper talks about the complexity of CNN required to have a successful word spotting. And, also answers to can the same results be obtained with a less complex and much smaller CNN.	The experiments conducted in this research shows that the deeper CNN architecture does not guarantee better performance. Neither increasing the number of parameters nor increasing the depth achieves higher nor a better performance.
H. Di and G. Alregib (2018)	[21]	The study uses the CNN for seismic fault detection. This architecture includes three major components: training image preparation, CNN classifier training and finally volumetric processing.	The research proves, that the trained CNN classifier is able to learn the targeted seismic features accurately. The network is also capable of connecting the seismic images with the target faults thereby, saving efforts in selecting and generating attributes.
B. Kayalibay, G. Jensen and P. v. d. Smagt (2017)	[22]	This work, a three-dimensional filters CNN- based method is applied to the brain and hand MRI. This work validates on data both from the central nervous system as well as, the bones in the hands.	In this work multiple segmentation maps, at different scales, were combined which led to speeding up the convergence. Despite the scarcity of the labelled medical images, a 3D trained CNN network achieves good quality results.
D. Bertero, Y. Wan, P. Fung, R. H. Y. Chan, C.-S. Wu and F. B. Siddique (2016)	[23]	An interactive dialogue system is implemented to recognise user emotions and sentiments in real time using CNN. A different, CNN-based sentiment model is analysed that recognises the sentiments from speech recognition.	The paper displays the use of CNN with a single filter at a high accuracy, on a six emotion categories, directly from the time-domain audio input. A sentiment analysis on the human-machine dialogue by using a keyword based method proved to have a slightly better accuracy and precision but, fell drastically on a larger dataset. Hence, some improvements are suggested for future training.
I. Rocco, R. Arandjelovic and J. Sivic (2017)	[24]	The CNN architecture proposed in this work is based on three major aspects: 1. Feature extraction, matching and simultaneous inlier detection and parameter estimation. 2. Analysing if these parameters can be trained without manual annotation. 3. The model performs well on a challenging dataset.	The fully trainable network architecture achieves state-of-the-art results for the category-level matching.

V. RECURRENT NEURAL NETWORKS (RNN)

RNN are kind of neural networks, which can redirect the feedback signals to form a directed cycle. [25] Feed- Forward Neural Networks [26] with recurrent connections, i.e. to feed signals from the previous time stamps, are known as Recurrent Neural Networks (RNN). A simple RNN structure is shown in Fig. 11.

Table II depict the existing research work on the Recurrent Neural Networks.

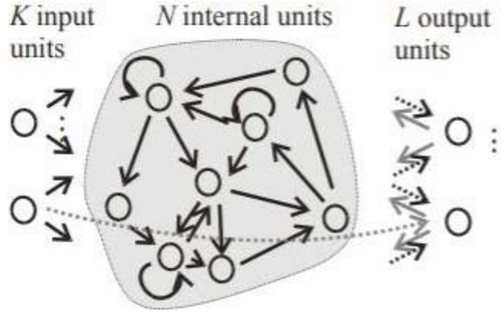


Fig. 11. The basic network structure of RNN [27].

Mathematically, they are the dynamic systems [28]. They consist of elementary building blocks called neurons with one or more feedback loop. These loops are recurrent cycles over time or sequence. These neurons are connected using synaptic links whose, strength is determined by the weights. [27] (As shown in Fig. 12.) The network consists of hidden states h_t which act as the memory of the whole network. It is dependent on the current observation $x_{t'}$ and also, on the previous hidden state h_{t-1}

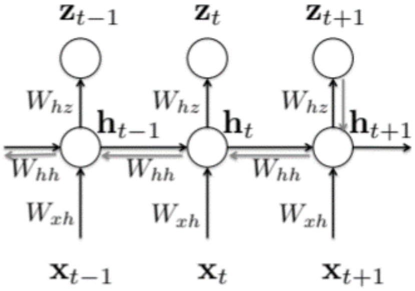


Fig. 12. Unit Structure of LSTM [25].

Therefore, we can represent h_t as-

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

Eq. (1) shows the recursive nature of RNN. Where, $f\ is\ the\ non-linear\ mapping$. And,

$$f \in \{sigm, tanh\}$$

Supposedly, we use $f\ as\ tanh$

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Long Short- Term Memory LSTM

RNNs can look back in time for approximately ten steps. This is because the fed back signal is vanishing in short gradient vanishing problem. Long Short-Term Memory Recurrent Neural Networks help in overcoming these problems. They help in introducing memory information and also, handle the long sequences in a better way. They are capable of avoiding long-term dependency problems.

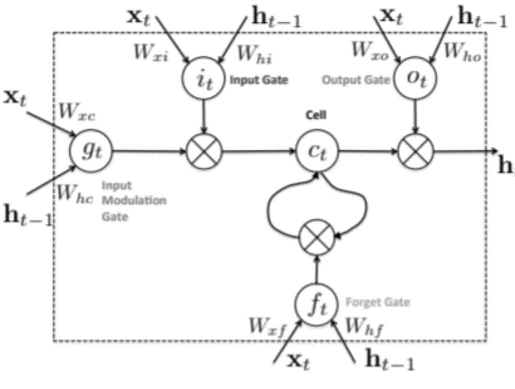


Fig. 13: Network with time lags [25].

In Fig. 13, the center of the LSTM unit is a memory block which contains memory cell $\boldsymbol{c_t}$ which, helps in encoding the information of the input that was observed upto that step.

Each memory block contains *an input gate, forget gate and output gate.* The input gate controls the input activations into the memory cell. While, the output gate controls the flow of the output activations into the rest of the network. The forget gate is to prevent the network from processing continuous input streams into subsequences.

The memory cells have the same inputs $(h_{t-1}) and\ x_t$ and output h_t. The output of the LSTM network can be shut via the output gate. [25]

Supposedly, let the sequence data be$\{x_{1......}x_T\}$. Then the gate definitions are -

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$
$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$
$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot g_t$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$
$$h_t = o_t \cdot tanh(c_t)$$

Output of the likelihood with softmax function-

$$z_t = softmax(W_{hz}h_t + b_z)$$

where, f_t is the forget gate; i_t is the input gate; o_t is the output gate; and g_t is the input moldulation gate.

By minimising the least square -$\frac{1}{2}(y_t - z_t)^2$ and we can take the derivative w.r.t $z_t\ and\ W_{hz}$

$$dz_t = (y_t - z_t)$$
$$dW_{hz} = \sum_t h_t dz_t$$
$$dh_T = W_{hz}d_{zT}$$

where, the gradient is only considered at the last time step T

And, the LSTM back propagate at current time step ***t*-**

$$do_t = tanh(c_t)dh_t$$
$$dc_t = (1 - tanh(c_t\,)^2)o_t dh_t$$
$$df_t = c_{t-1}dc_t$$
$$dc_{t-1} \mathrel{+}= f_t \cdot dc_t$$
$$di_t = g_t dc_t$$
$$dg_t = i_t dc_t$$

Activation functions over the whole sequence, and the weights are shared over the whole sequence-

$$dW_{xo} = \sum_t o_t(1 - o_t)x_t do_t$$
$$dW_{xi} = \sum_t i_t(1 - i_t)x_t di_t$$
$$dW_{xf} = \sum_t f_t(1 - f_t)x_t df_t$$
$$dW_{xc} = \sum_t (1 - {g_t}^2)x_t dg_t$$

Thus, we take the summation over *t* –

$$dW_{ho} = \sum_t o_t(1 - o_t)h_{t-1}do_t$$
$$dW_{hi} = \sum_t i_t(1 - i_t)h_{t-1}di_t$$
$$dW_{hf} = \sum_t f_t(1 - f_t)h_{t-1}df_t$$
$$dW_{hc} = \sum_t (1 - {g_t}^2)h_{t-1}dg_t$$

And, hidden at the current time stamp *t-1*. The source to derive dh_{t-1} is from the activation function-

$$dh_{t-1} = o_t(1-o_t)W_{ho}do_t + i_t(1-i_t)W_{hi}di_t + f_t(1-f_t)W_{hf}df_t + (1-{g_t}^2)W_{hc}dg_t$$

The source to derive dh_{t-1} is from the objective function-

$$dh_{t-1} = dh_{t-1} + W_{hz}dz_{t-1}$$

The least square objective function-

$$\pounds(x,\theta) = \min \sum_t \frac{1}{2}(y_t - z_t)^2$$

where,

$$\theta = [W_{hz}, W_{xo}, W_{xi}, W_{xf}, W_{xc}, W_{ho}, W_{hi}, W_{hf}, W_{hc}]$$

with ignoring the biases.

At the *T* time stamp, we take the derivative w.r.t. c_T

$$\frac{\partial \pounds(T)}{\partial c_T} = \frac{\partial \pounds(T)}{\partial h_T}\frac{\partial h_T}{\partial c_T}$$

At the time stamp *T-1*, we take the derivative of $\pounds(T-1)$ w.r.t c_{T-1} as,

$$\frac{\partial \pounds(T-1)}{\partial c_{T-1}} = \frac{\partial \pounds(T-1)}{\partial h_{T-1}}\frac{\partial h_{T-1}}{\partial c_{T-1}}$$

The error is back-propagated via $\pounds(T-1)$ but, also from c_{T-1}

$$\frac{\partial \pounds(T-1)}{\partial c_{T-1}} = \frac{\partial \pounds(T-1)}{\partial c_{T-1}} + \frac{\partial \pounds(T)}{\partial c_{T-1}}$$

And, finally if we use Stochastic Gradient Descent (SGD) then,

$$\theta = \theta - ŋd\theta$$

where, ŋ $is\ the\ learning\ rate.$

TABLE II: MATRIX OF EXISTING RESEARCH WORK-RNN

Author/Year	Reference	Theoretical/Conceptual Framework	Conclusions
W. Yin, K. Kann and H. Schutze (2017)	[29]	CNN and RNN are explored to handle various Natural Language Processing (NLP) tasks. This work compares CNNs, LSTMs and GRUs on NLP tasks like sentiment/relation classification, textual entailment, answer selection, matching of question related.	The experiments support the following findings: Learning rate changes the performance effortlessly. And, CNNs and RNNs provide information for text classification. The author concluded that RNNs perform well and robust in a wide range of tasks.
H. Sak, A. Senior and F. Beaufays (2014)	[30]	Long Short-Term Memory (LSTM) is a specific RNN outlined for temporal sequence and long range dependencies. In this paper, author explores the 2 layer LSTM RNN architecture for large scale acoustic modeling in speech recognition using distributed training.	The author concludes by saying that LSTM RNN framework achieves state-of-the-art results. The author showed that these models can be quickly trained using distributed training. Also, the model effectively uses the parameters by addressing the efficiency for training large networks.
H. Sak, A. Senior and F. Beaufays	[31]	The new LSTM based RNN architectures are presented which effectively uses the model parameters to train the framework for large vocabulary speech recognition with a large number of context dependent states. In this research paper, we introduce two architectures: 1. Initiating a recurrent projection layer between the LSTM layer and the output layer. 2. Introducing another non-recurrent projection layer in order to increase the projection layer size.	The proposed framework proves the improvement in the performance and flexibility than any other DNNs with a large number of output states.
C. Che, C. Xiao, J. Liang, B. Jin, J. Zhou and F. Wang	[26]	RNN architecture with Dynamic Matching Temporal patterns in the patient sequences is proposed for Personalised Predictions of Parkinson’s Disease. It learns the similarities between the two patient records sequences.	The author confirms with this model on the real-world patients demonstrates promising utility and efficacy. The architecture has a large performance gain.

VI. PROPOSED SYSTEM- CRNN

The hybrid version of *Convolutional and Recurrent Neural Networks* is known as *CRNN*. Both the CNN and RNN layer can be added to the multiple layer architecture. The RNN layer can extract the continuous dependency features from the output of the CNN layer [32].

CRNN is widely used in variety of projects, as depicted in Table III, few of them are: *music classification* [33], *recognition method for bank card number* [34], *audio event detection* [35].

The CRNN architecture is also being used in the medical industry. In [36], the author addresses the problems associated with *Big Data for medical care*. They put forward a data collection method, which adopts OCR recognition algorithm on the basis of CRNN, together with human assisted man-machine recognition method.

The proposed architecture of CRNN in [37], helps in reconstructing high quality dynamic cardiac *MR Images*. The framework roots the structure of traditional iterative algorithms, efficiently modelling the recurrence of the iterative reconstruction stages by using recurrent hidden connections across each iteration step. And further, incorporating bidirectional convolutional recurrent units emerge over time to make use of the temporal dependency of the sequence.

In [38] and [39] the author addresses the challenge faced in *text detection and recognition in natural scene images*. In [39] the main contribution is to present a method that directly transcribes scene text image to text without the use of sophisticated character segmentation.

In the network structure shown in Fig. 14, an end-to-end text spotting architecture, which together acknowledges and identifies words in natural scene images. The architecture shown in [38] consists of a number of convolutional layers, modified by VGG-16 Net [40], a Region-of-Interest, a Recurrent Neural Network, a MLP for regression and a RNN-based decode for word recognition.

In [32], the authors propose CRNN mixed model for Image Classification. This paper uses RNN to calculate the Dependency and Continuity Features of the Intermediate Layer Output of the CNN Model, connect the characteristics of these middle tiers to the final full-connection network for classification prediction.

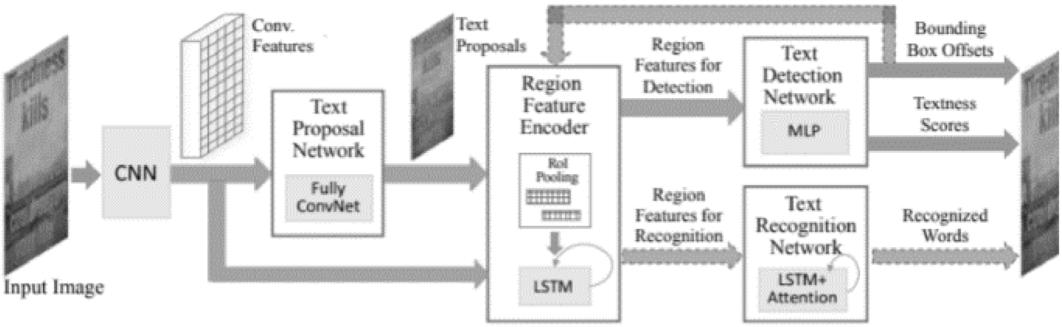


Fig. 14. Network Model [38].

The CNN-RNN model, as shown in Fig. 15, uses the RNN for calculating the Dependency and Continuity features of the middle layer of the CNN model as shown in Fig. 14.

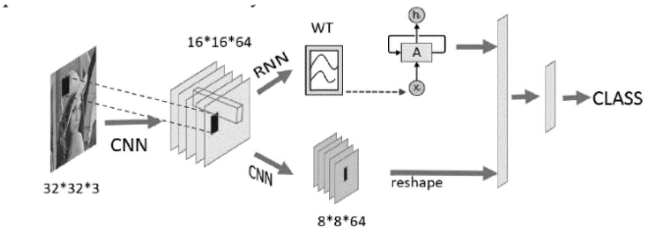


Fig. 15. CNN-RNN Model [32].

For a better understanding, a basic 1-Layer RNN structure is shown in Fig. 16. [32]

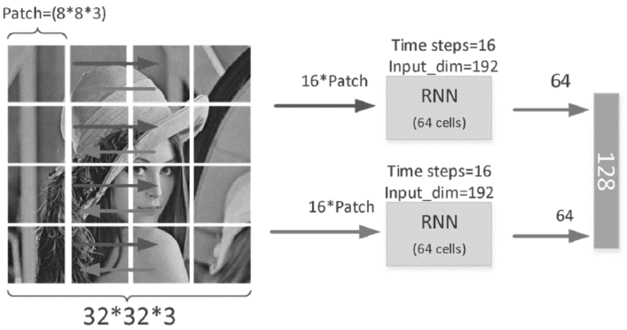


Fig. 16. 1-Layer RNN [32].

Input Data of the RNN layer is proposed as $x \in R^{w*h*c}$ where, w and h denote the width and height of the image inputted, x, and, c represents the number of color channels.

The convolution is the process of proceeding with the characteristics of the various distinct frequency bands of the images. The output of convolutional layer is the result of the texture, edge and shape of the input image i.e. the original image after the filtering process. Filtering process, is the sum of the pixels of the image and the filter. As seen from the image, the patches (the identification area) can be split into $M \times N$ patches where, $x_{m,n} \in R^{w_p \times h_p \times c}$ where,

$M = \frac{w}{w_p}$ and $N = \frac{h}{h_p}$ and, w_p and h_p are the width and height of the patches.

When we input $x_{m,n}$ into a bidirectional RNN network in a sequence gives-

After the deconvolution process,

$$S_{i,j}^{F} = f_{fwd}\left(S_{m,n-1}^{F}, x_{m,n}\right) for\ n = 1 \ldots\ldots N$$
$$S_{i,j}^{R} = f_{rev}\left(S_{m,n+1}^{F}, x_{m,n}\right) for\ n = N \ldots\ldots 1$$

where, f_{fwd} and f_{rev} return the forward and reverse sequence, recursively. They strive to return the final state result.

We can use the Fourier transform to utilise the Image wave which is, a non-periodic form of discrete signal in the time domain. But, Fourier Transform doesn't distinguish the characteristics of the signal in the time domain.

The transfer of the spatial domain to the frequency domain is given as-

$$F(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)e^{-i(ux+uy)}\,dudv$$

where, $f(x,y)$ is the 2- dimensional function of the image.

Because, of the limitations of the Fourier Transform, wavelet transform could also be used, as shown in Fig. 14. ,

$$WT(\alpha,\tau) = \frac{1}{\sqrt{\alpha}}\int_{-\infty}^{+\infty} f(t) * \varphi\left(\frac{t-\tau}{a}\right)dt$$

where, φ denotes the wavelet basis.

TABLE III: MATRIX OF LITERATURE REVIEWS OF CRNN

Author/ Year	Reference	Theoretical/ Conceptual Framework	Conclusions	Implications for Future Research
B. Zhao, X. Li, X. Lu and Z. Wang (2018)	[41]	The authors talk about the CRNN model for *Weather Prediction*. A CNN model is extended with a channel-wise attention model to extract correlated features and RNN is used to excavate the dependencies among weather classes.	Their overall results depicted that the architecture performs well in most of the cases, but sometimes fails when the weather cues are not obvious.	Introduction to a prediction task that can classify images with multi labels and also describes the weather conditions more comprehensively.

X. Fu, E. Ch'ng, U. Aickelin and S. See (2017)	[42]	The authors discuss the use of CRNN for *Redundancy Detection*. Character aware CNN and RNN are incorporated. Salient features from Char-CNN is selected and fed to Char-RNN that learns long sequence semantics via update mechanism.	The authors conclude that the CRNN architecture enhances the detection accuracy. And, also obtains best scores on precision rate and recall ratio.	In the future, they hope to extend the model to a variety of data for generic redundancy detection framework.
Y. Xin, Y. Lin, P. Shi, S. Han and B. Tian (2017)	[43]	The authors talk about Vehicle License Plate Recognition using ConvNet-RNN. The author states that the CNN-RNN framework incorporates the advantages of feature and label sequences	The architecture was evaluated on Malaysian Vehicle License Plate dataset and it demonstrated higher accuracy and a better performance.	The authors propose using LSTM module instead of RNN module such that it is better able to remember long term dependencies. Also, improve the models that simulate Gaussian noise and translations present in the real-world data.
G. Keren and B. Schuller (2016)	[44]	In this paper, authors propose a model that enhances the Feature Extraction process, from convolutional layers, for the case of sequential data, by feeding patches of the data into RNN and using the outputs to compute the extracted features. By doing so, it is exploited that a window containing a few frames of the sequential data is a sequence itself and this might encapsulate valuable information.	The model proposed yield improvements in classification results as compared to traditional convolutional layers for the same number of extracted features.	The authors propose using CRNN model for larger models and various other applications.
Z. Xie, Z. Sun, L. Jin, H. Ni and T. Lyons (2017)	[45]	In this research, a novel solution, multi-spatial-context fully convolutional recurrent network, is proposed to address the problem of Online Handwritten Text Recognition in order to achieve strong robustness and high accuracy.	In the experiments, the architecture outperforms the existing methods and thereafter, reducing the error and therefore, improving the performance of the system.	In the future, the authors plan to incorporate the over- segmentation strategy as preceding knowledge to improve the recognition performance. And, also want to investigate the challenge of out-of-vocabulary issue.
M. Zihlmann, D. Perekrestenko and T. Michael (2017)	[46]	The authors evaluated two deep neural networks for ECG Classification: 1. CNN and 2. CNN combined with LSTM layers for temporal aggregation of features. Both the architectures work on the data processing of the input, stack of Convolutional layers for feature extraction, temporal aggregation of features across time using averaging in CNN and a use of bidirectional LSTM in the CRNN architecture and a linear classifier.	The results shown in the research, depict as CRNN has more parameters, and therefore, is a higher capacity model than the conventional CNN. CRNN has been proven to have higher accuracy than CNN.	The authors suggest using the architecture with different pathology and also, refining and extending data augmentation scheme i.e. taking actual heart rate for resampling rather than fixing the heart rate.
Yaguang Li, Rose Yu, Cyrus Shahabi, Yan Liu (2018)	[47]	Traffic Forecasting have the following challenges: 1. Complexity in spatial dependency on the networks of road. 2. Trouble in long-term forecasting. 3. Non-Linear temporal dynamics with the changing road conditions. The authors use diffusion Convolutional Recurrent Neural Network that addresses both the spatial and temporal dependencies in the flow of traffic.	A random walk bidirectional graph is used for the spatial component and recurrent neural networks are used for seizing the temporal dynamics. And, additional sampling techniques and encoder-decoder architecture are used for long-term forecasting. And, the authors observed that this model has a 12-15% of improvement than baselines.	The authors suggest of using this model for other forecasting tasks that involve spatio-temporal dependency. And, also further proposes in improvement of the graph structure in the model for various moving objects.
K. Liang, N. Qin, L. Ma, A. H. Kemp and D. Huang (2018)	[48] [49]	CRNN is used to diagnose various faults in a High Speed Train (HST) Bogie in a timely manner. The framework of CRNN helps in extracting features from the bogie signals and further passing them to the stacked recurrent layers to obtain hidden features with time series correlation. And further these hidden features are used to calculate the probability of signal classification.	The authors concluded in his research that CRNN has a higher accuracy than the conventional model structure like 1D- CNN and LSTM. It also reduces the time spent during the training. Thereby, inferring that CRNN is much more efficient and time convenient as compared to other models.	The authors suggest further analysis on the results. CRNN further used to detect early dangers of HST by facing the deterioration law of fault states. Authors also suggest, using CRNN methods to solve pattern recognition of the measured and systematic deterioration of the key components that occur during the various operations of HST.

VII. Conclusions and Future Works

In this research paper, a factual and pragmatic study of different applications and assessments has been introduced. This research provides and presents important understanding on the association among the existing algorithms and points on the subsequent and future researches.

The proposed literature review explains and collaborates the usage of the prevailing machine learning architecture of CNN-RNN in various applications and further, describes the gap between utilising this artificial intelligence technology in the postal industry.

The future work specifically will focus on the utilisation of this algorithm CRNN in the postal industry.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contributions

Verma. P conceived and designed the work; Verma. P collected the information and relevant data; all authors drafted and wrote the paper; Foomani G.M guided and reviewed the presentation of the work; Foomani G.M approved the collected information and data; all authors did the critical revision of the paper; all authors approved the final version.

REFERENCES

[1] E. Borovikov, "A survey of modern optical character recognition techniques," arXiv:1412.4183v1, Lanham, 2004.

[2] S. Kumar, N. Sahu, A. Deep, K. Gavel and R. Ghost, "Offline handwritting character recognition (for use of medical purpose) using neural networks," *International Journal of Engineering and Computer Science*, vol. 5, no. 10, pp. 18612-18615, 2016.

[3] Y. M. Alginaih and A. A. Siddiqi, "Multistage hybrid arabic/ indian numeral OCR system," *International Journal of Computer Science and Information Securtiy (IJCSIS),* vol. 8, no. 1, 2010.

[4] N. Islam, Z. Islam and N. Noor, "A survey on optical character recognition system," *Journal of Information and Communication Technology- JICT,* vol. 10, no. 2, December 2016.

[5] A. Kornai, "An experimental HMM-Based postal OCR system," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Los Alamitos, 1997.

[6] N. S. Srihari, A. Shekhawat, and W. S. Lam, *Optical Character Recongition (OCR)*, 2003.

[7] F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates Inc., pp. 1097-1105, 2012.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[9] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 141-142, 2012.

[10] B. Kessab, C. Daoui, B. Bouikhalene, M. Fakir, and K. Moro, "Extraction method of handwritten digit recognition tested on the MNIST database," *International Journal of Advanced Science and Technology,* vol. 50, pp. 99-110, 2013.

[11] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville and Y. Bengio, "ReNet: A recurrent neural network based alternative to convolutional networks," arXiv preprint arXiv:1505.00393, 2015.

[13] S. McCann and J. Reesman. Object detection using convolutional neural network. [Online]. Available: http://cs229.stanford.edu/proj2013/ReesmanMcCann-Vehicle%20Detection.pdf

[14] M. Elleuch, R. Maalej, and M. Kherallah, "A new design based- SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition," *Procedia Computer Science,* vol. 80, pp. 1712-1723, 2016.

[15] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural network," *Pattern Recognition,* vol. 77, pp. 354-377, 2018.

[16] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE,* vol. 105, no. 12, pp. 2295-2329.

[17] T. N. Sainath, A.-R. Mohammed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[18] F. Siddique, S. Sakib, and M. A. B. Siddique, "Recognition of handwritten digit using convolutional neural network in python with tensorflow and comparison of performance for various hidden layers," in *Proc. 5th International Conference on Advances in Electrical Engineering (ICAEE)*, Dhaka, 2019.

[19] M. C. G. Neri, J. H. S. Azuela, V. G. C. Sánchez, O. O. V. Villegas, and M. Nandayapa, "A convolutional neural network for handwritten digit recognition," *International Journal of Combinatorial Optimisation Problems and Informatics*, vol. 11, no. 1, pp. 97-105, 2020.

[20] E. Rusakov, S. Sudholt, F. Wolf, and G. A. Fink, "Exploring architectures for cnn-based word spotting," *Dortmund*, 2018.

[21] H. Di and G. Alregib, "Seismic fault detection from post-stack amplitude by convolutional neural networks," in *Proc. 80th EAGE COnference and Exhibition 2018*, Copenhagen, 2018.

[22] B. Kayalibay, G. Jensen, and P. V. D. Smagt, "CNN-based segmentation of medical imaging data," arXiv preprint arXiv: 1701.03056, 2017.

[23] D. Bertero, Y. Wan, P. Fung, R. H. Y. Chan, C.-S. Wu, and F. B. Siddique, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *Proc. the 2016 Conference on Empirical Methods in Natural Language Processing,* 2016.

[24] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] G. Chen, "A gentle tutorial of recurrent neural network with error backpropagation," arXiv preprint arXiv: 1610.02583, 2016.

[26] C. Che, C. Xiao, J. Liang, B. Jin, J. Zhou, and F. Wang, "An RNN architecture with dynamic temporal matching for personalised predictions of parkinson's disease," in *Proc. the 2017 SIAM International Conference on Data Mining*, 2017.

[27] H. Jaeger. (2002). A tutorial on training recurrent neural networks, covering BPPT,RTRL, EKF and the "echo state network" approach. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.378.4095&rep=rep1&type=pdf

[28] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM – a tutorial into long short-term memory recurrent neural networks," arXiv:1909.09586v1, 2019.

[29] W. Yin, K. Kann and H. Schutze. (2017). Comparative study of CNN and RNN for natural language processing. [Online]. Available: https://arxiv.org/abs/1702.01923

[30] H. Sak, A. Senior, and F. Beaufays. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Google. [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43905.pdf

[31] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," arXiv preprint arXiv: 1402.1128, 2014.

[32] Q. Yin, R. Zhang, and X. Shao, "CNN and RNN mixed model for image classification," in *Proc. MATEC Web of Conferences*, 2019.

[33] K. Choi, F. Gyorgy, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* New Orleans, 2017.

[34] Y. Xin, Y. Lin, P. Shi, S. Han, and B. Tian, "An automated recognition method for bank card number," *Journal of Physics Conference Series 1345:022049,* 2019.

[35] F. Jia , C. Shi, Y. Wang, C. Wang, and B. Xiao, "Grayscale-projection based optimal character segmentation for camera -captured faint text recognition," in *Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* Kyoto, 2017.

[36] L. Zhao and K. Jia, "Application of CRNN based OCR in health records system," in *Proc. the 3rd International Conference on Multimedia Systems and Signal Processing*, New York, 2018.

[37] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural network for dynamic mr image reconstruction," *IEEE transactions on Medical Imaging,* vol. 38, no. 1, pp. 280-290, 2018.

[38] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. the IEEE International Conference on Computer Vision*, 2017.

[39] G. Qiang, T. Dan, L. Guohui, and L. Jun. (2016). Memory matters: Convolutional recurrent neural network for scene text recognition. [Online]. Available: https://arxiv.org/abs/1601.01100

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. The International Conference on Learning Representations*, 2015.

[41] B. Zhao, X. Li, X. Lu, and Z. Wang, "A CNN–RNN architecture for multi-label weather recognition," *Neurocomputing,* vol. 322, pp. 47-57, December 2018.

[42] X. Fu, E. Ch'ng, U. Aickelin, and S. See, "CRNN: A Joint Neural Network for Redundancy Detection," in *Proc. IEEE International Conference on Smart Computing (SMARTCOMP)*, Hong Kong, 2017.

[43] T. K. Cheang, Y. S. Chong, and Y. H. Tay, "Segmentation-free vehicle license plate recognition using convNet-RNN," arXiv preprint arXiv:1701.06439, 2017.

[44] G. Keren and B. Schuller, "Convoltuional RNN: An enhanced model for extracting features from sequential data," in *Proc. 2016 International Joint Conference on Neural Networks (IJCNN)*, 2016.

[45] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, *Learning Spatial-Semantic Context with Fully Convolutional Recurrent Network for Online Handwritten Chinese Text Recognition,* 2017.

[46] M. Zihlmann, D. Perekrestenko, and T. Michael, "Convolutional recurrent neural networks for electrocardiogram classification," *Computing in Cardiology (CinC),* Rennes, 2017.

[47] Y. Li, R. Yu, S. Cyrus and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-driven traffic forecasting," arXiv preprint arXiv: 1707.01926, 2017.

[48] K. Liang, N. Qin, L. Ma, A. H. Kemp, and D. Huang, "Multiple convolutional recurrent neural networks for fault identification and performance degradation evaluation of high-speed train bogie," *IEEE Transactions on Neural Networks and Learning Systems.*, 2018.

[49] K. Liang, N. Qin, D. Huang and Y. Fu, "Convolutional recurrent neural network for fault diagnosis of high- speed train bogie," *Complexity*, 2018.

Palak Verma was born in Faridabad, India in 1992. She received her master of science in electrical communication engineering from the University of Kassel, Kassel, Hessen, Germany in 2016; a post graduate diploma in advanced computing from the Centre for development of Advanced Computing (C-DAC Acts), Pune, Maharashtra, India, 2014; and a bachelor of technology from YMCA University of Science and Technology, Faridabad, Haryana, India, 2013.

Currently, she works as a technical advisor in engineering systems for Canada Post, Ottawa, Ontario, Canada. She is also a P.ENG (professional engineering) instructor for Global Innovative Campus, Canada. Furthermore, she had worked as an engineering intern in Siemens.

Matin G. Foomani was born on Jan. 22nd. He is currently pursuing a PhD in civil engineering in transportation systems from Concordia University, Montreal, Quebec, Canada. He received his MASc, engineering in intelligent transport systems from FH (UAS) Technikum Wien- Vienna, Austria, 2011. He was awarded a M.B.A in business administration, Warwick Business School, Coventry, UK, 2008.

Currently, he works as a manager in engineering systems for Canada Post in Ottawa, Ontario, Canada. He is also a research and teaching assistant at the Concordia University, Montreal, Quebec, Canada. Furthermore, he was a transportation network engineer at Canada Post. He also worked at Orange Traffic INC as a transportation engineer. He was the consultant at Ekium and worked as a project coordinator and ITS Engineer. He has also worked as a consultant at Gli (Gli.Fr.) as ITS engineer.