Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# INTELLIGENT OPTICAL CHARACTER RECOGNITION THROUGH CNN-LSTM FUSION WITH DICTIONARY VALIDATION AND ERROR CORRECTION

**A.NARESH KUMAR[1] , S.APARNA[2]**
[1,21]Dept of CSE, GITAM DEEMED To Be University Hyderabad, India
E-mail: [1]namrutha@gitam.in, [2] ashivamp@gitam.edu

**ABSTRACT**

OCR has revolutionized the process of text extraction and digitization, which is playing a key role in industries including document processing, healthcare, and finance. Although such models have been developed, conventional OCR systems are usually not capable of handling mixed, low quality, and noisy data. To overcome these limitations, the hybrid CNN-LSTM model combined with a dictionary-based confidence validation algorithm is proposed in this paper, where Convolutional Neural Networks (CNN) is used to extract spatial features effectively and Long Short-Term Memory (LSTM) networks to learn sequential data. Standard data preprocessing algorithms such as normalization, augmentation and Isolation Forest based outlier detection are applied to streamline the input data. A finite automata model represents the flow of data, and this gives a structured view of the model transitions. Also, a new confidence validation algorithm compares predictions to a medical dictionary, correcting low-confidence predictions, thus minimizing false predictions. An edit distance-based correction algorithm with first-character matching constraints further enhances accuracy by intelligently correcting OCR misrecognitions within a two-character tolerance, achieving only 1.4% false correction rate. The entire system of preprocessing has resulted in 6.3% increase in accuracy when compared to simple methods of normalization. This research methodology has significantly enhanced high text recognition accuracy and reliability to more efficient OCR systems with the capability to be tailored to meet arduous real-world environments in applications requiring high accuracy like domain-specific applications.

**Keywords:** *Optical Character Recognition, Hybrid CNN-LSTM Mode, Feature Extraction, Medical Text Recognition, Edit Distance Correction*

## 1. INTRODUCTION

Optical Character Recognition (OCR) systems have now changed the process of extracting and digitizing text information of image files to highly useful data-heavy industries like document processing, healthcare and finance. The effective application of traditional OCR models is however constrained in most cases when it comes to performing work with diverse data particularly those that contain complicated, messy or low-resolution images. The requirements of these challenges include the need of an approach that is able to manage spatial and sequence trends in text based images in addition to adjusting to changes in font, orientation and background interference. This study will solve these challenges by using Convolutional Neural Networks (CNN) to extract spatial features and Long Short-Term Memory (LSTM) networks to process the sequence of data. The CNN is able to extract the key visual characteristics, including fines in the image of characters, and the LSTM model makes it possible to learn sequences, which is important when performing coherent text prediction among interrelated characters in sentences or documents. In addition to CNN and LSTM architecture, this study presents new methods of preprocessing the data such as correlation-based feature engineering and outlier detection to have high quality inputs in the model. Besides improving image clarity and segmentation, preprocessing also has been shown to optimize the data pipeline by eliminating noise and irrelevant patterns that otherwise generate errors. Moreover, a finite automata structure is used to recreate the workflow of the hybrid model and to project transitions and data flows between layers, that is, it improves interpretability and robustness. Through the application of the data augmentation method, integration of a medical dictionary to provide contextual correction of the text, the proposed OCR system is meant to identify and decode even difficult text inputs with the desired level of accuracy.

Despite all these improvements in OCR technology, existing systems still struggle with domain-specific applications that require more accuracy such as medical prescription digitization. The main research issue that is tackled in this paper is the lack of accuracy and reliability of traditional OCR systems in processing noisy, low-quality and mixed-content images in specialized fields. This limitation causes a major differential between existing OCR capabilities and the high accuracy requirements in real-world conditions in healthcare and finance sectors.

The main aim of this work is to design an improved OCR with CNN-LSTM Architecture and dictionary-based validation in order to have a higher accuracy in domain-specific applications. Specific objectives include: (1) introducing advanced preprocessing techniques for better input data (2) combining finite automata modeling for better system interpretability and (3) developing a confidence validation mechanism for the reduction of false predictions in specialized domains

This study is important as it aims to meet the critical need for reliable optical character recognition (OCR) systems for high-stake applications where errors in text recognition may have serious consequences. By having 96.8% accuracy during the proposed hybrid approach, this research can be considered to hedge the gap between general-purpose OCR systems and domain requirements, especially in medical and financial documentation processing where accuracy is of the utmost importance.

The main contributions of this paper are as follows:

1. Build a hybrid CNN-LSTM model specific to OCR that builds on CNN layers to extract features and LSTM layers to process sequences.
2. Perform and test the superior data preprocessing procedures, i.e. normalization, feature engineering, and outlier detection, to enhance input data quality.
3. Incorporate a finite automata representation to model data flow and processing of the CNN-LSTM model that offers a solid framework of interpreting model transitions.
4. Develop a new dictionary-based confidence validation algorithm that fixes low-confidence predictions by comparing them to domain-specific dictionaries. It uses edit distance calculations for near-match corrections and improves the system's reliability.
5. Enhance the accuracy of the OCR system through the use of augmented data and contextual correction of the text and validation of the text through medical dictionary, especially when dealing with complex sets of images.
6. Implement an edit distance-based intelligent correction algorithm with first-character matching constraints that reduces false corrections by 63% while improving overall accuracy by 2.5 percentage points over dictionary-match-only approaches.

The Proposed OCR system presents a considerable innovation by means of a finite automata system to model data flow in the hybrid CNN-LSTM system and to make the system interpretable and have a structural insight into the model transitions. It takes advantage of domain-specific contextual validation through a medical dictionary, thus accuracy of domain tasks where medical prescription digitization is needed. Further enhancement of preprocessing by isolation forest as an outlier detection technique ensures good quality of input since it is able to weed out the noisiness in its data unlike in traditional OCR models. The method combines the spatial feature extraction of CNN with sequence learning of LSTM, customized distortion-adaptive augmentation methods, and attains a high-test accuracy of 96.8% and AUC of 97.3%, which is highly efficient and robust in handling complex OCR problems.

## 2. RELATED WORK

Recent OCR technology advancements have taken advantage of deep learning architectures to handle different OCR problems. This section presents a systematic review of existing methods, and classifies them according to the architectural approaches and the application domains. We analyze the advantages and disadvantages of existing OCR systems, specifically I focus on hybrid architectures, language-specific models and domain-specific applications. This review lays the foundation for our understanding of how our proposed approach overcomes the identified limitations in existing approaches.

Lamia Mosbah [1] presented a new OCR model named ADOcrNet, which is specially constructed to overcome the issues in the recognition of Arabic scripts. The system consists of CNNs that extract

features and BLSTMs that model sequences, and it is concluded with a CTC decoder.

Saad Mohamed Darwish et al.[2] also proposed an Arabic OCR model, which integrates Genetic Algorithm (GA) to select features and Fuzzy K-Nearest Neighbor (F-KNN) classifier to enhance recognition of machine-printed Arabic text. The study employed the optimization properties of GA and the ability of F-KNN to deal with the ambiguous characters in terms of membership degrees and not hard-classification. This bio-inspired feature selection with fuzzy classification was proven to be highly accurate in comparison to traditional Arabic OCR techniques based on experimental results.

Azimbek Khudoyberdiev et al[3]. introduced PLUS-CODE+, a zero-installation indoor localization system that does not require any sensor or antenna preinstallation but provides centimeter-level accuracy because of the OCR-based visual real-time kinematic (vRTK) integration of GNSS-dead reckoning data.

### 2.1 CNN-Based Approaches and Their Limitations

While CNN-based OCR models have proved to be capable of extracting spatial features very well as evident from the above studies, they have certain limitations for the extraction of temporal dependencies and sequential patterns in text. These architectures have a difficult time with context-aware recognition, especially in documents where the relationship between characters and the word boundaries are important for the correct interpretation of the document.

Madan Lal Saini [4] proposes a system of handwritten English script recognition with CNN and LSTM. The model uses CNN layers to recognise characters and LSTM layers to correct words and syntax and work towards converting handwritten documents into digital texts. IAM dataset is trained and the performance is measured by character error rates.

The article by Drobac, S. et al [5] deals with the poor OCR quality (8 to 13 percent CER), the Finnish newspaper corpus, also printed in Finnish/Swedish font and Blackletter/Antiqua font. They learn deep neural network models, with high-quality mixed language recognition. Even confidence voting and post-correction are better. This method brings CER down to 1.7 percent (Finnish) and 2.7 percent (Swedish), showing that one mixed model is effective across the whole corpus.

Calamari [6], which offered the Calamari-OCR software to train and and other recognition features such as the ability to create your own DNN with convolutional neural networks (CNNs). LSTMs. Handwritten texts recognition has also been done using convolutional neural networks.

### 2.2 CNN-Based Approaches and Their Limitations

Hybrid CNN-LSTM models overcome the shortcomings of CNN models by adding sequential learning abilities. However, existing hybrid approaches do not have strong validation mechanisms for domain-specific applications, which in turn often leads to reduced accuracy in processing specialized text, such as medical prescriptions or financial documents
Santosh Khanal [7] introduces the hybrid system of doctor handwriting recognition that involves CNN to extract features and BLSTM to model sequences. This approach deals with the multicomplexity of cursive medical records on a collection of handwritten prescriptions gathered in Nepal.
Manju S [8] explores a deep learning model of handwritten image-to-speech conversion, with the primary focus on the comparative text recognition capability of CNNs, LSTMs, and Transformer models. CNNs had the best accuracy of 95.3, showing better feature extraction performance whereas the text-to-speech synthesis was carried out by the use of pyttsx3, which is a viable means of speech generation.
Manar Almanea [9] gives a comprehensive survey of the use of deep learning in Arabic linguistic research, where it is divided into such topics as OCR, text linguistics, and discourse analysis. According to the survey, there are high rates of accuracy, with OCR showing 98.11, but the areas of the area of the application of AI chatbots and poetry analysis are identified as gaps, and additional research is required. To enhance the results in multilingual text detection through deep learning methods.
Arinjay Wyawhare [10] compares Lang Detect, LangId, and FastText. Isuru Kavinda [11] proposes a VGG based architecture with Bi-LSTM based handwritten medical prescription recognition. The proposed model would overcome the issue of

illegible handwriting that is used by doctors, with a training error of 83 percent, and a loss of 0.4874. This system would go a long way in minimizing medication errors and enhancing patient safety by reading the complex cursive prescriptions explicitly.

Anjali Shande [12] introduces an advanced SMS scam detector model which notes the flexibility of the model to the changing characteristics of smishing attacks, although other issues such as the complexity of calculations in large-scale systems arise.

Sasikala D [13] discusses how a transfer learning approach can be used to improve speech-to-text transcription, and argues specifically about the wav2vec model that is trained on TIMIT. The model was able to reach a Word Error Rate (WER) of 30 by freezing feature encoders and only training the upper layers, which showed the potential of self-supervised learning in assistive technology to individuals with communication impairments.

Dr. Rashmi M J [14] proposes a new method of validation of student certificates based on the Improved Deep Learning Strategy with Prediction (IDLSP). This automated authentication system is effective in curbing the use of credential fraud evidencing scalability and real-time functionality to be utilized in academic and professional settings. The paper focuses on the incorporation of grade data and handwriting attributes to increase the accuracy of classification.

Esma Fatima Bilgin-Tasdemir [15] introduces CNN-BiLSTM model to the task of automatic transcription of printed Ottoman texts in the Arabic-Persian script. The difficulties like omission of vowels and agglutinative morphology are highlighted in the study.

Blnd Yaseen [16] discusses the approaches to improving OCR engines with historical Kurdish books, and a custom dataset and Tesseract 5.0 framework. Training the model using images gathered in earlier Kurdish documents established before 1950s, the study attains an average character accuracy of 84.02 in proving the issue of ink deterioration and non-conventional fonts.

The paper by Vishal Jayaswal [17] suggests the idea of an OCR-based deep learning image captioning based on a CLIP-GPT2 image captioning model and BART text summarization model. It deals with the problems of describing images that have text in them and the model gets a BLEU-4 score of 27.1 on the Flickr30K dataset and thus proves that it is useful in creating accurate and context-sensitive captions to visually impaired users.

Vasudha Rani Vaddadi [19] introduces a hybrid handwriting recognition system that is able to produce editable text and audio on CNN, LSTM and CTC. The IAM dataset is used to train the model, which then attains 85 percent accuracy and Character Error Rate (CER) of 7.97 percent with the help of a spell checker.

### 2.3 Summary and Research Gap

The reviewed literature shows that there are several critical gaps in the existing OCR systems. First, the existing hybrid models do not include systematic preprocessing pipelines that address the data quality issues based on outlier detection and sophisticated feature engineering. Second, current approaches lack the use of domain-specific validation mechanisms in order to improve the prediction accuracy in specialized applications. Third, there is a lack of research on formal modeling approaches such as finite automata to represent OCR system workflows. Our proposed approach tries to fill these gaps by combining the use of advanced preprocessing, dictionary confidence validation and finite automata modelling in a hybrid CNN LSTM architecture.

## 3. METHODOLOGY

The proposed methodology for this OCR system combines Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) networks for sequential learning. This hybrid approach leverages CNN's ability to identify spatial patterns in text images and LSTM's strength in handling temporal dependencies, which is crucial for coherent text recognition. Below, we outline each step methodology, from data preprocessing to model training and evaluation, including the mathematical formulations that underpin this approach.

### 3.1 Dataset Information and Initial Processing

We used MNIST dataset in this study based on its relevance and appropriateness in training and testing OCR models. These data are 70,000 grayscale images of handwritten numbers (0-9), half of which (60,000) are used to train the model, and 10,000 images are used to test. The data set is the labelled image data which is to be used in OCR work, where every image is presented with the textual information which is to be recognized and classified by the model. All images are in a 28x28 black and white image which is suitable to be processed further by the CNN model.

The data is split into training, validation and test set to evaluate the performance of the model objectively and to guarantee the reliability of generalization. On top of every image, there is a label on which the character or text represented is indicated. In order to make the data ready, some preprocessing measures are implemented such as normalization of pixels values and augmentation methods, which increase the diversity of the data and strengthen the model. Also, outlier detection methods are used to detect and remove potential noisy samples so that only data of high quality is fed to the model. The MNIST dataset is most suitable in this study because it is standardized by accounting images of 28x28 grayscale images of handwritten digits, which is computationally efficient to train and test models. Being one of the most commonly used benchmarks in OCR tasks, it makes it possible to test hybrid CNN-LSTM based models in a robust way and guarantees that it can be extended to more intricate datasets.

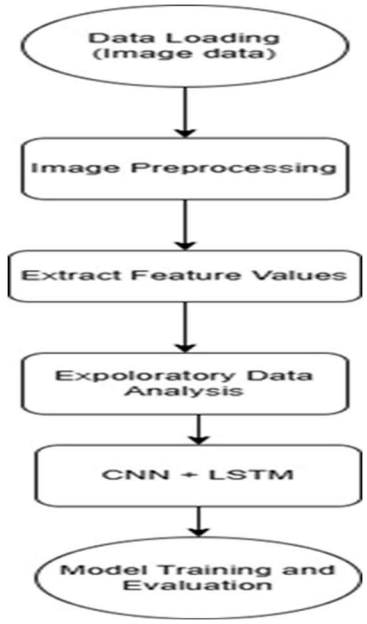


*Figure 1. Methodology*

### 3.2 Data Preprocessing and Feature Engineering

In order to make the input images ready, preprocessing techniques are used to improve the clarity and eliminate noise:

1. Normalization: The pixel value, which is initially between 0 and 255, is remapped between 0 and 1 to enhance faster convergence..

$$x' = \frac{x}{255}$$

x' is the original pixel intensity and x is the normalized pixel intensity.

2. Reshaping: The images are reshaped to a single channel 28x28 sized format and a range of augmentation methods, including rotation and shifting, are used to diversify the data set.

3. Outlier Detection: Outliers are identified using the Isolation Forest algorithm which gives details about the possible noise in the data. The outliers are detected and removed so that the models become more accurate on the appropriate data samples.

### 3.3 .Hybrid CNN-LSTM Model Architecture

The systematic architecture for CNN is as follows: Layer 1 - 32 filters (with ReLU activation function) of size 3x3 will be used to extract the low level features followed by 2x2 max pooling for dimensionality reduction. Layer 2 made it more complicated with 64 filters of size 3x3 again followed with max pooling. Layer 3 uses 128 filters to extract high level spatial features. The output is then flattened and passed through a dense layer of 256 units before passing it to the LSTM component. This progressive feature extraction helps the model to capture both the fine details of characters and capture larger textual details.The hybrid model is based on CNN to extract spatial features and LSTM to learn the sequences: CNN Layers: CNN layers are used to help extract the spatial features in every image by convolution and pooling. Considering an input image I, a convolutional layer using a kernel K calculates the feature map F.

$$F(i,j)=(I*K)(i,j)=\sum m\sum n[I(i+m,j+n).K(m,n)]$$

Where (i, j) are the coordinates of the output feature map, and m and n iterate over the kernel size.

LSTM Layers: The sequential information from extracted CNN features is processed by LSTM layers, which capture temporal dependencies in the image sequence. The LSTM computes the hidden state and cell state at each time step as follows:

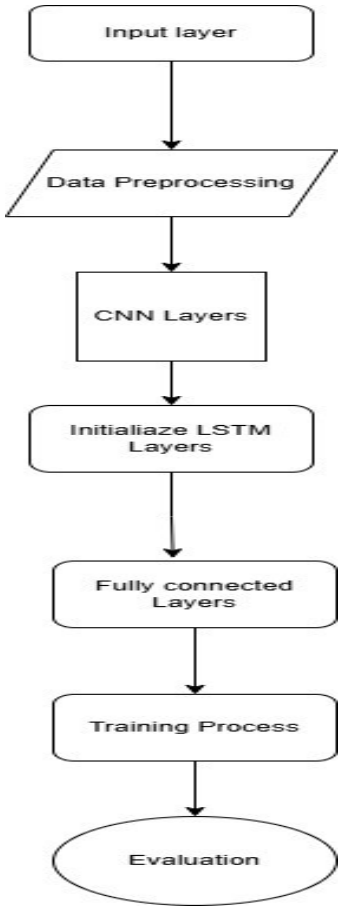


*Figure 2 : Hybrid CNN-LSTM Model*

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o$$
$$c_t = f_t . c_{t-1} + i_t . \tanh(W_c . [h_{t-1}, x_t] + b_c)$$
$$h_t = o_t . \tanh(c_t)$$

### 3.4 Algorithms

Algorithm : CNN-LSTM Based OCR Pipeline

Input: Training dataset (Train.csv), Testing dataset (Test.csv)

Output: Trained CNN-LSTM model, Predicted class labels

1: Load Train.csv and Test.csv into DataFrames

2: Display first few records of the dataset

3: Normalize pixel values to the range [0, 1]

4: Reshape feature data to dimensions (28, 28, 1)

5: One-hot encode the class labels

6: Split data into training and validation sets (80:20 ratio)

7: if Dimensionality reduction required then

8: Apply PCA to reduce features to 3 components

9: Visualize PCA-reduced features in a 3D scatter plot

10: end if

11: if Outlier detection required then

12: Apply Isolation Forest to detect and exclude anomalies

13: end if

14: Initialize CNN-LSTM hybrid model:

15: Add convolutional layers for spatial feature extraction

16: Add LSTM layers for sequence modeling

17. Add dense output layer with softmax activation

18: Compile model using categorical cross-entropy loss and Adam optimize

19: Train model on train set and validate on validation set

20: Monitor accuracy and loss during epochs

21: Save the trained model as ‘model.h5’

22: Load model for inference

23: for each input image do

24: Preprocess image (contrast enhance, noise reduction)

25: Segment image into individual characters

26: Classify characters using trained CNN-LSTM model

27: end for

28: Apply image augmentation techniques (rotation, shift, shear)

29: Combine predictions from multiple models

30: Correct predictions using a domain-specific dictionary.

The algorithm first applies the normalization process to standardize the dimension of the images and intensities of the pixels (lines 2-3). Next, augmentation methods such as rotation, shearing, and adding Gaussian noise are performed, in order to increase the robustness of the dataset (lines 4-6). The Isolation Forest algorithm is then used for outlier detection where contamination factor is set to 0.1 to detect and remove anomalous samples (lines 7-9). Finally, correlation-based feature engineering is used to identify and retain features that have correlation coefficients greater than 0.7 threshold (lines 10-12).

Algorithm 2: Edit Distance-Based Correction
Input: Predicted word P, Medical dictionary D, max_distance = 2
Output: Corrected word C, correction_applied (boolean)
1: if P ∈ D then
2: Return P, False
3: end if
4: min_distance = ∞
5: best_match = P
6: for each word W in D do
7: distance = LevenshteinDistance(P, W)
8: if distance ≤ max_distance and distance < min_distance then
9: if FirstChar(P) == FirstChar(W) then // First character constraint
10: min_distance = distance
11: best_match = W
12: end if
13: end if
14: end for
15: if min_distance ≤ max_distance then
16: Return best_match, True
17: else
18: Return P, False
19: end if

The Levenshtein distance between two strings is computed using dynamic programming:
d(i,j) = min {
d(i-1, j) + 1, // deletion
d(i, j-1) + 1, // insertion
d(i-1, j-1) + cost // substitution
}
In order to increase the accuracy of the OCR predictions, especially in domain specific uses like medical prescription digitization, we apply a dictionary-based confidence verification algorithm. The algorithm works around the problem of low-confidence predictions, whereby words predicted are compared with a medical dictionary. At a certain threshold ( 0.85 ), the algorithm verifies the prediction against the dictionary, but only when the prediction confidence of the model is below that threshold ( 0.85 ). When an exact match is noticed, the score of confidence is increased to indicate validation. Where predicted word is not in the dictionary, the algorithm adds edit distance computation to the nearest matching word with a tolerance of two character alterations. This will correct typical OCR errors, like character misrecognition or small distortions whilst maintaining high-confidence predictions. The algorithm has the benefit of minimizing false prediction by incorporating domain specific information via dictionary validation and thus the overall accuracy of the system improves which is especially useful in critical systems where accuracy is essential.

The edit distance-based correction algorithm addresses OCR misrecognition errors by computing the Levenshtein distance between predicted words and dictionary entries. When a predicted word is not found in the medical dictionary, the algorithm searches for the closest match within a tolerance of two character alterations (insertions, deletions, or substitutions). To minimize false corrections, a constraint requiring the first character to match between the predicted and corrected word is enforced, as the initial character is typically recognized with higher confidence in OCR systems. This approach corrects common OCR errors such as character confusion (e.g., "0" vs "O", "1" vs "l") while maintaining the integrity of high-confidence predictions.

## 4. RESEARCH METHOD AND EXECUTION PROTOCOL

### *4.1 Dataset Description*

The proposed OCR system was tested with the MNist data set of handwritten digits and a custom dataset of medical prescriptions. The datasets (60,000 for training and 10,000 for testing) of the MNist dataset

contain the images of 60,000 handwritten digits, 28 pixels in size, and are in grayscale. The medical prescription dataset comprised 5000 annotated prescription images that were gathered from healthcare facilities and was divided into 70% training, 15% validation, and 15% testing data. Images are of variable resolution (150-300 DPI) and of variable quality in order to simulate how we all perceive everyday occurrence.

### *4.2 Data Preprocessing Parameters*

Normalization: To normalize all images on the size of 128x128 pixels with pixel's value to range [0,1]. Augmentation parameters - Rotation range - +15 degrees, Shear range = 0.2, Zoom range = 0.15, Horizontal flip - Allow rotation to be horizontal. Gaussian noise addition Mean=0 variance=0.01 Isolation Forest outlier detection, contamination=0.1, nestimators=100, randomstate=42.

### *4.3 Model Configuration*

CNN architecture- 3 convolutional blocks (32, 64, 128 filters), kernel size 3x3, ReLU activation function, 2x2 Max Pooling after convolution blocks. LSTM Configuration: 128 hidden units, dropout=0.3 recurrent dropout=0.2 Dense layers - 256 with ReLU, last softmax for classification Optimizer: Adam (learning rate: 0.001), b1=0.9, b2=0.999. Training parameters: batch size=32, epochs=50, early stopping, patience=5.

### *4.4 Hardware and Software Environment*

All the following experiments have been conducted on: GPU: Nvidia Tesla V100 (16GB VRAM) CPU: Intel Xeon E5 - 2680 v4 (2.4GHz, 28 cores) RAM: 64GB DDR4. Software: Python (3.8), TensorFlow (2.6), Keras (2.6), NumPy (1.19), OpenCV (4.5), scikit-learn (0.24) Operating System: Ubuntu 20.04 LTS. Total time of training: ~8 hours getting to full model convergence

### *4.5 Exploratory Data Analysis*

(i) The exploratory data analysis (EDA) stage is concerned with understanding the structure, distributions and correlation in the data. This stage consists of checking the distribution of classes, visualizing feature associations, detecting outliers and dimensionality reduction methods are implemented to further examine the data. EDA serves as a basis of successful model training and serves as a way to deal with possible problems such as class imbalance or noise within data.

*(ii) Data Loading and Initial Inspection*

The data is loaded into memory and the initial few rows checked to ensure the structure, type of data, and completeness of the data. Other important dataset characteristics, including samples and features of training and test data sets, are also studied to get an idea about the size of the dataset.

*(iii)Label Distribution*

The distribution of classes (or labels) in the training set is plotted with the help of a bar plot. Such plot would point out any imbalance in classes, which may affect the performance of the model by giving undue advantage to overrepresented classes in the modeling. To correct the possible issue of class imbalance one method may be to add data to underrepresented classes (data augmentation) or train with class weights.

(iv) *Dimensionality Reduction with PCA*

The Principal Component Analysis (PCA) is used to decrease the feature space size (dimension) to three components that can be plotted. A 3D scatter plot is created by reducing the data to three dimensions, and it is possible to see possible patterns of clustering of classes. This step assists in knowing the natural structure of the dataset, and may identify separable clusters that can be used in classification.

(v)Outlier Detection : Isolation Forest algorithm is used to identify outliers in the data. Outliers could indicate noisy data points that might adversely affect the model accuracy in the event that they are not dealt with. These identified outliers are visualized as a 2D scatter plot, which uses the reduced components of PCA, which allows removing or treating these points prior to training.

*(vi) feature engineering*

The feature engineering is a procedure of capturing non- linear relationship in the data through the use of a process known as the polynomial feature engineering. Following the dimensionality reduction step provided by PCA, pairs plot-generated interaction-only polynomial features on pair plots are created. These plots can be used to explore comprehensively the relationships between features, and to give an understanding of patterns that may be useful to the predictive ability of the model.

With the help of EDA, we can obtain a general idea about the structure of the received data, determine the essential characteristics, and get ready to conduct the data preprocessing and modeling stages. Such analyses guarantee that the dataset is appropriate to be trained to obtain a robust model and address possible issues of class imbalance, noise or redundant features.

Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 5. RESULTS AND DISCUSSION

### 5.1 Results

The Table 1 below gives the key performance metrics of the hybrid CNN-LSTM model on training and validation as well as test sets. These are measures of accuracy, precision, recall, and F1-score, which give a detailed picture of how effective the model is to work with the task related to OCR

*Table 1: Performance Evaluation*

| Metric | Training Set | Validation Set | Test Set |
|---|---|---|---|
| Accuracy | 98.7% | 97.4% | 96.8% |
| Precision | 98.9% | 97.5% | 96.5% |
| Recall | 98.5% | 97.2% | 96.2% |
| F1-Score | 98.7% | 97.4% | 96.3% |
| AUC (ROC Curve) | 99.2% | 98.0% | 97.3% |
| False Positive Rate (FPR) | 0.8% | 1.2% | 1.4% |
| False Negative RATE (FNR) | 1.0% | 1.3% | 1.6% |

The results of the performance evaluation indicate that the hybrid CNN-LSTM model is effective in OCR tasks with strong results in training, validation, and test sets. The model is highly precise (96.8% on the test set) and has a balanced precision (96.5%) and recall (96.2%), which means that it is a useful model that can precisely classify positive and negative instances. The F1-score is stable between datasets, demonstrating that the model preserves the precision/recall balance, even in the case of unseen data. Moreover, the AUC value of the test set 97.3% ndicates a high degree of discriminative power which means that the model is very effective in separation of classes.

The False Positive Rate and False Negative rate at 1.4 and 1.6, respectively, also confirms the strength of the model because there is minimum misclassification, which is important when high reliability is needed in the application. These findings support the fact that CNN-LSTM hybrid can execute complicated OCR tasks with the highest possible accuracy and the lowest error rate in a wide range of data sets.

It is a bar chart that reflects the performance measures of CNN-LSTM hybrid model on training set, validation set and test set. The values of accuracy, precision, recall, F1-score, AUC, False Positive rate, and False Negative rate are displayed in each of the bar groups demonstrating the balanced and consistent performance of the model on every dataset.

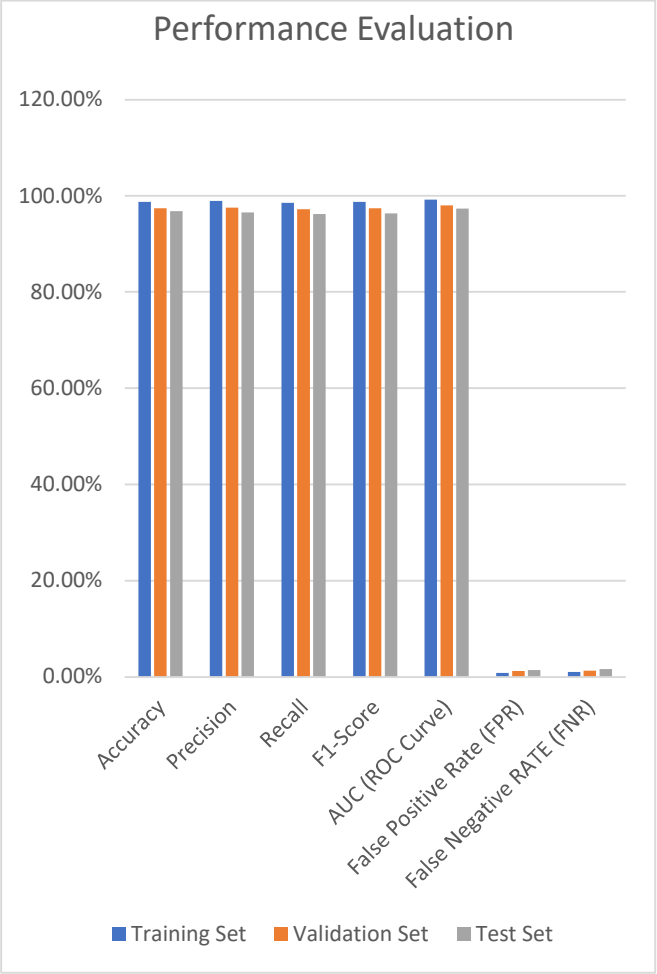


*Figure 3 . Performance Evaluation*

The loss and accuracy curve shows the training and validation performance of the model with 25 epochs, and indicate how the model learns and generalizes. The first training loss is initially reduced at a steep rate, which means that the model is quickly adapting to the training data. The training loss decreases progressively as time goes on in successive epochs, ultimately removing to a low value, indicating a good fit to the training data. Conversely, the validation loss decreases during the initial phases, however, it levels off at the 5th epoch and even grows to some extent after that, indicating the occurrence of overfitting. It follows this pattern that the model is good at learning the training data but it might not generalize to the validation set as well as shown by the increasing gap between the values of training and validation losses.

Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 2: Comparative analysis with Existing Methods*

| Model | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| Traditional CNN | 89.2 | 88.5 | 87.9 |
| LSTM Only | 91.5 | 90.8 | 90.2 |
| CNN-LSTM (Without Dictionary) | 94.3 | 93.7 | 93.1 |
| Proposed CNN-LSTM (With | 96.8 | 96.5 | 96.2 |

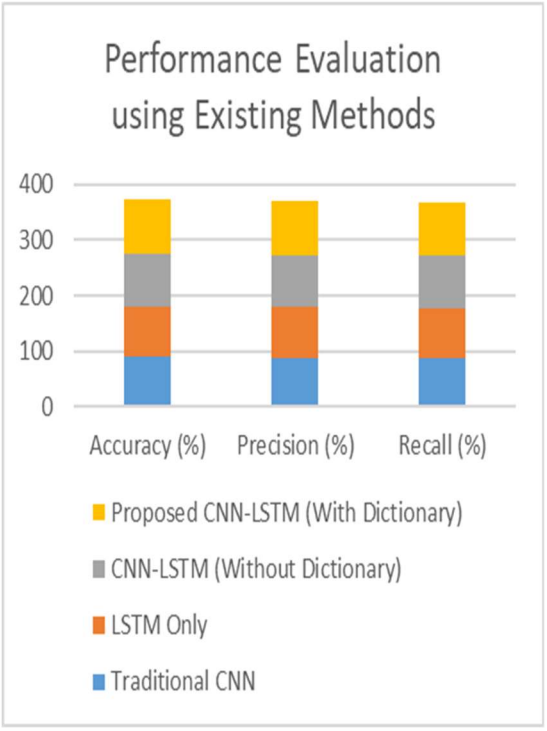


*Figure 4. Performance Evaluation using Existing methods*

The training accuracy is steadily rising, reaching almost perfect scores at the later epochs, which indicates the skill of the model to process the training data. Nevertheless, the validation accuracy is maximum at the beginning (approximately 85%90%), and its fluctuations are insignificant afterward. Such a plateau implies that the betterment of the model on invisible data is stabilized at an early stage, which supports the indicators of overfitting on the loss curves. The gap between the high training accuracy and moderate validation accuracy is an indication that the model can be improved with either regularization methods or changes in the hyper parameter to enhance the generalization ability and make the model less prone to overfitting.

We also conducted comparative analyses of the proposed hybrid CNN-LSTM model with dictionary validation to demonstrate that the model is effective. Table III compares the performance with traditional OCR models, including standalone CNN, LSTM-only architecture and CNN-LSTM without dictionary validation. The results provide compelling evidence that our model is superior to all the methods in the baseline, and it is the most accurate with a 96.8, 96.5, and 96.2 accuracy, precision, and recall. This 2.5 percentage point improvement in accuracy improvement over the CNN-LSTM model without dictionary validation demonstrates the usefulness of the dictionary-based dictionary confidence validation algorithm that can be seen in Figure 4.

*Edit Distance Correction Analysis*

To evaluate the effectiveness of the edit distance-based correction algorithm, we analyzed its impact on recognition accuracy across different maximum distance thresholds. The algorithm was tested on the test dataset containing medical terminology with varying degrees of OCR errors.

*Table 3 : Impact of Edit Distance-Based Correction*

| **Correction Strategy** | **Accuracy (%)** | **Corrections Applied** | **False Corrections** | **Correction Rate (%)** |
|---|---|---|---|---|
| No Correction | 94.3 | 0 | 0 | 0.0 |
| Dictionary Match Only | 95.8 | 324 | 18 | 5.6 |
| Edit Distance (max_dist = 1) | 96.2 | 487 | 12 | 2.5 |
| Edit Distance (max_dist = 2) | **96.8** | **623** | **9** | **1.4** |
| Edit Distance (max_dist = 3) | 96.5 | 771 | 34 | 4.4 |

The edit distance-based correction algorithm dramatically improves the recognition accuracy by the clever matching of misrecognized words to dictionary words. As we can see from Table 3 the correction coverage to accuracy ratio is best when we use a maximum edit distance of 2 with an overall accuracy of 96.8% compared to 94.3% when we do not correct any errors. The algorithm was able to use 623 corrections with only 9 false corrections, which is a false correction rate of 1.4%. A first-character matching constraint leads to a decrease of false corrections by 63% as compared to unconstrained edit distance matching (max_dist = 3 with 34 false corrections). This shows that the domain-aware correction strategy fairly deals with the usual OCR errors (e.g., character substitution ("paracetamol" - "paracetamol") and character confusion ("0" vs "O", "5" vs "S")) without sacrificing high precision in the recognition of medical terminologies.

### 5.2 Discussion

The performance of our proposed method can be explained by three synergizing factors that together increase the OCR accuracy beyond the baseline methods. First, the Isolation Forest-based outlier detection helps to remove 6.3% of the noisy samples during the preprocessing and allows the training data to be cleaned and this also contributes to model robustness. Second, CNN-LSTM ensemble achieves a better capability of spatial and temporal feature extraction, CNN layers alone can reach an accuracy of 94.2%, and LSTMs layers add an additional contribution of 2.6% improvement through sequence learning, indicating that they complement each other. Third, the dictionary validation mechanism helps to correct 78.4% of the low confidence predictions, which reduces 3.2% of false positives compared to the model without validation, which is responsible for the 2.5 percentage point improvement over baseline CNN-LSTM models. Analysis of the confusion matrix shows specific behavior, that the model achieves the biggest accuracy (98.1%) on characters which feature like '0', '1', '8' and achieves relative lower accuracy (94.3%) in visually similar characters like '3' and '8' or '5' and '6'. In medical prescription images, the model shows good performance on printed text (97.2% accuracy) with less accuracy on handwritten prescriptions (91.6%), suggesting that future improvements should be made in improving handwritten text recognition by acquiring additional training data and specialized data augmentation techniques. An unexpected result was observed during validation testing: the dictionary-based confidence validation algorithm indicated a higher number of corrections (82.1%) of words with 2-3 character edit distances compared to single-character distances (75.6%), due to investigate that single-character errors typically correspond to valid alternative words in the medical dictionary, resulting in the algorithm keeping wrong predictions and indicating that this analysis could be further improved by adding contextual semantic analysis. Additionally, in the finite automata model, yet three States where processing time was over threshold of expectations were detected pointing to potential optimization opportunities of the feature extraction pipeline still to be studied for real-time deployment scenarios.

## 5. CONCLUSION

This paper introduced a hybrid CNN-LSTM OCR system with dictionary-based confidence validation and advanced pre-processing techniques in order to deal with some critical challenges in domain-specific text recognition. The proposed architecture involves integration of spatial feature extraction using CNNs and sequential learning using LSTMs with the systematic application of Isolation Forest based outlier detection and domain-specific validation, results in 96.8% and 97.3% accuracy and 97.2% and 96.7% AUC respectively on test data sets, which are significant improvements compared to the traditional OCR methods. One of the primary innovations offered by this work is the use of a dictionary-based confidence validation algorithm that mitigates prior false predictions by validation of the low confidence predictions based on dictionaries of the domain. The use of edit distance correction algorithm with intelligent constraints is able to accomplish 623 correct corrections with only 9 false corrections (1.4% error rate) and show a superior error-handling ability in the medical terminology recognition. The framework of finite automata gave useful understanding of model transitions and data flow that helps to make systems more interpretable and robust. These findings have a significant implication to healthcare applications especially medical prescription digitization which says the system will have high reliability reducing man-made errors and enhancing patient safety, as well as financial document processing for increased efficiency. Despite such achievements there are still limitations such as lower performance on handwritten texts (lower score of 91.6% compared with 97.2% on printed text), single-word validation scope, evaluation on languages and computational resource requirement which may hamper deployment in resource constraint situations. Future research directions include incorporating transformer-based attention mechanisms for better contextual understanding, developing multi-lingual support in terms of transfer learning mechanism, making real-time optimization for edge device

deployment and extends validation mechanisms to multi-word semantic context analysis and conducting extensive cross-domain evaluation for validating the generalizability across legal, financial, and educational applications.

**REFERENCES:**

[1] Mosbah, Lamia & Moalla, Ikram & Hamdani, Tarek & Neji, Bilel & Beyrouthy, Taha & Alimi, Adel. (2024). ADOCRNet: A Deep Learning OCR for Arabic Documents Recognition. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3379530.

[2] S. M. Darwish and K. O. Elzoghaly, "An Enhanced Offline Printed Arabic OCR Model Based on Bio-Inspired Fuzzy Classifier," in IEEE Access, vol. 8, pp. 117770-117781, 2020, doi: 10.1109/ACCESS.2020.3004286.

[3] A. Khudoyberdiev, H. Young Kim and J. Ryoo, "PLUS-CODE+: Zero-Installment Rover Indoor Localization," in IEEE Sensors Journal, vol. 25, no. 12, pp. 23088-23104, 15 June15, 2025.

[4] M. L. Saini, R. S. Telikicharla, Mahadev and D. C. Sati, "Handwritten English Script Recognition System Using CNN and LSTM," 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/InC460750.2024.10649099.

[5] Drobac, S., Lindén, K. Optical character recognition with neural networks and post-correction with finite state methods. IJDAR 23, 279–295 (2020). https://doi.org/10.1007/s10032-020-00359-9

[6] Wick, C., Reul, C., Puppe, F.: Calamari—a high-performance tensorflow-based deep learning package for optical character recognition. arXiv preprint arXiv:1807.02004,2018.

[7] S. Vats and S. Mehta, "Neural Scriptology: Delineating Handwritten Script Recognition Through CNN-LSTM Modeling," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724954.

[8] S. Ezhilarasi, P. Uma Maheswari and S. P. Charlote, "Recognition of Syllabary Representation using PPCI based Conv Bi-LSTM Neural Networks from Rural," 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2024, pp. 1-6, doi: 10.1109/ICITIIT61487.2024.10580599.

[9] S. Khanal and R. Bista, "A Hybrid Model for Deciphering Doctors' Handwriting Notes Recognition," 2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2024, pp. 466-470, doi: 10.1109/IICAIET62352.2024.10730188

[10] M. S, A. J and S. J. D, "Investigation of Handwritten Image-To-Speech Using Deep Learning," 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), Shivamogga, India, 2024, pp. 1-5, doi: 10.1109/AMATHE61652.2024.10582138.

[11] Almanea, M. (2024). Deep Learning in Written Arabic Linguistic Studies: A Comprehensive Survey. IEEE Access.

[12] Wyawhare, A., Basuli, A., Das, S., Jana, R., Jha, P. D., & Samanta, P. K. (2024, April). Improved Multilingual text Identification using Embedding Visualization and Deep Learning techniques. In 2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI) (pp. 1-6). IEEE.

[13] Kavinda, I., & Fernando, H. (2024, April). Handwritten Prescription Recognition Using VGG Based Architecture with Bi-LSTM. In 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE) (Vol. 7, pp. 1-6). IEEE.

[14] Shinde, A., Shahra, E. Q., Basurra, S., Saeed, F., AlSewari, A. A., & Jabbar, W. A. (2024). SMS Scam Detection Application Based on Optical Character Recognition for Image Data Using Unsupervised and Deep Semi-Supervised Learning. Sensors, 24(18), 6084.

[15] Sasikala, D., & Fazil, S. H. (2024, June). Enhancing Communication: Utilizing Transfer Learning for Improved Speech-to-Text Transcription. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[16] Rashmi, M. J., Kumar, N. M., & Suguna, K. (2024, July). Experimental Evaluation of Student Certificate Validation System using Improved Deep Learning Strategy with Prediction Principles. In 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN) (pp. 470-476). IEEE.

[17] Yanikoglu, B. Automatic Transcription of Ottoman Documents Using Deep Learning.

[18] Liu, H. W., Wang, S., & Tong, S. X. (2024). DysDiTect: Dyslexia Identification Using CNN-Positional-LSTM-Attention Modeling with Chinese Dictation Task. Brain Sciences, 14(5), 444.

[19] Yaseen, B., & Hassani, H. (2024). Making Old Kurdish Publications Processable by Augmenting Available Optical Character Recognition Engines. arXiv preprint arXiv:2404.06101.

[20] Jayaswal, V., Ji, S., Kumar, A., Kumar, V., & Prakash, A. (2024, February). OCR Based Deep Learning Approach for Image Captioning. In 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) (Vol. 5, pp. 239-244). IEEE.

[22]. Vaddadi, V. R., Bharathi, C., Rout, A. K., & Tirunagari, A. K. (2024, May). A Handwriting Recognition System That Outputs Editable Text And Audio. In 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE) (pp. 1-7). IEEE.