Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 1, 2026, 13-24.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Comparative Evaluation of Deep Learning Architectures for Printed and Handwritten Farsi OCR

Fatemeh Asadi-Zeydabadi[1], Ali Afkari-Fahandari[1], Elham Shabaninia[2*], and Hossein Nezamabadi-pour[1]

1. Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.
2. Department of Applied Mathematics, Graduate University of Advanced Technology, Kerman, Iran.

Article Info

Article History:
Received 04 May 2025
Revised 28 July 2025
Accepted 24 October 2025

DOI:10.22044/jadm.2025.16098.2728

Keywords:
Farsi OCR, Persian Text Recognition, Deep Learning, CRNN, Transformer, Printed Text, Handwritten Text, Benchmark Fatasets

**Corresponding author: e.shabaninia@kgut.ac.ir (E. Shabaninia).*

Abstract

Farsi optical character recognition remains challenging due to the script's cursive structure, positional glyph variations, and frequent diacritics. This study conducts a comparative evaluation of five foundational deep learning architectures widely used in OCR—two lightweight CRNN-based models aimed at efficient deployment and three Transformer-based models designed for advanced contextual modeling—to examine their suitability for the distinct characteristics of Farsi script. Performance was benchmarked on four publicly available datasets: Shotor and IDPL-PFOD2 for printed text, and Iranshahr and Sadri for handwritten text, using word-level accuracy, parameter count, and computational cost as evaluation criteria. CRNN-based models achieved high accuracy on word-level datasets—99.42% (Shotor), 97.08% (Iranshahr), 98.86% (Sadri)—while maintaining smaller model sizes and lower computational demands. However, their accuracy dropped to 78.49% on the larger and more diverse line-level IDPL-PFOD2 dataset. Transformer-based models substantially narrowed this performance gap, exhibiting greater robustness to variations in font, style, and layout, with the best model reaching 92.81% on IDPL-PFOD2. To the best of our knowledge, this work is among the first comprehensive comparative studies of lightweight CRNN- and Transformer-based architectures for Farsi OCR, encompassing both printed and handwritten scripts, and establishes a solid performance baseline for future research and deployment strategies.

1. Introduction

Optical Character Recognition (OCR) is a fundamental technology that transforms text from printed or handwritten images and scanned documents into machine-readable form. Beyond enabling digitization, OCR enhances search capabilities, accessibility, and analytical procedures, supporting seamless integration of textual data into diverse applications [1]. Typical OCR systems (Figure 1) consist of several consecutive stages: pre-processing, segmentation, text detection, text recognition, and postprocessing. From another perspective, OCR can be classified into online and offline systems [2]. Online OCR receives pen-stroke coordinates during writing and performs recognition in real time, whereas offline OCR operates on static text images—printed or handwritten—after acquisition [3] . This work focuses exclusively on offline OCR, whose input is fixed images containing textual content.[3]

OCR technology is deployed in a broad range of applications, including automatic license plate recognition [4], bank check processing [5], digital libraries, and CAPTCHA verification [6].

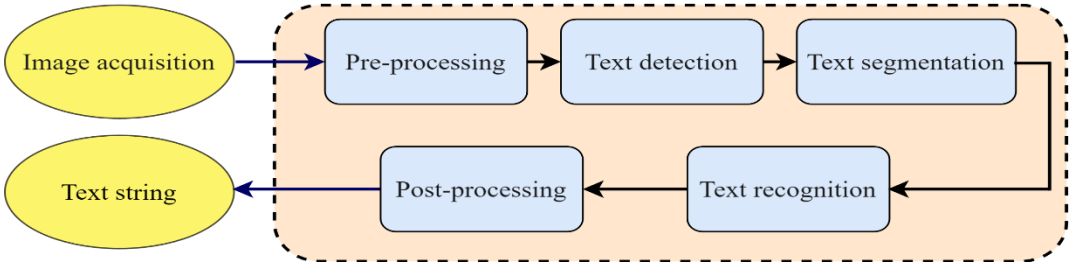


Figure 1. General overview of an OCR system [7].

However, despite decades of research, Farsi OCR remains a challenging task.

The Farsi script is cursive, with inherently connected characters whose shapes vary according to positional context (initial, medial, final, or isolated). Frequent diacritics, complex ligatures, and font diversity further complicate character segmentation and recognition.

Recent advances in deep learning have shifted OCR research toward end-to-end recognition pipelines that learn visual and sequential features jointly. Two architecture families have become particularly influential: CRNN (Convolutional Recurrent Neural Network) based models, which combine convolutional layers for spatial representation with recurrent layers for sequential modelling, and Transformer-based models, which use self-attention to capture long-range dependencies across input sequences. CRNNs are favoured for their favourable trade-off between accuracy and computational efficiency, making them suitable for deployment on resource-constrained devices. In contrast, Transformers often achieve state-of-the-art performance on large and diverse datasets, demonstrating robustness to variations in font, style, and layout.

Despite their success in OCR for Latin, a comprehensive, dataset-aware comparison of these architectures for Farsi OCR—covering both printed and handwritten text—has been largely absent. This work aims to address that gap. We evaluate five representative deep learning architectures: two lightweight CRNN-based models [8], [9] selected for efficiency, and three Transformer-based models [2], [10], [11] chosen for their advanced contextual modelling capacity. Evaluations are conducted on four benchmark datasets such as Shotor and IDPL-PFOD2 for printed text, Iranshahr and Sadri for handwritten text, using standardized accuracy metrics, parameter counts, and computational cost (GFLOPs).

To ensure a fair comparison, we rely on publicly available, pre-processed datasets, focusing exclusively on the recognition stage. Recognition involves mapping extracted visual features to character sequences using deep neural networks, followed by optional postprocessing such as error correction, language modelling, or spell checking.

The primary contributions of this work are as follows:

- Comprehensive comparative analysis of five widely used deep learning architectures for Farsi OCR, spanning both printed and handwritten scripts.
- Dataset-type-aware evaluation, presenting separate results for printed and handwritten corpora to reveal architecture-specific strengths and weaknesses.
- Inclusion of IDPL-PFOD2, a large and heterogeneous printed corpus, to assess robustness under realistic font, style, and layout variations.
- Quantitative trade-off assessment between recognition accuracy, parameter count, and computational complexity, informing deployment in resource-limited scenarios.

Henceforth, the remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the selected architectures and experimental setup; Section 4 presents and analyses the experimental results; and Section 5 concludes the paper with final remarks and directions for future research.

2. Related work

2.1. Abjad Script

Farsi, as an Abjad right-to-left language, shares similarities with Arabic, Urdu, and Pashto, including several common characters. Moreover, owing to the parallels among these languages, OCR techniques applied to one language can often be adapted for other Farsi-like languages as well.

Several researchers have employed Convolutional Neural Networks (CNNs) for character recognition. These approaches are particularly focused on individually written characters, thereby limiting their application in OCR to the recognition of discrete letters and digits [12]. For instance, the authors of [13] employed pre-trained AlexNet and GoogleNet models on the ImageNet dataset to

classify 54 handwritten characters and digits in the Urdu language. The study presented in [14] combines CNN and RNN (recurrent neural network) architectures for recognizing Arabic text in natural scene images. This approach incorporates an attention mechanism to focus on the most informative regions of the image, thereby enhancing text recognition accuracy. The authors of [15] apply a CNN-based classifier for printed Arabic text recognition, utilizing a combination of texture, shape, and statistical features.

Utilizing a multi-dimensional LSTM (long short-term memory) with a right-to-left sliding window in the Urdu Nastaliq language led to achieving a recognition accuracy of 94.97% for printed text lines from the UPTI dataset [16]. Furthermore, researchers applied a multi-dimensional LSTM to Pashto image data, enabling their OCR system to achieve high recognition accuracy directly from raw image data. The study presented in [17] employs an Arabic OCR model that utilizes BLSTM networks in conjunction with CTC (Connectionist Temporal Classification) to predict relevant Arabic text sequences, incorporating a linguistic model during prediction to enhance output accuracy.

The study in [14] employs a convolutional structure to extract features from discrete Arabic characters, followed by classification using a support vector machine. In 18, a two-way multi-directional LSTM (MD BLSTM) combined with CTC is used to recognize Arabic printed texts without requiring character separation. Additionally, the approach presented in [19] utilizes a CNN-BLSTM-CTC architecture for recognizing Arabic handwritten texts.

2.2. Farsi Language

Reference [20] employs deep neural networks, utilizing DenseNet and Xception architectures, to develop a handwritten Farsi text recognition system. The study in [21] constructs a dictionary of primary characters and subwords based on Farsi sub-word markers, which is subsequently used to train a classifier. This approach integrates a CNN-based classifier with an autoencoder (AE) network for feature extraction, enhancing recognition performance. The study presented in [22] proposes an innovative method for recognizing handwritten Farsi words, employing a sequential indexing strategy alongside a deep convolutional neural network for feature extraction. The extracted feature sequences are then processed through a stack of BiLSTM (bidirectional LSTM) layers, culminating in a CTC layer to predict the corresponding characters. The authors of [23] present an efficient method for recognizing handwritten Farsi phone numbers, combining computer vision operations with deep convolutional layers. This process includes a segmentation algorithm to decompose numeric strings into individual digits, which are subsequently classified using a specialized recognition algorithm.

The study presented in [24] develops a deep learning-based document layout analysis (DLA) and text line detection (TLD) system tailored for Farsi scripts, which improved Tesseract OCR accuracy by 2.8% on Iranian newspaper datasets. Their method combines deep models with post-processing heuristics such as font size normalization and line curvature elimination. The authors of [25] design a lightweight deep convolutional network (LWDCNN) to predict personality traits from Farsi handwriting, highlighting the richness of information embedded in handwritten Farsi beyond textual content.

These studies further highlight the need for robust, flexible models capable of handling the complexities of Farsi document structure and handwriting style.

3. Methodology

In recent years, deep learning has significantly advanced OCR technology, achieving state-of-the-art performance in text recognition tasks [7]. In this study, we examine different deep learning architectures for Farsi text recognition, focusing on established models that deliver reliable performance and serve as strong baselines for comparative analysis (see Fig. 2). To this end, we categorize text recognition methods into two main groups: Transformer-free and Transformer-based approaches [7]. Transformer-free methods include CRNN-based and attention-based CRNN architectures, where the former employs the CTC algorithm for sequence-to-text alignment, and the latter leverages attention mechanisms to emphasize informative regions of the image. Transformer-based methods, on the other hand, rely on self-attention mechanisms for recognition and can be further divided into convolutional Transformer-based (CvT) and convolution-free (vanilla) variants. In this study, both families of architectures are considered to comprehensively evaluate their effectiveness.

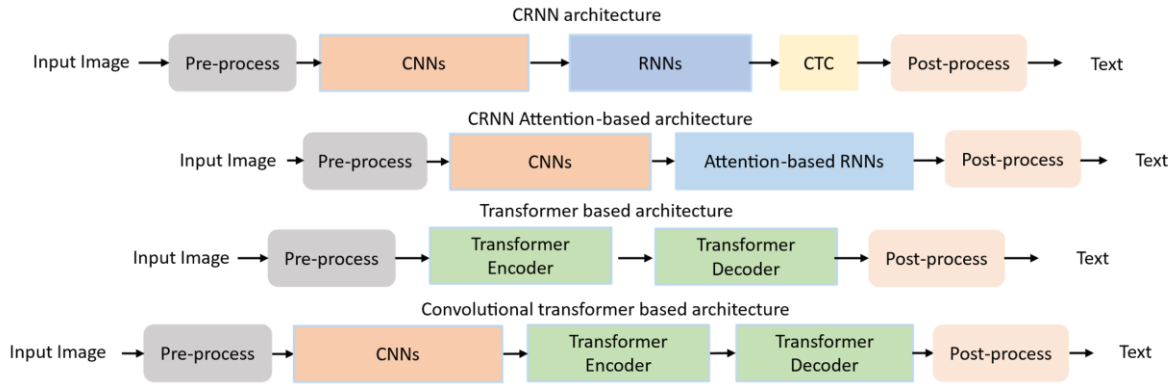


Figure 2. General overview of the selected text recognition approaches based on their core building blocks.

3.1. Transformer-free approaches

3.1.1. lightweight CRNN

As shown in Figure 2(a), the CRNN model is composed of three main components. First, a convolutional backbone (i.e., MobileNetV3) serves as the encoder by extracting meaningful features from input images. Next, BiLSTM layers function as the decoder by capturing contextual dependencies across the sequence.

Finally, a Connectionist Temporal Classification module acts as the transcription layer, which converts the sequential features into readable text. Subsequently, to improve the recognition accuracy, we use an internal language modeling, similar to [26]. Input images are first scaled to a fixed dimension, then feature maps are extracted and converted into sequential feature vectors for the recurrent layers. Each feature vector corresponds to a local receptive field in the original image. To handle sequences of varying lengths, the CNN output is fed into the RNN, which models the sequential dependencies. Upon top of the convolutional layers, the recurrent layer consists of a deep bidirectional recurrent neural network. The task of the recurrent layers is to predict the label distribution y_t for each frame x_t, in the feature sequence $x = x_1 ... x_t$ RNN networks are able to back-propagate the error from their output layer to the previous layers in the network, allowing for integrated training of the entire network.

This CRNN architecture uses a bidirectional version of LSTM as the recurrent layer to avoid the problems of both vanishing gradient and explosion, which usually arise when training long sequences. This allows the network to process extracted sequences with variable lengths, making it easier for the network to handle variable-length text images. LSTM is a directed network that uses only past data.

However, when dealing with image-based sequences, accessing future contexts is just as useful and complementary as accessing past contexts. Therefore, similar to the method in [27], we chose the BLSTM network to access bidirectional information. The transcription layer takes the per-frame predictions made by the RNN and uses a fully connected network with a SoftMax activation.

3.1.2. Attention-based CRNN

The CRNN attention-based model extends the standard CRNN by replacing the CTC transcription with an attention mechanism. While the traditional CRNN with CTC allows alignment-free sequence prediction, it often struggles with long-range dependencies, alignment inaccuracies, and redundant outputs that require post-processing. In contrast, the attention-based CRNN explicitly aligns the input and output sequences through attention weights, effectively capturing contextual information and long-range dependencies [8]. This leads to improved recognition accuracy and greater flexibility in handling variable-length inputs and outputs, as the model adaptively focuses on the most relevant parts of the input at each decoding step (see Figure 2(b)).

In the structure of the CRNN attention-based method, the encoder combines convolutional and recurrent layers to extract sequential features from word images of variable size. Initially, convolutional layers (i.e., MobileNetV3) generate feature maps that capture local visual patterns, which are then converted into sequences of vectors aligned from left to right, where each vector corresponds to a receptive field in the original image. To expand the contextual understanding beyond local patterns, a two-layer BiLSTM processes these vectors, capturing long-range dependencies in both forward and backward directions and producing a refined sequential representation. Then, the decoder generates the

character sequence step by step using the encoder's output. At each time step, an attention mechanism assigns weights to the encoded vectors, highlighting the most relevant regions for predicting the next character. A GRU-based recurrent structure then uses this focused representation to output the next character, repeating the process until the end-of-sequence token is reached. This explicit alignment enables the model to handle variable-length sequences more effectively and improves recognition accuracy compared to CTC-based CRNNs. Note that a post-processing approach similar to that used in CRNN-based models is applied here.

3.2. Transformer-based approaches

Transformers, neural network architectures based on the self-attention mechanism, have demonstrated remarkable success in both computer vision and NLP (natural language processing) tasks [26]. Initially developed for language modeling and machine translation, they excel at capturing long-range dependencies and offer strong representation capabilities [28], [29]. Unlike CRNN-based models, Transformers can process sequences in parallel, enabling significantly faster training.

3.2.1. Vision Transformer Text Recognition Architecture (ViT)

The Vision Transformer is an encoder-only transformer-based architecture adapted for computer vision tasks, including OCR (see Figure 2(c)) [10]. Unlike traditional CRNN-based networks, ViT divides an input image into non-overlapping patches, treating each patch as a "token" similar to words in NLP. These patches are embedded into 1D vectors and combined with positional encodings to preserve spatial information. The sequence of embeddings is then processed through a multi-head self-attention Transformer encoder, capturing long-range dependencies and contextual relationships across the image. For OCR, the encoder outputs a sequence of feature vectors representing the text, which is fed into a decoder to generate the corresponding characters or words. ViT enables end-to-end training, parallel processing, and effective handling of long-range dependencies, making it highly suitable for recognizing complex textual patterns without extensive post-processing steps.

3.2.2. The SVTR model

The SVTR model [2] introduces a novel approach for text recognition by decomposing input images into small 2D patches called character components, which are then tokenized and processed using self-attention. The core building blocks of this approach follow Figure 2(d). Its three-stage backbone employs mixing, merging, and combining operations to extract both local stroke-level features (i.e., convolutional operations) and long-range dependencies across characters. Global mixing captures relationships among different characters, while local mixing encodes morphological details of individual characters. This patch-wise design eliminates the need for explicit sequence modeling and relies on a single visual model, enabling flexible, interpretable, and multi-grained feature extraction. SVTR's parameters can be estimated using Bayesian inference, maximum likelihood, or iterative optimization methods, and its architecture effectively balances local and global context for accurate text recognition.

3.2.3. The CvT combined with iterative language modeling approach

The proposed method consists of two main units: a vision model (VM) and a language model (LM). The vision model is based on a convolutional Transformer architecture, which enhances both performance and memory efficiency compared to standard Transformer models. It combines a hierarchical Transformer structure with convolutional token embedding and convolutional Transformer blocks (see Figure 2(d)). This design captures both local and global visual features, enabling effective text recognition while using fewer parameters and computational resources. The language model operates under an autonomous strategy, treating the LM as an independent spelling correction module. This allows for flexible integration of more advanced models and ensures explicit acquisition of linguistic knowledge [11].

A bidirectional representation is employed to capture richer contextual information compared to conventional unidirectional or ensemble approaches. This is achieved using a Bidirectional Character Network (BCN), a variant of a Transformer decoder, which efficiently models character sequences while preventing information leakage through attention masking. To handle noisy inputs from the vision model, an iterative correction mechanism is applied. The LM is executed multiple times, progressively refining predictions and improving overall text recognition performance. Essentially, the VM and the language model (LM) operate in separate modalities, with the VM processing images and the LM processing text. To effectively align visual and linguistic

information, a gated fusion mechanism is used for the final prediction [11].

4. Experimental results

This section provides the experimental results, and setting their implications. An Intel(R) Core (TM) i9-10900k @ 3.70GHz processor, 32 GB RAM, 1024 GB SSD, and a dedicated NVIDIA GeForce RTX 3070 GPU with 8GB GDDR6 memory are used in every experiment. The complete implementations of studied methods are publicly available at GitHub Repository[1] to ensure reproducibility and allow further extensions by the research community.

4.1. Datasets

This research assesses the approach by employing four well-known Farsi word-level datasets, namely Sadri [30], Iranshahr [31], IDPL-PFOD2 [10], and Shotor[2].

4.1.1. Shotor dataset

The collection includes printed Farsi literature and is mostly focused on word-level analysis. It has 120,000 grayscale images, all standardized to a 50*100 scale, that show significant text in a range of typefaces and sizes. This large collection of images was assembled using texts from Ganjoor and Farsi Wikipedia. Figure 3 shows some examples of Shotor dataset images.

Figure 3. Examples of Shotor dataset images.

4.1.2. Iranshahr dataset

The Iranshahr dataset, consisting of 19,583 images, features 503 names of cities in Iran. Each class within the dataset contains at least 20 samples, resulting in approximately 30 to 40 distinct images per class. These images exhibit a diverse array of handwriting styles. Figure 4 shows sample images collected from the Iranshahr dataset.

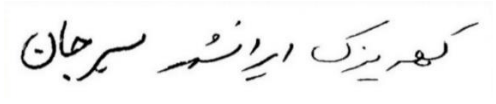


Figure 4. Sample images collected from the Iranshahr dataset.

4.1.3. Sadri dataset

Images with a wide variety of textual elements, such as words, text, numbers, letters, dates with text and numbers, signs, and symbols, are included in this collection. 62,500 words total, divided into 125-word classes. To compile this dataset, 500 Farsi language writers, comprising both male and female individuals, were enlisted to randomly generate the specified number of words.

Notably, 10% of these writers were left-handed. Figure 5 shows some examples from the Sadri dataset.

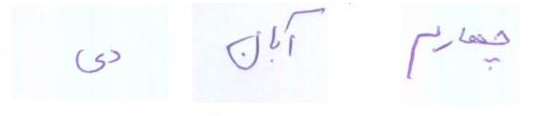


Figure 5. Some examples of Sadri dataset.

4.1.4. IDPL-PFOD2 dataset

IDPL-PFOD2 is a large-scale artificial image dataset of printed Farsi text, containing 2,003,541 images in `.png` format. Each image represents a single line of real Farsi text with dimensions of 300 × 50 pixels. The dataset includes a wide variety of fonts, styles, and sizes, extending the previous IDPL-PFOD dataset in both volume and diversity. It is designed to support the training and evaluation of deep learning-based OCR models for Farsi text. Figure 6 shows some examples from the IDPL-PFOD2 dataset.

Figure 6. Some examples of IDPL-PFOD2 dataset.

4.2. Training details

To ensure a fair comparison, all models were trained under carefully standardized protocols, with consistent preprocessing, optimization strategies, and evaluation procedures.

The CRNN-based model resized input images to 100×32 for both training and testing. Training employed the Adam optimizer with an initial

[1] https://github.com/ftmasadi/BehNasr-OCR

[2] https://github.com/amirabbasasadi/Shotor

learning rate of 0.0005 and L2 regularization, while validation was performed at the end of each epoch. In the Vision Transformer model, input images were resized to 32×128. The model used the Adadelta optimizer with a learning rate of 1.0, alongside L1 regularization to stabilize training.

The SVTR architecture incorporated a rectification module to correct distorted text, with images resized to 32×64. Training utilized the AdamW optimizer with a weight decay of 0.05 and a cosine learning rate scheduler with a 2-epoch warm-up. The initial learning rate was scaled relative to batch size ($\frac{5e^{-4} \times batchsize}{2048}$), with extensive data augmentations (rotation, perspective distortion, motion blur, Gaussian noise).

Finally, the CvT-based model resized images to 32×128 and applied augmentations such as perspective, affine, and rotation transformations. The network employed a model dimension (D) of 512, with a BCN block of 4 layers and 8 attention heads. Training used the Adam optimizer with L2 regularization and a learning rate starting at $1e^{-4}$, reduced to $1e^{-5}$ after 6 epochs.

All models were trained with a batch size of 64 and number of epochs equals to 300. Also, to mitigate overfitting an early stopping strategy with patience of 10 epochs is applied.

4.3. Evaluation on Accuracy

Here, the efficacy of the models is evaluated using word-level accuracy (WLA). Specifically, a word is considered correct only if it is predicted entirely without error. The accuracy is thus defined as the ratio of the number of completely correct words (CWrds) to the total number of words (AllWrds) in the ground truth, as follows:

$$\frac{CWrds}{AllWrds} \times 100 \quad (1)$$

Table 1 summarizes the accuracy results for all models across the four benchmark datasets. The first five rows correspond to models directly evaluated in this study. Some accuracy values for these models have been previously reported in earlier works [2], [8], [10], [11] while those marked with * indicate new results first presented in this paper. The analysis reveals several notable patterns:

High accuracy of CRNNs on homogeneous datasets: Both CRNN-CTC and CRNN-Attention achieve WLA above 96% on Shotor, Iranshahr, and Sadri, with CRNN-CTC reaching up to 99.42% on Shotor. This indicates their ability to effectively recognize text in relatively small and less diverse datasets.

Performance drops on large and diverse corpora: On IDPL-PFOD2, which contains over 2 million printed text lines with substantial variation in font, style, and layout, CRNN models experience a sharp performance decline to 78.49% for CRNN-CTC highlighting their limited generalization under high variability.

Table 1. Accuracy (%) and model complexity comparison of evaluated Farsi OCR architectures and prior works on four datasets. (* Results first reported in this paper).

	Method	IDPL-PFOD2	Shotor	Sadri	Iranshahr	FLOPs ($\times 10^9$)	Params ($\times 10^6$)
Evaluated models	**CvT** [11]	**92.81***	**99.97**	**99.25**	97.59	7.20	37.00
	SVTR [2]	78.12*	99.75	99.23	**97.87***	3.55	22.66
	ViT [10]	81.32	99.91*	97.70*	81.15*	4.60	21.50
	CRNN attention [8]	–	99.33*	98.99	96.83	2.00	10.80
	CRNN	78.49	99.42*	98.86*	97.08*	**1.40**	**8.50**
State-of-the-art	CTC + BiLSTM + CNN [22]	–	–	98.80	–	–	–
	Various CNN models [20]	–	–	98.80	–	–	–
	DCNN [32]	–	–	98.60	94.60	–	–
	RNN [33]	–	–	–	84.30	–	–
	CNN + AE [21]	–	–	–	91.09	–	–

Robustness of Transformer architectures: Transformer-based models demonstrate greater resilience to dataset diversity. SVTR achieves 92.81% on IDPL-PFOD2, substantially reducing the gap to homogeneous dataset performance. ViTOCR and CvTLM also exceed 90%, maintaining superior results in both printed and handwritten domains.
Domain-independent consistency: While Shotor (printed) and Sadri (handwritten) differ in script source, Transformers maintain high WLA across both, suggesting better adaptability to cross-domain variations compared to CRNNs.
These results underscore that dataset characteristics—size, diversity, and complexity—should guide model selection. CRNNs remain competitive in constrained scenarios with limited computational resources, whereas Transformers are preferable for large-scale, real-world deployments requiring robustness to extensive visual variability.

4.4. Qualitative Analysis

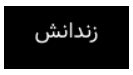
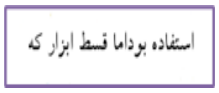
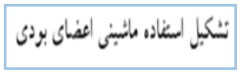
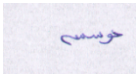
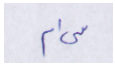
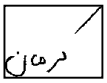
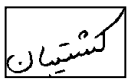
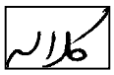
To complement the numerical findings in Section 4.4, a qualitative analysis was carried out using representative samples from all four datasets. Table 2 illustrates example input images, their corresponding ground-truth labels, and the recognition outputs produced by the five evaluated models. In the table, green-shaded cells indicate correct predictions, while non-highlighted cells denote erroneous outputs.
Dataset-wise observations
Shotor (printed, word-level): The majority of models, including CRNN and Transformer architectures, produced perfectly correct predictions for simple, clear prints (e.g., "زندانش"). Minor issues appear in compound words or where diacritic placement is subtle, as seen in "نمخاست" without spacing by CRNN-Attention, reflecting its less consistent word-boundary handling.
IDPL-PFOD2 (printed, line-level, high diversity): This dataset revealed the most pronounced differences between architectures. Long lines such as "که ابزار قسط بوداما استفاده" degraded heavily for CRNN models, which lost ligatures ("استده") or substituted glyphs incorrectly ("قسفظ") and generated artifacts. Transformer-based models—especially CvT and ViT—preserved more of the sequence structure but still occasionally merged or split words ("استفادهببوام"). Similar behavior was observed for "نمایند فارس بیرنگ که بازخوانی," where CRNN variants produced incomplete reconstructions, while SVTR approximated the text more closely but misrecognized certain words ("ناین" instead of "نمایند").
Sadri (handwritten, mixed content): In handwritten samples such as "ام سی" and "موسسه," most models achieved correct recognition, but CRNN-Attention dropped entire characters ("سام") or omitted diacritics in complex calligraphy. Transformers—especially CvT—retained better character continuity in variable stroke widths.
Iranshahr (handwritten, city names): All models succeeded on simpler tokens like "کلاله," but confused visually close letters in certain city names. For example, "کرن" was substituted for "کرمان" by ViT and CRNN, while CvT and SVTR maintained correctness. Errors here are often due to incomplete loop closure or faint strokes in handwriting, which Transformers handle better due to their global context modeling.

Error patterns across models: Similar-looking glyphs: Omissions or substitutions involving dot-placement (e.g., ب vs. پ and ق vs. ف) were frequent in CRNN models, modestly present in ViT, and least frequent in CvT and SVTR.
Loss of ligatures in printed text: Especially in dense fonts, CRNNs either split characters incorrectly or merged them into meaningless tokens.
Boundary inconsistencies in handwritten text: CRNN-Attention, in particular, merged adjacent words more often, suggesting limitations in contextual disambiguation.
These qualitative findings align with the accuracy analysis: CRNN models excel on short, uniform inputs, while Transformers—though not immune to errors—consistently maintain text structure and readability under higher variability. This complementary perspective emphasizes that OCR model selection for Farsi should consider dataset diversity, input complexity, and resource constraints.

4.5. Model Complexity Analysis

A detailed comparison of model sizes and computational demands is provided in Table 2, reporting the number of parameters (Params, in millions) and floating-point operations (FLOPs, in billions) alongside accuracy results for the four benchmark datasets.
CRNN recorded the smallest footprint with 8.5 M parameters and 1.4 B FLOPs, confirming its efficiency in resource-constrained settings. Attention-based CRNN increased slightly to 10.8 M parameters and 1.9 B FLOPs due to the attention mechanism.

Table 2. Sample text recognition predictions versus ground truth from different Farsi text datasets.

Dataset	Input image	Proposed methods				
		CRNN	CRNN attention	ViT	SVTR	CvT
Shotor		نم خاست	نمخاست	نم خاست	نم خاست	نمی خاست
		زندانش	زندانش	زندانش	زندانش	زندانش
		قله ا	قله ا	قله ا	قله ا	قلهای
IDPL-PFOD2		استفادهبیوام قسط ابزار ک	-	استفادهبیوام قسط ابزار ک	استده بواما قسفظ آبزآر که	استفاده بوداما قسط ابزار که
		بازخانی که بینگفرس ننایی		بازخانی که بینگفرس ننایی	باش که بیرنگ فاران ناین	بازخوانی که بیرنگ فارس نمایند
		تشکل اتفاه مماییضضیی بوو		تشکل اتفاه مماییضضیی بوو	نکیل استفا ماش عضای بردی	تشکیل استفاده ماشینی اعضای بود
Sadri		وسسه	موسسه	وسسه	موسسه	موسسه
		سام	-	س ام	س ام	سی ام
Iranshahr		کرن	کردان	کرن	کرمان	کرمان
		کتیان	کشتیبان	کتیان	کشتیبان	کشتیبان
		ملاله	کلاله	ملاله	کلاله	کلاله

SVTR and ViT had larger sizes (22.66 M and 21.5 M parameters, respectively), requiring 3.55 B and 4.6 B FLOPs.

CNN + ViT + LM represented the most complex configuration with 37 M parameters and 7.2 B FLOPs, a result of its integrated language model module.

This quantitative profile establishes the computational characteristics of each architecture.

4.6. Comparison with the State-of-the-art

Table 1 contrasts the five evaluated architectures CvT, SVTR, ViT, CRNN attention, and CRNN with representative prior methods.

While earlier works report competitive accuracies on individual datasets, they often lack full coverage of the four benchmark corpora or omit computational metrics, making direct one-to-one comparison incomplete.

Our evaluation uniquely provides:

Uniform protocol across datasets — all five architectures were tested on Shotor, IDPL-PFOD2,

Sadri, and Iranshahr under identical conditions, ensuring that observed differences stem from architectural design rather than dataset or preprocessing variations.

Joint accuracy–efficiency view — prior works rarely report both recognition performance and complexity; here, model size and FLOPs accompany every result, enabling a fair assessment of efficiency.

Broad dataset scope — unlike studies focused on a single printed or handwritten dataset, this work includes both highly homogeneous corpora (e.g., Shotor) and the more challenging, heterogeneous IDPL-PFOD2.

Overall, while some earlier systems still hold edge cases of superior accuracy on specific datasets, this is — to our knowledge — the first SOTA comparison in Farsi OCR that spans multiple architectures, printed and handwritten scripts, and integrates computational cost into the performance discussion. This broader framing offers a more practical basis for model selection in real-world deployments than accuracy alone.

5. Discussion

Our experimental results highlight that there is no one-size-fits-all architecture for Farsi OCR. Instead, the optimal choice depends on dataset characteristics and deployment constraints.

CRNN-based models remain competitive on smaller or more homogeneous datasets such as Shotor, Iranshahr, and Sadri, achieving up to 99.42% word-level accuracy with substantially fewer parameters and operations. However, their performance drops sharply on large, heterogeneous datasets like IDPL-PFOD2 (–21% accuracy gap compared to SVTR), indicating lower robustness to high intra-dataset variability.

In contrast, Transformer-based approaches, particularly SVTR, resist accuracy degradation under such diversity, reaching 92.81% WLA on IDPL-PFOD2 while maintaining a moderate computational footprint (3.55 G FLOPs, 22.66 M parameters). CvT delivers leading accuracy on clean printed text but at the cost of nearly five times the FLOPs of the CRNN baseline.

Section 4.4.1 revealed model-specific error patterns—such as merged letter boundaries in handwritten text or visual confusion between similar glyphs—that explain some performance gaps. CRNN architectures tend to suffer from segmentation ambiguities in cursive handwritten samples, while Transformer-based models better preserve text structure and legibility under challenging layouts. However, the latter can misinterpret visually similar letters in tightly spaced text, as observed in the Iranshahr results.

For OCR deployment on resource-constrained devices (mobile scanning apps, low-power embedded systems), CRNN or CRNN attention offers the best balance—high accuracy on focused corpora, with low RAM and power requirements.

For broad-coverage, large-scale digitization projects (e.g., national archives, documents of mixed quality), a Transformer-based solution like SVTR may justify higher computational costs to reduce error propagation in downstream indexing and search systems.

CvT may be preferable when the highest possible accuracy on clean, high-resolution input is a priority, and hardware resources are abundant.

Our evaluation focused on offline text recognition with fixed input dimensions; real-world deployments may require handling variable-length sequences, multilingual content, and noisy acquisition. Enhancing models with domain-adaptive pretraining, synthetic data augmentation, and more sophisticated postprocessing (e.g., dictionary-constrained decoding) could further improve robustness. Evaluating the architectures in low-resource training scenarios would also inform their suitability for rapid adaptation to other Abjad scripts.

Overall, this study emphasizes that matching model architecture to the operational and data profile of the problem is key to optimal Farsi OCR, and that balanced evaluation across both accuracy and complexity dimensions provides the clearest guidance for practitioners.

6. Conclusion

This work presented the first dataset-inclusive, protocol-consistent comparison of five representative deep learning architectures for Farsi OCR—two CRNN-based models (with and without attention) and three Transformer-based models—across four publicly available datasets covering both printed and handwritten scripts (Shotor, IDPL-PFOD2, Iranshahr, Sadri). By jointly reporting word-level accuracy, parameter count, and computational cost (FLOPs), the study established a transparent performance–efficiency profile for each architecture.

Results confirmed that CRNN-based models, despite their lightweight design (as low as 8.5 M parameters and 1.4 G FLOPs), can achieve near-perfect accuracy on smaller, more homogeneous corpora, making them ideal for resource-constrained deployments such as mobile and embedded OCR. Conversely, Transformer-based solutions—particularly SVTR—demonstrated greater robustness on the

highly diverse IDPL-PFOD2 dataset, closing the accuracy gap under challenging layout and font variability, while CvT achieved the best accuracy on clean printed text.

Compared to previous studies, our unified evaluation protocol and inclusion of computational metrics deliver a more complete benchmark for both academic comparison and real-world decision-making. The findings emphasize that model choice should be dataset-aware, balancing accuracy and computational demands according to operational needs.

Future work will explore domain-adaptive pretraining, multilingual extensions to other Abjad scripts, and deployment-oriented optimizations to further bridge the gap between cutting-edge accuracy and efficiency.

References

[1] C. Indravadanbhai Patel, D. Patel, C. Patel Smt Chandaben Mohanbhai, A. Patel, S. Chandaben Mohanbhai, and D. Patel Smt Chandaben Mohanbhai, "Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study Scholar Model of Images, Objects and Superpixels View project CHARUSAT Apps (Mobile Application) View project Optical Character Recognition by Open Source OCR Tool Tesseract: A," *Artic. Int. J. Comput. Appl.*, vol. 55, no. 10, pp. 975–8887, 2012.

[2] F. Asadi-Zeydabadi, E. Shabaninia, H. Nezamabadi-Pour, and M. Shojaee, "Farsi Optical Character Recognition Using a Transformer-based Model," in *2023 13th International Conference on Computer and Knowledge Engineering, ICCKE 2023*, IEEE, 2023, pp. 293–299. doi: 10.1109/ICCKE60553.2023.10326255.

[3] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of Post-OCR Processing Approaches," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, 2022, doi: 10.1145/3453476.

[4] E. Shabaninia, F. Asadi, and H. Nezamabadi_pour, "Enhancing License Plate Recognition Using a Language Model-Based Approach," *J. Mach. Vis. Image Process.*, vol. 11, no. 4, pp. 15–26, 2025.

[5] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014, doi: 10.1016/j.eswa.2014.07.008.

[6] R. Gossweiler, M. Kamvar, and S. Baluja, "What's up CAPTCHA? A CAPTCHA based on image orientation," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 841–850.

[7] A. Afkari-Fahandari, E. Shabaninia, F. Asadi-Zeydabadi, and H. Nezamabadi-Pour, "A Comprehensive Survey of Transformers in Text Recognition: Techniques, Challenges, and Future Directions," *ACM Comput. Surv.*, vol. 58, no. 5, p. 42, 2025.

[8] A. Afkari-Fahandari, F. Asadi-Zeydabadi, E. Shabaninia, and H. Nezamabadi-pour, "Farsi Handwritten Text Recognition via a Lightweight Attention-Driven Sequence Recognition Network," in *2024 19th Iranian Conference on Intelligent Systems (ICIS)*, IEEE, 2024, pp. 24–29.

[9] D. V. Sang and L. T. B. Cuong, "Improving CRNN with EfficientNet-like feature extractor and multi-head atention for text recognition," in *ACM International Conference Proceeding Series*, 2019, pp. 285–290. doi: 10.1145/3368926.3369689.

[10] F. Asadi-zeydabadi, A. Afkari-Fahandari, A. Faraji, E. Shabaninia, and H. Nezamabadi-Pour, "IDPL-PFOD2: A New Large-Scale Dataset for Printed Farsi Optical Character Recognition," *arXiv Prepr. arXiv*, 2023, [Online]. Available: https://arxiv.org/abs/2312.01177.

[11] A. Afkari-Fahandari, F. Asadi-Zeydabadi, E. Shabaninia, and H. Nezamabadi-Pour, "Enhancing Farsi Text Recognition via Iteratively Using a Language Model," in *2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing, AISP 2024*, IEEE, 2024, pp. 1–6. doi: 10.1109/AISP61396.2024.10475269.

[12] A. Nasr-Esfahani, M. Bekrani, and R. Rajabi, "Robust Persian Digit Recognition in Noisy Environments Using Hybrid CNN-BiGRU Model," *J. AI Data Min.*, vol. 13, no. 3, pp. 337–345, 2025, doi: 10.22044/jadm.2025.15932.2707.

[13] M. A. KO and S. Poruran, "OCR-nets: variants of pre-trained CNN for Urdu handwritten character recognition via transfer learning," *Procedia Comput. Sci.*, vol. 171, pp. 2294–2301, 2020.

[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 Comput. Sci.*, vol. 80, pp. 1712–1723, 2016.

[15] L. Bouchakour, F. Meziani, H. Latrache, K. Ghribi, and M. Yahiaoui, "Printed arabic characters recognition using combined features and cnn classifier," in *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, IEEE, 2021, pp. 1–5.

[16] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, and T. Breuel, "Scale and rotation invariant OCR for Pashto cursive script using MDLSTM network," in *2015 13th international conference on document analysis and recognition (ICDAR)*, IEEE, 2015, pp. 1101–1105.

[17] S. Rawls, H. Cao, E. Sabir, and P. Natarajan, "Combining deep learning and language modeling for segmentation-free OCR from raw pixels," in *2017 1st international workshop on Arabic script analysis and recognition (ASAR)*, IEEE, 2017, pp. 119–123.

[18] S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, and F. Shafait, "Urdu Nasta'liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks," *Springerplus*, vol. 5, no. 1, p. 2010, 2016.

[19] R. Maalej and M. Kherallah, "Convolutional neural network and BLSTM for offline Arabic handwriting recognition," in *2018 International Arab conference on information technology (ACIT)*, IEEE, 2018, pp. 1–6.

[20] M. Bonyani, S. Jahangard, and M. Daneshmand, "Persian handwritten digit, character and word recognition using deep learning," *Int. J. Doc. Anal. Recognit.*, vol. 24, no. 1, pp. 133–143, 2021.

[21] S. Khosravi and A. Chalechale, "Recognition of Persian/Arabic handwritten words using a combination of convolutional neural networks and autoencoder (AECNN)," *Math. Probl. Eng.*, vol. 2022, no. 1, p. 4241016, 2022.

[22] V. M. Safarzadeh and P. Jafarzadeh, "Offline Persian handwriting recognition with CNN and RNN-CTC," in *2020 25th international computer conference, computer society of Iran (CSICC)*, IEEE, 2020, pp. 1–10.

[23] M. Akhlaghi and V. Ghods, "Farsi handwritten phone number recognition using deep learning," *SN Appl. Sci.*, vol. 2, no. 3, p. 408, 2020.

[24] A. Fateh, M. Fateh, and V. Abolghasemi, "Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection," *Eng. Reports*, vol. 6, no. 9, p. e12832, 2024.

[25] M. S. Anari, K. Rezaee, and A. Ahmadi, "TraitLWNet: a novel predictor of personality trait by analyzing Persian handwriting based on lightweight deep convolutional neural network," *Multimed. Tools Appl.*, vol. 81, no. 8, pp. 10673–10693, 2022.

[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ACM International Conference Proceeding Series*, 2006, pp. 369–376. doi: 10.1145/1143844.1143891.

[27] N. A. M. Isheawy and H. Hasan, "Optical character recognition (OCR) system," *IOSR J. Comput. Eng. (IOSR-JCE), e-ISSN*, pp. 661–2278, 2015.

[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, pp. 4171–4186.

[29] T.Brown *et al.*, "Language models are few-shot learners, in *Advances in neural information processing systems*,2020, pp. 1877–1901.

[30] F. olimanpour, J. Sadri, and C. Y. Suen, "Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language," in *Tenth International workshop on Frontiers in handwriting recognition*, Suvisoft, 2006.

[31] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban, and S. M. Golzan, "A comprehensive isolated Farsi/Arabic character database for handwritten OCR research," in *Tenth international workshop on frontiers in handwriting recognition*, Suvisoft, 2006.

[32] A. Zohrevand and Z. Imani, "Holistic persian handwritten word recognition using convolutional neural network," *Int. J. Eng.*, vol. 34, no. 8, pp. 2028–2037, 2021.

[33] M. F. Y. Ghadikolaie, E. Kabir, and F. Razzazi, "Sub-word based offline handwritten farsi word recognition using recurrent neural network," *ETRI J.*, vol. 38, no. 4, pp. 703–713, 2016.

مقایسه روش‌های یادگیری عمیق در سامانه‌های OCR فارسی چاپی و دست‌نویس

فاطمه اسدی زیدآبادی[۱]، علی افکاری فهندری[۲]، الهام شعبانی نیا[۳،*] و حسین نظام آبادی پور[۴]

[۱] دانشکده مهندسی برق، دانشگاه شهید باهنر کرمان، کرمان، ایران.

[۲] دانشکده ریاضی کاربردی، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران.

ارسال ۲۰۲۵/۰۵/۰۴ بازنگری ۲۰۲۵/۰۷/۲۸؛ پذیرش ۲۰۲۵/۱۰/۲۴

چکیده:

بازشناسی نوری حروف فارسی همچنان چالشی است؛ زیرا خط فارسی ماهیتی پیوسته دارد، شکل نویسه‌ها بسته به موقعیت آن‌ها تغییر می‌کند و استفاده از اعراب نیز در آن رایج است. در این مطالعه، ارزیابی مقایسه‌ای پنج معماری بنیادی یادگیری عمیق که به‌طور گسترده در سامانه‌های OCR به‌کار می‌روند انجام شده است: دو مدل سبک مبتنی بر CRNN با هدف پیاده‌سازی کارآمد، و سه مدل مبتنی بر ترنسفرمر با قابلیت بالاتر در مدل‌سازی بافتار. هدف، بررسی میزان تناسب این معماری‌ها با ویژگی‌های خاص خط فارسی است. عملکرد مدل‌ها با استفاده از چهار پایگاه داده عمومی ارزیابی شد: دو پایگاه داده شتر و IDPL-PFOD2 برای متون چاپی و دو پایگاه داده ایرانشهر و صدری برای متون دست‌نویس. معیارهای ارزیابی شامل دقت سطح کلمه، تعداد پارامترها و هزینه محاسباتی بود. نتایج نشان داد مدل‌های مبتنی بر CRNN در پایگاه‌های داده سطح کلمه، دقت بالایی کسب کرده‌اند—۹۹/۴۲ درصد در شتر، ۹۷/۰۸ درصد در ایرانشهر و ۹۸/۸۶ درصد در صدی—در حالی که حجم مدل و نیاز محاسباتی در آن‌ها کمتر است. با این حال، دقت آن‌ها در پایگاه داده بزرگ‌تر و متنوع‌تر سطح خط یعنی IDPL-PFOD2 به ۷۸/۴۹ درصد کاهش یافت. در مقابل، مدل‌های مبتنی بر ترنسفرمر این فاصله عملکرد را به‌طور قابل‌توجهی کاهش دادند و با برخورداری از پایداری بیشتر نسبت به تغییرات قلم، سبک و چیدمان، به بهترین دقت ۹۲/۸۱ درصد در IDPL-PFOD2 رسیدند. تا جایی که ما اطلاع داریم، این پژوهش از نخستین مطالعات جامع در مقایسه معماری‌های سبک مبتنی بر CRNN و معماری‌های مبتنی بر ترنسفرمر برای OCR فارسی است که هر دو متن چاپی و دست‌نویس را دربر می‌گیرد و یک خط مبنای عملکردی معتبر برای تحقیقات و برنامه‌های کاربردی آینده فراهم می‌سازد.

کلمات کلیدی: OCR فارسی، تشخیص متن فارسی، یادگیری عمیق، CRNN، ترنسفرمر، متن چاپی، متن دست‌نویس، مجموعه داده‌های معیار.