7. Model architectures

Figure 9 presents the default architectural configurations of the three HTR systems evaluated in this work: Retsinas et al. [17], PyLaia [16], and HTRFlor [3]. The model proposed by Recsinas et al. (Figure 9a) employs a deep residual convolutional neural network with multiple ResBlocks and intermediate max pooling layers, followed by a column-wise max pooling operation. Sequence modeling is performed by a single BiLSTM layer with 256 hidden units, followed by a dense output layer. Additionally, the architecture incorporates a CTC shortcut path consisting of a 1×1 convolution for intermediate supervision. PyLaia (Figure 9b) adopts a relatively shallow CNN feature extractor composed of five convolutional layers with batch normalization and LeakyReLU activations. This is followed by a stack of five BiLSTM layers, each with 256 hidden units per direction. The final output is produced by a softmax layer applied after a dense transformation.

HTRFlor (Figure 9c) is based on gated convolutions and uses PReLU activations combined with Batch Renormalization and MaxNorm regularization. The encoder consists of six gated convolutional blocks with increasing filter sizes, interleaved with dropout and normalization layers. Max pooling and tiling are applied before the recurrent decoder, which comprises two stacked BGU layers, each with 128 hidden units. The final dense layer maps the outputs to character probabilities, followed by a softmax operation. In our experiments, we use the unmodified, default versions of all three architectures as provided in their respective implementations.

8. PyLaia model architectures

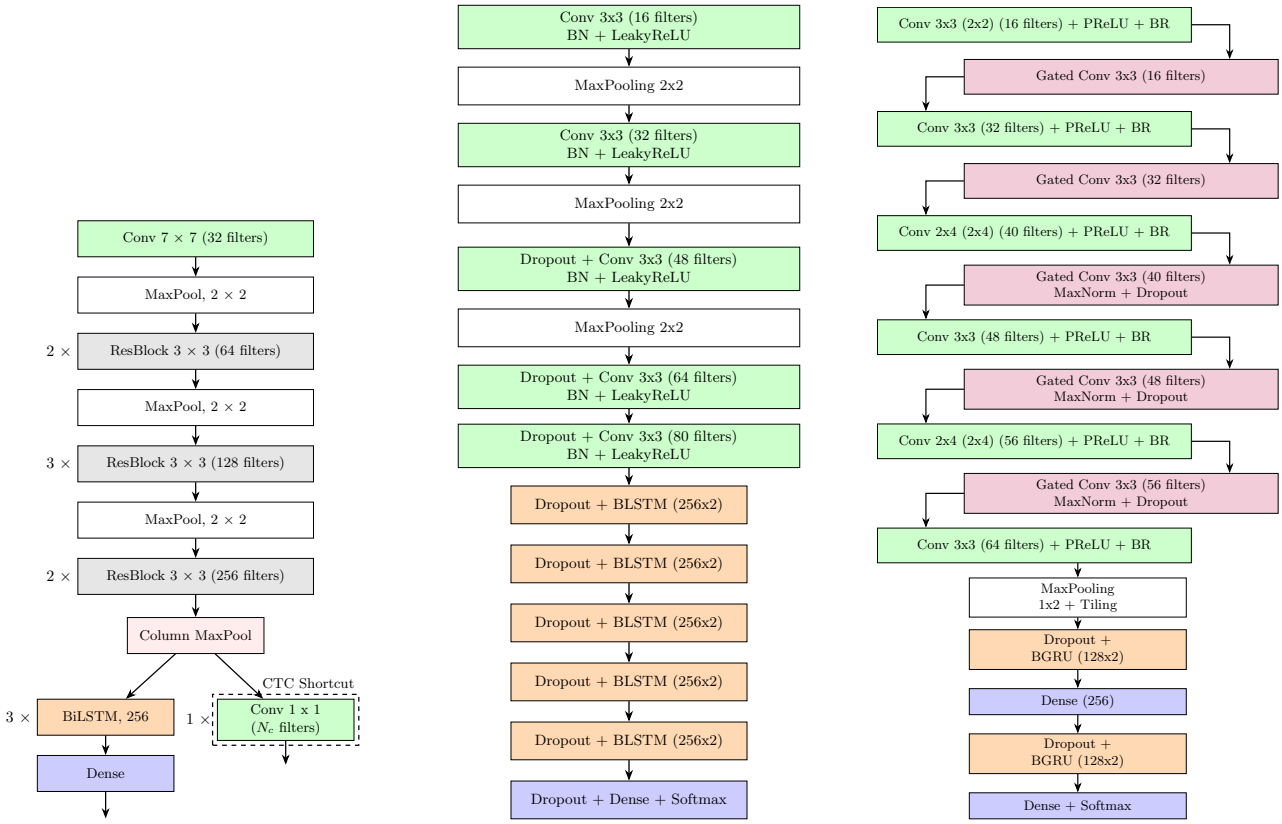
In Table 8, the detailed architectures of the PyLaia model for the ablation study on the model size are reported. Large corresponds to the default architecture and is also the default provided in the PyLaia Framework. The main difference between the mid and large model is the reduced size of the CNN backbone (3 vs. 4), while the small and mid model differ mainly in terms of the reduced size of the RNN.

8.1. Subset size for finetuning

Figure 10 illustrates the number of training samples retained for finetuning at various confidence thresholds. As the confidence threshold increases from 0 to 90, the number of samples decreases across all three models: HTRFlor, PyLaia, and Retsinas. HTRFlor consistently maintains a higher number of samples compared to the other models at most thresholds, while Retsinas shows the steepest decline. This shows how stricter confidence criteria reduce the training dataset size, which impacts the model finetuning by reducing the amount of faulty data (while also having less data).

Parameter	**Small**	**Mid**	**Large**
CNN			
Features	[16, 24, 48]	[16, 24, 36]	[16, 16, 32, 32]
Kernel Size	[3, 3, 3]	[3, 3, 3]	[3, 3, 3, 3]
Stride	[1, 1, 1]	[1, 1, 1]	[1, 1, 1, 1]
Dilation	[1, 1, 1]	[1, 1, 1]	[1, 1, 1, 1]
Pool Size	[2, 2, 2]	[2, 2, 2]	[2, 2, 2, 0]
Dropout	0	0	0
Activation	LeakyReLU	LeakyReLU	LeakyReLU
RNN			
Layers	2	3	3
Units	128	256	256
Type	LSTM	LSTM	LSTM
Dropout	0.5	0.5	0.5
Linear			
Dropout	0.5	0.5	0.5

Table 8. Comparison of Small (1.3M), Mid (4.9M), and Large (6.4M) PyLaia model configurations. (T = True, F = False)



(a) Retsinas et al. [17] (b) PyLaia [16] (c) HTRFlor [3]

Figure 9. Comparison of the three architectural designs used for text recognition.

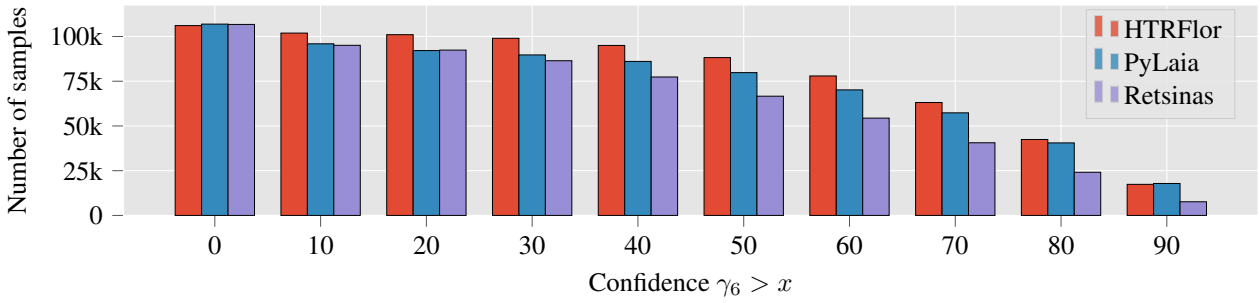


Figure 10. Number of training samples used for finetuning for different thresholds.