ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

Handwritten Text Recognition: A Comprehensive Survey of Evolution and Architectures

Subodh Kant[1]; Dr. Gaurav Harit[2]

[1]Department of Computer Science and Engineering Indian Institute of Technology Jodhpur, Rajasthan, India
[2]Co-Author; Department of Computer Science and Engineering Indian Institute of Technology Jodhpur, Rajasthan, India

[1]ORCID: 0009-0008-7099-2304
[2]ORCID: 0000-0001-7943-0123

Publication Date: 2026/07/10

Abstract: This survey presents a comprehensive synthesis of contemporary research in handwritten text recognition (HTR), encompassing bibliometric analysis, systematic methodological reviews, and novel architectural innovations spanning line-level, document-level, and multi-lingual recognition systems. Drawing from 25 original research works, survey papers, and empirical studies, this work categories existing literature into ten thematic clusters: survey and review studies, meta-learning and adaptation methods, self-supervised learning approaches, vision-language models, transformer-based architectures, word and keyword spotting methods, document-level recognition systems, low-resource and Indic script recognition, foundational deep learning architectures, and out-of-distribution generalization studies. The survey reveals that while significant progress has been achieved through deep learning and transfer learning techniques, critical challenges persist in handling domain shifts, low-resource languages, and complex document layouts. Furthermore, emerging paradigms including self-supervised vision transformers, meta-learning frameworks, and foundation models demonstrate substantial promise for enabling more adaptive, efficient, and generalisable HTR systems. This paper synthesises these developments, identifies cross-cutting methodological themes, analyses comparative performance trends, and delineates key research gaps that warrant future investigation.

Keywords: *Handwritten Text Recognition, Deep Learning, Meta-Learning, Vision Transformers, Self-Supervised Learning, Domain Adaptation, Foundation Models, Indic Scripts.*

How to Cite: Subodh Kant; Dr. Gaurav Harit (2026) Handwritten Text Recognition: A Comprehensive Survey of Evolution and Architectures. *International Journal of Innovative Science and Research Technology*, 11(7), 28-33. https://doi.org/10.38124/ijisrt/26jul106

I. INTRODUCTION

➢ *Research Context and Motivation*

Handwritten text recognition (HTR) represents a fundamental challenge in computer vision and document analysis, with profound implications for cultural heritage preservation, administrative automation, and accessibility technologies. The task encompasses the conversion of handwritten textual content—characterized by inherent variability in stroke characteristics, slant, spacing, and writing style—into machine-encoded format suitable for digital storage, retrieval, and analysis. Despite more than four decades of intensive research and the emergence of deep learning methodologies, achieving robust and generalisable HTR performance across diverse handwriting styles, scripts, and document types remains an open research problem.

The motivation for comprehensive examination of the HTR landscape is multi-faceted. First, the field has undergone substantial transformation through the adoption of deep neural networks, sequence-to-sequence models, and transformer architectures, warranting systematic analysis of these methodological shifts. Second, the emergence of meta-learning frameworks, self-supervised approaches, and foundation models represents a paradigm shift from traditional supervised learning assumptions. Third, the increasing recognition of out-of-distribution challenges and low-resource language applications has expanded the scope of HTR research beyond benchmark-driven optimisation toward more practically generalisable systems. Finally, the proliferation of handwritten digital ink technologies in tablets and styluses necessitates novel approaches to real-time, on-device recognition.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

➢ *Scope and Organization*

This survey synthesises findings from 25 primary research contributions spanning multiple research directions within HTR and adjacent domains. The selected works represent a balanced portfolio of theoretical innovations, empirical investigations, methodological surveys, and application-specific systems. The organizational structure reflects ten identified thematic clusters that capture the principal research directions within contemporary HTR literature:

- Survey and Review Studies — Meta-analyses of HTR field evolution and bibliometric trends.
- Meta-Learning and Adaptation Methods — Writer-adaptive and test-time adaptation frameworks.
- Self-Supervised and Unsupervised Learning — Representation learning without labeled annotations.
- Vision-Language Models and Foundation Models — Multimodal approaches for text recognition.
- Transformer-Based HTR Architectures — Vision transformer applications in recognition.
- Word and Keyword Spotting — Sub-document-level retrieval and indexing.2
- Document-Level and Page-Level Recognition — End-to-end systems for complex layouts.
- Low-Resource and Indic Script Recognition — Multilingual and endangered language HTR.
- General Deep Learning Architectures — Foundational models enabling HTR advances.
- Generalization and Robustness Studies — Out-of-distribution performance analysis.

Following this introduction, Section 2 establishes the contextual background within existing literature. Section 3 presents methodological reviews of the surveyed works, organized thematically. Section 4 provides comparative analysis across approaches, identifying performance trends and technical trade-offs. Section 5 synthesizes identified research gaps and challenges, while Section 6 proposes future research directions. Finally, Section 7 concludes with key insights and recommendations for advancing the field.

II. BACKGROUND AND LITERATURE CONTEXT

➢ *Historical Evolution of HTR Approaches*

The evolution of HTR reflects broader trends in machine learning and computer vision, progressing through distinct technological phases. Early approaches, dominated by heuristic and handcrafted feature representations, relied on domain expertise to engineer features such as chain codes, gradient-based descriptors, and morphological characteristics. Hidden Markov Models (HMMs) emerged as the dominant paradigm for several decades, leveraging the model's capacity to capture sequential dependencies in handwritten traces and character sequences.

The transition to neural network-based methods accelerated following the 2012 AlexNet breakthrough in image classification. Convolutional Neural Networks (CNNs) proved particularly effective for feature extraction from handwritten document images, providing learnable representations that obviated manual feature engineering. The integration of recurrent architectures—specifically Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs)—enabled HTR systems to capture both visual features and sequential dependencies crucial for accurate text transcription.

The most recent paradigm shift toward sequence-to-sequence models with attention mechanisms introduced encoder-decoder architectures capable of handling variable-length outputs and complex spatial arrangements. Concurrently, transformer architectures, initially developed for natural language processing, have begun to penetrate the HTR landscape, offering both computational efficiency and enhanced feature expressivity through multi-head self-attention mechanisms.

➢ *Challenges Specific to Handwritten Text Recognition*

Several factors distinguish HTR from related recognition tasks such as printed text OCR or scene text recognition. Handwriting exhibits high intra-personal and inter-personal variability, influenced by factors including individual writing habits, emotional states, educational background, age, and neuromotor characteristics. This inherent variability creates a continuous rather than discrete recognition problem, where models must accommodate spectrum-like variations rather than discrete categorical distinctions.

The data scarcity problem represents another critical challenge. High-quality handwritten datasets require labour-intensive manual annotation, and publicly available datasets remain orders of magnitude smaller than their counterparts in printed text recognition or general computer vision. This asymmetry in data availability has historically disadvantaged deep learning approaches relative to traditional methods, though recent advances in transfer learning and self-supervised techniques have substantially mitigated this limitation.

➢ *Evaluation Metrics and Standardization*

Standardized evaluation metrics are essential for facilitating meaningful comparisons across research contributions. The HTR field has converged on several metrics borrowed from automatic speech recognition and adopted to text recognition contexts.

Character Error Rate (CER) represents the most granular evaluation metric, measuring the minimum edit distance (Levenshtein distance) between predicted and ground truth character sequences, normalized by reference sequence length:

$$CER=\frac{S+D+I}{N}\times 100\% \quad (1)$$

Where S denotes substitutions, D denotes deletions, I denotes insertions, and N denotes total reference characters.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

Word Error Rate (WER) provides recognition accuracy at the word level, similarly computed over word sequences rather than character sequences, providing insight into semantic error patterns.

Sequence Error Rate (SER) indicates the proportion of entire sequences (lines, paragraphs, or documents) incorrectly recognized in their entirety, providing a stringent evaluation criterion.

Additionally, domain-specific metrics have emerged for particular applications. Word spotting research employs Mean Average Precision (mAP) to evaluate retrieval accuracy, while document layout analysis incorporates pixel-level metrics from semantic segmentation literature.

III. METHODOLOGICAL REVIEW OF CONTEMPORARY RESEARCH

➢ *Survey and Review Studies*

The field of HTR has accumulated sufficient empirical results and methodological diversity to warrant systematic reviews and meta-analytical examinations. Three primary survey contributions structure existing knowledge within thematic frameworks.

The first comprehensive bibliometric and systematic review of handwritten document recognition (HDR) techniques encompasses not solely textual content but also auxiliary document elements including numerals, mathematical expressions, diagrams, and tables [1]. The bibliometric analysis reveals quantitative trends in publication volume, research focus areas, and citation patterns across the Scopus database. A principal finding is the identification of substantial opportunities for handling complex, real-world documents with variable styles and quality—opportunities that remain largely unaddressed despite advances in isolated recognition tasks.

A comprehensive examination of HTR evolution spanning from early heuristic-based approaches to modern deep learning techniques establishes a conceptual framework distinguishing between two primary recognition granularities [2]. The survey emphasizes the historical borrowing of techniques from automatic speech recognition (ASR), noting fundamental similarities in problem formulation. A critical contribution is comprehensive dataset enumeration, cataloging public benchmarks, evaluation protocols, and performance trends across temporal progression, documenting a 37% reduction in average CER over recent years through architectural and methodological innovations.

➢ *Meta-Learning and Adaptation Methods*

Meta-learning frameworks represent a fundamental paradigm shift in HTR, enabling systems to rapidly adapt to new writers or document styles with minimal additional examples

MetaHTR [3] introduces model-agnostic meta-learning (MAML) for writer-adaptive handwritten text recognition. The framework discovers that different characters from the same writer exhibit varying levels of style discrepancy, motivating character-specific weighted loss functions that focus adaptation effort on the most challenging characters for each writer. Critically, MetaHTR employs learnable layer-wise learning rates rather than fixed learning rates, enabling differential adaptation across network layers. Empirical evaluation demonstrates 2.1–2.4% accuracy improvements over standard MAML across three baseline architectures.

DocTTT [4] extends meta-learning principles to document-level recognition through a test-time training framework combining primary and auxiliary task branches. The primary branch performs the core recognition task, generating sequences of alphabetic characters and layout markers interpretable as flattened XML representations of document structure. A bi-level MAML optimisation procedure enables rapid test-time adaptation to specific document characteristics.

MetaWriter [5] addresses practical limitations by eliminating the requirement for labeled adaptation examples. Instead, the framework leverages self-supervised image reconstruction using masked auto encoders (MAE) combined with meta-learned prompt tuning. Prompts are injected as learnable image padding, achieving extreme parameter efficiency—0.08M trainable parameters during adaptation versus 1.7M for MetaHTR. Evaluation reveals a 20.4% relative CER improvement over TrOCR, with superior performance to MetaHTR despite using unlabeled data.

➢ *Self-Supervised and Unsupervised Learning*

The emergence of self-supervised learning approaches represents a significant methodological advance, particularly important given persistent data scarcity in HTR.

ST-KeyS [6] addresses data scarcity in historical document analysis through masked auto encoder pre-training combined with vision transformers. The framework employs unsupervised pre-training on unlabeled historical document collections to learn word representations, followed by supervised fine-tuning for word spotting. Evaluation across multiple datasets demonstrates competitive or superior performance compared to existing keyword spotting methods.

Self-Training for Synthetic-to-Real Adaptation [7] employs self-training (pseudo-labelling) to bridge the domain gap between synthetic handwritten data and real-world handwriting. The approach generates synthetic training data using 398 handwriting-like fonts with randomized stroke variations, subsequently training models and applying confidence-based selection to pseudo-label real unlabeled data.

DINO [8] introduces self-distillation with no labels, revealing emergent properties in self-supervised vision transformers. DINO-trained ViT features explicitly encode semantic segmentation and object boundary information not present in supervised ViTs or CNNs. The framework achieves 78.3% ImageNet top-1 accuracy through k-nearest neighbour classification without fine-tuning.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

DINOv2 [9] scales self-supervised vision transformers through curated data collection and optimized training procedures. Training a 1-billion parameter ViT model produces general-purpose features outperforming specialised supervised approaches on most benchmarks without task-specific fine-tuning.

➢ *Vision-Language Models and Foundation Models*

Emerging foundation models capable of processing multi-modal inputs represent a distinct methodological direction with particular promise for low-resource recognition scenarios.

Parameter-efficient fine-tuning of open-source vision-language models enables effective OCR system development for endangered languages [10]. Using Manchu as a case study, the authors employ synthetic handwritten word images combined with structured prompts guiding dual-script output. Key findings indicate that open-source models maintain robust real-world performance following synthetic data fine-tuning, with only a 5.2 percentage point accuracy decline, whereas traditional CRNN baselines suffer a 27.3 percentage point degradation.

InkFM [11] unifies multiple handwriting understanding tasks within a single foundational model based on the PaLI-Gemma 3B architecture. The model simultaneously performs text segmentation at hierarchical levels, text recognition across 28 scripts, and mathematical expression recognition, plus auxiliary classification tasks. Evaluation demonstrates state-of-the-art results on 5 of 6 datasets.

CLIP [12] establishes the feasibility of vision-language foundation models through contrastive learning on 400 million internet image-text pairs. The model learns to match images with corresponding captions, enabling zero-shot classification by computing similarity between image and text embeddings.

➢ *Transformer-Based HTR Architectures*

The adoption of transformer architectures in HTR represents a fundamental architectural transition from CNN+LSTM paradigms.

HTR-VT [13] demonstrates that vision transformers with minimal modifications achieve state-of-the-art HTR performance without requiring external pre-training data. The framework processes handwritten text line images through CNN feature extraction, subsequently applying transformer encoder processing and CTC loss optimization. Critical innovations include span masking enabling broader attention receptive fields.

GatedLexiconNet [14] addresses limitations of traditional segmentation-based HTR by introducing end-to-end paragraph recognition without requiring explicit text line segmentation. The framework integrates gated convolutional layers within a Vertical Attention Network architecture, achieving 54.7–70.4% relative CER improvements over the baseline VAN.

HTR-JAND [15] combines advanced CNN feature extraction with multi-head self-attention and knowledge distillation. The framework employs a multi-component loss combining CTC loss, cross-entropy loss, knowledge distillation loss, and auxiliary losses. Evaluation demonstrates a 90% CER reduction on the Bentham dataset while maintaining computational efficiency.

➢ *Document-Level and Low-Resource Recognition*

Meta-DAN [16] improves the Document Attention Network for page-level recognition by proposing novel multi-token and windowed-decoding strategies. MT-DAN predicts multiple tokens per decoding iteration with dynamic policy adjustment, while W-DAN processes multiple queries through a sliding window approach.

PLATTER [17] introduces a comprehensive framework specifically designed for Indic language handwritten OCR. The framework employs DBNet for language-agnostic text detection combined with language-specific recognition models for ten Indic languages. Evaluation of six recognition architectures identifies optimal configurations.

Research on Devanagari character recognition achieves 96.58% accuracy through VGG16 transfer learning [18], while CNN-based Devanagari numeral recognition achieves 98.07% accuracy [19].

➢ *Out-of-Distribution Generalization*

A watershed contribution is the first large-scale systematic investigation of out-of-distribution (OOD) generalization [20]. The evaluation of 8 state-of-the-art HTR. models across 7 datasets reveals catastrophic generalization failures in OOD scenarios.

Standard CER performances in-distribution average 3.2%, yet increase to 37.6% in OOD settings—an 11.75× error increase. Critical findings include: (1) textual divergence constitutes the primary generalization barrier; (2) synthetic data substantially helps but incompletely solves OOD challenges; (3) model size demonstrates no consistent relationship to OOD robustness; and (4) OOD performance is predictable using domain-agnostic metrics.

IV. COMPARATIVE ANALYSIS

➢ *Technical Trade-offs Across Methodologies*

Contemporary HTR research exhibits diverse technical approaches reflecting different design philosophies. Key trade-offs include the following.

- *Architectural Complexity vs. Data Efficiency.* Traditional CNN+LSTM approaches prioritize simplicity and proven small-dataset performance. Transformer-based approaches offer superior data efficiency through parallelization. Foundation models enable competitive performance with limited target-domain labeled data through transfer learning.
- *Computational Cost vs. Recognition Accuracy*. Knowledge distillation approaches achieve a 48% complexity reduction while maintaining competitive

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

accuracy. Meta-learning frameworks achieve extreme parameter efficiency (0.08M tunable parameters) during adaptation. Foundation models achieve superior accuracy but imposepre-training computational costs.

- *Labeled Data Requirements vs. Generalization*. Self-supervised approaches reduce labelled data requirements but require large unlabelled datasets. Meta-learning frameworks reduce few-shot adaptation requirements. Combined approaches achieve both reduced labeled data requirements and superior few-shot generalization.

➢ *Performance Trends and Benchmarking*

On the IAM English handwriting database, state-of-the-art recognition achieves character error rates spanning 2.1–3.4% depending on architectural approach. The French RIMES dataset demonstrates similar convergence patterns, with state-of-the-art approaches achieving 0.9–2.2% CER. Recognition of degraded historical documents presents substantially higher error rates, with a 90% error reduction demonstrated through knowledge distillation approaches.

As revealed by the systematic OOD generalization study [20], recognition accuracy degrades catastrophically when models encounter domain shifts. In-distribution performance averaging 3.2% CER increases to 37.6% in OOD scenarios—an 11.75× error multiplier.

V. RESEARCH CHALLENGES

➢ *Technical Trade-offs Across Methodologies*

The catastrophic performance degradation in OOD scenarios represents the most fundamental unresolved challenge. Current approaches achieve 3.2% in-distribution CER yet degrade to 37.6% in OOD settings. The root causes remain incompletely understood, requiring principled theoretical frameworks and training procedures explicitly optimizing for OOD robustness.

➢ *Foundation Models as HTR Base Systems*

Scaling self-supervised learning to massive model sizes suggests potential for general purpose visual foundation models to serve as HTR base systems. Future research should explore systematically leveraging foundation models through efficient fine tuning, prompt tuning, or in-context learning approaches.

➢ *Low-Resource Language Coverage*

Fundamental asymmetries persist in resource availability across writing systems. While emerging frameworks begin addressing this asymmetry, coverage remains limited. Transfer learning between related scripts could substantially improve low-resource script recognition.

➢ *Human–AI Collaboration Frameworks*

Research should emphasize human–AI collaboration where systems provide high confidence predictions for verification while flagging uncertain regions for human transcription, maximization productivity while maintaining quality guarantees.

VI. FUTURE RESEARCH DIRECTIONS

Based on the identified challenges, several high-priority research directions are recommended. First, principled OOD training objectives should be developed that incorporate domain-agnostic evaluation metrics as optimization targets. Second, parameter-efficient adaptation methods—such as prompt tuning and low-rank adaptation—should be further explored to enable scalable deployment of large foundation models under constrained computational budgets. Third, the development of synthetic data pipelines specifically targeting under-represented scripts and historical document styles will be essential for low-resource language coverage. Finally, the integration of uncertainty quantification into HTR pipelines is a prerequisite for practical human–AI collaborative transcription workflows.

VII. CONCLUSION

The contemporary landscape of handwritten text recognition reveals a field in transition, characterized by methodological innovation and growing recognition of practical deployment limitations. This survey synthesized 25 primary research contributions across ten thematic clusters.

Key insights include: (1) methodological evolution from heuristic approaches toward transformer-based and foundation model approaches; (2) emergence of meta-learning frameworks as a standard for handling writer-specific variations; (3) foundation models and self-supervised learning representing fundamental shifts in the field; (4) catastrophic OOD generalization failures requiring re-conceptualization of evaluation practices; and (5) growing recognition of biases toward Latin scripts necessitating multilingual framework development.

The field stands at a crucial juncture where addressing authentic deployment challenges—particularly OOD robustness and low-resource language coverage—requires fundamental methodological innovation. The emergence of meta-learning frameworks, self-supervised approaches, and foundation models provides promising directions, yet their full potential remains substantially unexploited.

ACKNOWLEDGMENTS

The author would like to thank the faculty and research community at IIT Jodhpur for their guidance and the reviewers for their valuable feedback.

REFERENCES

[1]. V. Agrawal, J. Jagtap, and M. V. V. Prasad Kantipudi, “Exploration of advancements in handwritten document recognition techniques,” Intelligent Systems with Applications, vol. 22, Art. no. 200346, 2024.

[2]. C. Garrido-Muñoz, A. Rios-Vila, and J. Calvo-Zaragoza, “Handwritten text recognition: A survey,” arXiv preprint arXiv:2502.08417, 2025.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul106

[3]. A. K. Bhunia, S. Khan, A. Fischer, F. S. Khan, and L. Shao, "MetaHTR: Meta-learning for writer-adaptive handwritten text recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1581–1590.

[4]. W. Gu, L. Gu, Z. Wang, C. Y. Suen, and Y. Wang, "DocTTT: Test-Time Training for Handwritten Document Recognition Using Meta-Auxiliary Learning," arXiv preprint arXiv:2501.12898, 2025.

[5]. W. Gu, L. Gu, C. Y. Suen, and Y. Wang, "MetaWriter: Personalized handwritten text recognition using meta-learned prompt tuning," arXiv preprint arXiv:2505.20513, 2025.

[6]. S. K. Jemni, M. E. Souibgui, Y. Kessentini, and A. Fornés, "ST-KeyS: Self-supervised transformer for keyword spotting in historical documents," in Proc. Digital Humanities Conference, 2023.

[7]. F. Wolf and G. A. Fink, "Self-training of handwritten word recognition for synthetic-to-real adaptation," in Proc. International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1–8.

[8]. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, et al., "Emerging properties in self-supervised vision transformers," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9650–9660.

[9]. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, et al., "DINOv2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.

[10]. Y. H. M. Chung and D. Choi, "Finetuning vision-language models as OCR systems for low-resource languages: A case study of Manchu," arXiv preprint arXiv:2507.06761, 2025.

[11]. A. Fadeeva, V. Coriou, D. Antognini, C. Musat, and A. Maksai, "InkFM: A foundational model for full-page online handwritten note understanding," arXiv preprint arXiv:2503.23081, 2025.

[12]. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in Proc. International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.

[13]. Y. Li, Y. Zhu, and Z. Lian, "HTR-VT: Handwritten text recognition with vision transformer," arXiv preprint arXiv:2409.08573, 2024.

[14]. L. Kumari, P. Mazumdar, and P. Bhattacharyya, "GatedLexiconNet: End-to-end handwritten paragraph recognition," arXiv preprint arXiv:2404.14062, 2024.

[15]. M. Hamdan, L. S. Saoud, and N. Houmani, "HTR-JAND: Joint attention network and knowledge distillation for handwritten text recognition," arXiv preprint arXiv:2412.18524, 2024.

[16]. D. Coquenet, "Meta-DAN: Towards an efficient prediction strategy for page-level handwritten text recognition," arXiv preprint arXiv:2504.03349, 2025.

[17]. B. V. Kasuba, M. K. Chinnakotla, and C. V. Jawahar, "PLATTER: Page-level recognition for Indic languages," arXiv preprint arXiv:2502.06172, 2025.

[18]. L. Sharma, A. Bhatt, and N. Garg, "Advancements in handwritten Devanagari character recognition using deep learning," Discover Applied Sciences, vol. 6, 2024.

[19]. K. V. C. Madhav, A. Veeramaneni, P. A. Sriya, K. Sumanth, J. C. Jyotsna, T. Singh, and P. Duraisamy, "Handwritten Devanagari numeral recognition using deep learning," in Proc. 2024 3rd International Conference for Innovation in Technology (INOCON), Karnataka, India, Mar. 2024, pp. 1–6, doi: 10.1109/INOCON60754.2024.10511947.

[20]. C. Garrido-Muñoz and J. Calvo-Zaragoza, "On the generalization of handwritten text recognition models," arXiv preprint arXiv:2411.17332, 2024.