MPHD: Multi-purpose Persian Handwriting Dataset for Text Recognition, Line Segmentation, Writer Identification, and Handwritten Analysis

Mahdi Jampour, Shervin Farridnejad, Amin KarimiSardar, Kourosh Champour, Azizeh Aghaee Meybodi, Faezeh Hematzadeh, and Ludwig Paul

Abstract—Despite remarkable advances in handwriting text recognition (HTR), progress has remained overwhelmingly concentrated on Latin scripts. Cursive, right-to-left writing systems such as Persian, whose 32 letters exhibit complex positional variants and are frequently distinguished only by subtle diacritical marks, continue to lack large-scale, publicly available resources capable of supporting modern deep learning pipelines. Existing Persian datasets are typically restricted to isolated characters or words, omit line-level transcriptions, or provide no demographic metadata, thereby impeding both technical progress and broader studies in writer analysis. We introduce MPHD, a multi-purpose offline Persian handwriting dataset collected from 500 native writers. It contains 75,430 words and 333,545 characters across 5,021 carefully annotated text lines, released at three complementary levels of granularity (full-form pages, cropped text regions, and isolated characters) together with rich demographic attributes (age, gender, education, handedness, and spectacles usage). Extensive experiments on four core tasks, HTR, line segmentation, writer identification, and character/digit recognition, reveal a stark performance hierarchy: character and digit recognition exceed 96% accuracy, writer identification reaches 95.19% Top-1 accuracy, and line segmentation remains substantially harder (mIoU below 0.45), while HTR itself proves the most demanding task. State-of-the-art Vision Transformer models attain only 2.17% CER on fixed text yet degrade to 21.34% CER on variable text, a gap driven by heavy reliance on lexical memorization rather than robust visual understanding that exposes fundamental limitations in fine-grained diacritic discrimination and generalization to unseen lexical content. MPHD thus provides a rigorous and previously unavailable benchmark for advancing cursive script understanding.

Index Terms—Multi-Purpose, Multi-Granularity, Dual-Text, Handwritten Text Recognition, Line Segmentation, Writer Identification, Character Recognition, Handwritten Analysis, HTR, ViT, CRNN, Dataset, Benchmark.

1 Introduction

Handwriting Text Recognition (HTR) has long stood as a cornerstone challenge in pattern recognition and computer vision, with profound implications spanning historical document preservation, automated data entry, biometric authentication, and accessibility systems. While the past decade has witnessed remarkable advances, driven by the proliferation of deep learning architectures ranging from CNNs and RNNs to Transformers, as comprehensively surveyed in [1], and further exemplified by domain-specific benchmarks such as BNU-Exam-HTR [2] designed to tackle sequential and visual artifacts in examination scripts, these successes have been overwhelmingly concentrated on Latin scripts. Benchmark datasets such as BullingerDB [3], IAM [4] and RIMES [5] have provided the empirical foundation for these breakthroughs. The same cannot be said for cursive, right-to-left scripts with rich morphological complexity, where progress remains disproportionately slow. The Persian language, whose literary and scientific corpus represents one of the world's most enduring cultural heritages [6], is rendered in a cursive script that amplifies all the difficulties mentioned above. Its 32 letters assume context-dependent forms (i.e., initial, medial, final, or isolated) as illustrated in Table 1, and several letters share identical base shapes, distinguished solely by the number and placement of diacritical dots (e.g., ب، پ، ت، ث, ج، چ، ح، خ, ر، ز، ژ, etc.), making accurate recognition highly sensitive to image resolution, noise, and writing style. Table 2 further illustrates this challenge by presenting pairs of words that differ only by the presence or position of a single diacritical dot, demonstrating how a minor writing variation can lead to entirely different meanings. This intrinsic ambiguity, compounded by the natural variability of human handwriting such as slant, stroke thickness, and spacing, renders Persian handwriting recognition a uniquely difficult problem that cannot be adequately addressed by models trained on Latin or other script datasets without considerable adaptation.

Despite the growing demand for robust Persian handwriting recognition systems, fueled by cultural heritage digitization, administrative automation, and the proliferation of Persian digital content, the research community has long suffered from a scarcity of comprehensive, publicly available benchmark datasets. Existing Persian resources, while valuable in their specific niches, suffer from fundamental limitations that render them inadequate for modern HTR pipelines. The Hoda dataset [7], though widely cited, is confined to isolated digits and thus cannot support continuous text recognition. The IFN/Farsi database [8] is inspired by the Arabic IFN/ENIT corpus [9]; both are word-level datasets comprising city names. Moreover, the IFN/ENIT corpus lacks

Corresponding author: Dr. Mahdi Jampour, University of Hamburg, Hamburg, Germany (e-mail: mahdi.jampour@uni-hamburg.de).
The MPHD is available at DOI: 10.17632/2fcw6cd9gd.1 and page: https://github.com/jampour/MPHD-Persian-Handwriting-Dataset

TABLE 1: The 32 letters of the Persian alphabet and their contextual forms according to their position and joining behavior within a word. Here, Iso, Ini, Med, and Fin denote isolated, initial, medial, and final forms, respectively.

No.	Iso	Ini	Med	Fin	No.	Iso	Ini	Med	Fin	No.	Iso	Ini	Med	Fin	No.	Iso	Ini	Med	Fin
1	آ	آ	آ	آ	9	خ	خـ	ـخـ	ـخ	17	ص	صـ	ـصـ	ـص	25	ک	کـ	ـکـ	ـک
2	ب	بـ	ـبـ	ـب	10	د	د	ـد	ـد	18	ض	ضـ	ـضـ	ـض	26	گ	گـ	ـگـ	ـگ
3	پ	پـ	ـپـ	ـپ	11	ذ	ذ	ـذ	ـذ	19	ط	طـ	ـطـ	ـط	27	ل	لـ	ـلـ	ـل
4	ت	تـ	ـتـ	ـت	12	ر	ر	ـر	ـر	20	ظ	ظـ	ـظـ	ـظ	28	م	مـ	ـمـ	ـم
5	ث	ثـ	ـثـ	ـث	13	ز	ز	ـز	ـز	21	ع	عـ	ـعـ	ـع	29	ن	نـ	ـنـ	ـن
6	ج	جـ	ـجـ	ـج	14	ژ	ژ	ـژ	ـژ	22	غ	غـ	ـغـ	ـغ	30	و	و	ـو	ـو
7	چ	چـ	ـچـ	ـچ	15	س	سـ	ـسـ	ـس	23	ف	فـ	ـفـ	ـف	31	ه	هـ	ـهـ	ـه
8	ح	حـ	ـحـ	ـح	16	ش	شـ	ـشـ	ـش	24	ق	قـ	ـقـ	ـق	32	ی	یـ	ـیـ	ـی

several Persian-specific letters (e.g., پ، چ، ژ، گ), and like other word-level resources, neither dataset is optimally suited for modern line-level HTR pipelines. The Sadri dataset [10], representing a significant effort with 500 writers and seven-page forms, remains one of the most comprehensive Persian resources; however, it does not provide line-level transcriptions, limiting its direct applicability to modern HTR systems that rely on line-wise training and evaluation. More recently, the Khayyam dataset [11] has been introduced as a large unconstrained collection of 44,000 words. However, like most existing Persian resources, it is primarily word-level and does not provide line-level annotations, limiting its direct applicability to contemporary HTR systems. Other datasets, such as FHT dataset [12], while valuable and line-level annotated, are substantially smaller, lack isolated character samples and structured demographic metadata, and are stored in a legacy format, constraints that preclude their use as a comprehensive benchmark for modern multi-task HTR evaluation. Consequently, a clear and persistent research gap exists: no publicly available Persian handwriting dataset simultaneously provides (i) line-level annotated continuous text, (ii) both constant and variable text types, (iii) isolated character samples for fine-grained recognition, (iv) thematic text diversity, and (v) rich demographic metadata, all within a unified, well-annotated collection. This gap is particularly relevant for paleographic and stylistic analysis. Our recent work on end-to-end modeling of historical handwriting styles [13] highlighted the value of fine-grained structural annotations for style analysis, a capability that current Persian datasets largely lack and that MPHD aims to provide. Similarly, our previous deep learning framework for text-independent writer identification [14] demonstrated promising results on non-Persian scripts, yet lacked a sufficiently rich Persian benchmark for validation at both line and character levels. Our prior experience in designing historical manuscript collections [15] and large-scale visual datasets [16] has directly informed the data collection and annotation protocols of MPHD, ensuring both scalability and demographic representativeness.

To address this critical gap, we introduce the Multi-purpose Persian Handwriting Dataset (MPHD), a comprehensive, publicly available, and multi-purpose offline Persian handwriting dataset. MPHD comprises handwritten forms collected from 500 native Persian writers, with a balanced distribution across five education levels and an age range of 11 to 36 years. Each writer contributed three distinct components, a fixed text, a variable text and a set of 90 isolated characters organized at three levels of granularity: full-form images, cropped text regions, and individual character images. Comprehensive metadata, including writer demographics, is provided in a structured CSV file along with JSON ground truth files, enabling a wide range of downstream analyses beyond mere transcription.

TABLE 2: Examples of Persian letter groups with similar base shapes and diacritical differences, highlighting visual confusion through contrasting word pairs.

Letter Group	Words with similar base shapes
ب، پ، ت، ث	بخش (Department) – پخش (Broadcast)
ج، چ، ح، خ	چنگ (Harp) – جنگ (War) – خنگ (Dumb)
د، ذ	دلت (Your heart) – ذلت (Humiliation)
ر، ز، ژ	دزد (Thief) – درد (Pain)
س، ش	سیر (Garlic) – شیر (Milk)
ط، ظ	طاهر (Pure) – ظاهر (Apparent)
ع، غ	عیب (Fault) – غیب (Unseen)
ک، گ	کام (Mouth) – گام (Step)

MPHD offers several distinctive advantages over existing Persian handwriting datasets. First, its dual-text design, combining a fixed reference text with variable thematic texts, enables rigorous evaluation of HTR models under both controlled and realistic conditions, while also facilitating content-based retrieval tasks. Second, the inclusion of isolated characters (90 per writer) provides a dedicated resource for character-level analysis and for studying the influence of demographic factors on the shape of individual glyphs. Third, the dataset's balanced demographic composition enables systematic investigation of the relationship between handwriting and writer characteristics, including age, gender, education, handedness, and spectacles usage. This direction has received limited attention in the Persian context due to data scarcity. Fourth, all data components (full forms, cropped texts, isolated characters, transcriptions, and metadata) are complete, consistent, and publicly released with a permissive license, ensuring reproducibility and facilitating comparative research. To the best of our knowledge, and as evidenced by the comparative analysis in Table 3, MPHD is the first publicly available dual-text design Persian handwriting dataset that simultaneously supports HTR, line segmentation, writer identification, and character recognition within a unified framework, while also providing rich demographic annotations for auxiliary studies in graphology and behavioral biometrics. In total, the dataset comprises over 75,000 words and 330,000 characters (excluding spaces) across all text components,

TABLE 3: Comparison of representative handwriting datasets across different scripts. MPHD uniquely provides multi-level annotations alongside comprehensive demographic metadata, while remaining publicly available. For details, see Section 2.

Dataset	Script/Lang.	# Writers	# Samples	Text Level	Multi-Task	Metadata & Availability
KHATT [17]	Arabic	1,000	4,000 paragraphs	Paragraph/Line	Yes	Yes (gender, age, education, handedness); Public
HICMA [18]	Arabic	–	5,031 pg	Manuscript/Calligraphy	No	Yes (Style labels); Public
IFN/ENIT [9]	Arabic	411	26,459 words	Word	Yes	Yes (age, gender, etc.); Public
AHDB [19]	Arabic	100	3,000+ words	Line/Word	Yes	Limited; Public
IAM [4]	English (Latin)	657	1,539 pg, 13,353 ln	Paragraph/Line/Word	Yes	Limited (GT, writer ID); Public
RIMES [5]	French (Latin)	1,300+	12,723 pg, 12,000+ ln	Page/Line	Yes	Limited (writer ID); Public
OpenITI MAKHZAN [20]	Multi-language	–	1,500 pg	Page/Line	Yes	Yes(source, script); Public
NUST-UHWR [21]	Urdu	1,000+	10,608 ln	Line	Yes	Yes (age, gender, background); Public
Sadri [10]	Persian	500	3,500 pg	Page/Word/Char	Yes	Yes (age, education, handedness, gender); Public (upon request)
Khayyam [11]	Persian	400	44,000 words, 60,000 letters, 6,000 digits	Word/Isolated char	No	Limited metadata, Public (upon request)
Hoda [7]	Persian	–	102,352 digits	Digit	No	No metadata; Public
FHT [12]	Persian	250	1,000 pg	Page/Line	Yes	Limited; Public (.mat files)
MPHD (Ours)	**Persian**	**500**	**500 pg, 1,000 images, 2 texts, 5,021 ln, 45,000 chars**	**Full Form/Line/Char**	**Yes**	**Yes (gender, education, handedness, spectacles, age) Public**

TABLE 4: Summary statistics of the MPHD dataset, including word, character, and line counts for the fixed (Text 1) and variable (Text 2), along with their combined totals.

Metric	T1 (Fixed)	T2 (Var.)	Combined
Total Writers	500	500	500
Isolated Chars	—	—	45,000
Text Lines	2,745	2,276	5,021
Words	39,996	35,434	75,430
Chars (w/o spaces)	183,958	149,587	333,545
Avg. Words / Writer	80.0	70.9	150.9
Avg. Chars / Writer	367.9	299.2	667.1

along with 45,000 isolated character images and more than 5,000 line-level annotated text lines (see Table 4 for a detailed breakdown), making it one of the most comprehensive multi-purpose Persian handwriting resources to date.

The remainder of this paper is organized as follows. Section 2 reviews related work on existing datasets specifically for handwriting recognition in Persian with a quick review on the similar datasets for relevant languages such as, Arabic, Urdu and English-script languages, highlighting the research gaps that MPHD addresses. Section 3 provides a detailed description of the MPHD dataset, covering the data collection protocol, form design, annotation process, and data structure. It also presents comprehensive statistics and demographic analysis, offering insights into the dataset's composition and potential applications. Section 4 reports baseline experimental results for the primary tasks supported by MPHD, handwriting text recognition, writer identification, line segmentation, and character and digit recognition, using baseline and state-of-the-art approaches, including Vision Transformer (ViT)-based models. Finally, Section 5 concludes the paper with a summary of contributions, a discussion of limitations, and directions for future work.

2 Related Work

The development of robust handwriting recognition systems is fundamentally dependent on the availability of comprehensive, well-annotated benchmark datasets. Over the past three decades, numerous datasets have been introduced for various scripts, each contributing to advances in handwriting text recognition (HTR), writer identification, and document analysis. However, as highlighted in the introduction, the Persian language remains significantly underrepresented in this landscape. This section reviews the most relevant existing handwriting datasets and methodologies, with a particular focus on Persian and other representative languages, to contextualize the contributions of the proposed MPHD dataset.

2.1 Handwriting Datasets for Latin Scripts

The IAM Handwriting Database [4] stands as the most widely used benchmark for English handwriting recognition. It comprises 1,539 scanned document pages written by 657 distinct writers, providing a rich resource for both HTR and writer identification tasks. Similarly, the RIMES (REconnaissance d'Inscriptions Manuscrites et d'Écritures en-ligne) database [5] offers French handwriting samples, while the CVL database [22] provides multi-script German and English handwriting data. These datasets have been instrumental in advancing the state of the art in Latin-script handwriting recognition, enabling the development and evaluation of increasingly sophisticated deep learning architectures. The success of these resources underscores the critical need for equivalent benchmarks for non-Latin scripts, particularly those with cursive, right-to-left writing systems.

2.2 Handwriting Datasets for Arabic Script

Given the structural and visual similarities between Persian and Arabic scripts, Arabic handwriting datasets provide valuable reference points. The KHATT (KFUPM Handwritten Arabic TexT) database [17] is one of the most comprehensive Arabic handwriting resources. It consists of handwritten forms from 1,000 distinct writers from various countries, producing 2,000 similar-text paragraph images and 2,000 unique-text paragraph images. The dataset includes manually verified ground-truth text and covers diverse topics including arts, education, health, nature, and technology. Importantly, KHATT provides writer demographic information including gender, age group, handedness, and education level, making it suitable for demographic analysis studies similar to those enabled by MPHD. The database is freely available for academic and research purposes. More recently, the HICMA

(Handwriting Identification of Manuscripts and Calligraphy in Arabic) dataset [18] was introduced as the first publicly available dataset featuring real-world Arabic handwritten text from manuscripts and calligraphy. With more than 5,000 images across five different writing styles, HICMA provides image-text pairs and style labels, addressing a gap in the evaluation of Arabic OCR systems on historical and artistic texts. The OpenITI MAKHZAN dataset [20] further contributes to Arabic-script research as a large aggregation of ground-truth data drawn from Persian, Arabic, Ottoman Turkish, and Urdu print and handwritten documents. The multi-script handwritten database (MSHD) [23] also provides a valuable resource, containing 1,300 Arabic and French handwriting samples written by 100 writers. Additional Arabic-script resources include the Arabic Handwriting Database (AHDB) [19], which provides line- and word-level samples, and the multi-script handwritten database (MSHD) [23], containing 1,300 Arabic and French handwriting samples from 100 writers, further enriching the pool of publicly available resources for Arabic-script analysis.

2.3 Handwriting Datasets for Urdu Script

Urdu, which shares the Perso-Arabic script with Persian, has benefited from more recent dataset development efforts. The NUST-UHWR (National University of Sciences and Technology - Urdu Handwritten Recognition) dataset [21] is the first published large-scale, unconstrained Urdu handwriting corpus. It includes 10,608 text lines written by nearly 1,000 contributors. Similarly, the Khat-e-FAST dataset [25] and the UPTI benchmark corpus for printed Urdu text provide additional resources for Urdu OCR research. These datasets collectively encompass a wide range of script variations, capturing the diversity of writing styles and character-level complexities. The UCOM offline dataset [26] provides 600 pages of handwritten Urdu text written in the Nasta'liq style, while the CENIP-UCCP (Center for Image Processing-Urdu Corpus Construction Project) has developed an Urdu handwritten database for benchmarking purposes [27]. The existence of these resources for Urdu highlights the comparative scarcity of publicly available Persian handwriting datasets and underscores the value of the proposed MPHD dataset.

2.4 Persian Handwriting Datasets

Despite the rich history and widespread use of Persian, the number of publicly available handwriting datasets for this language remains remarkably limited. The Sadri Persian Handwritten Database [10], introduced in 2016, represents one of the most comprehensive efforts to date. It consists of forms completed by 500 native Persian writers (250 male, 250 female), randomly selected from across Iran. The database includes a vast collection of isolated digits, connected digits, dates, words, names, alphabet letters, free text, and mathematical symbols. Importantly, each sample is accompanied by writer metadata including gender, age, handedness, and education level. The dataset is available in three formats (color, grayscale, and binary) and is pre-divided into training, validation, and test sets. While the Sadri dataset is a valuable resource, its primary focus is on isolated characters and words rather than continuous text recognition, which limits its applicability for modern HTR systems that require line-level or paragraph-level training data. The recently introduced Khayyam Offline Persian Handwriting Dataset [11] (2024) represents another significant contribution. Khayyam contains 44,000 words, 60,000 letters, and 6,000 digits written by 400 native Persian writers. The dataset intentionally focuses on collecting unconstrained Persian word samples, which were relatively scarce in previously available resources. While Khayyam addresses the word-level gap in Persian handwriting resources, its availability for broad academic use requires verification. Other Persian datasets include the Hoda dataset [7], which is limited to isolated digits and does not address continuous text recognition; the FHT (Farsi Handwritten Text) dataset [12], which provides line-level annotations from 250 writers but is substantially smaller in scale, lacks isolated character samples and structured demographic metadata, and is primarily available in a legacy format that limits its usability in modern deep learning pipelines; and the IFN/Farsi [8] dataset, which is used primarily for word-level HTR evaluation. The CENPARMI-F (Farsi Dataset) [24] is another resource but is primarily focused on digit recognition. Despite these efforts, no publicly available Persian handwriting dataset simultaneously supports these four tasks within a unified, well-annotated collection.

2.5 Research Gap and Motivation

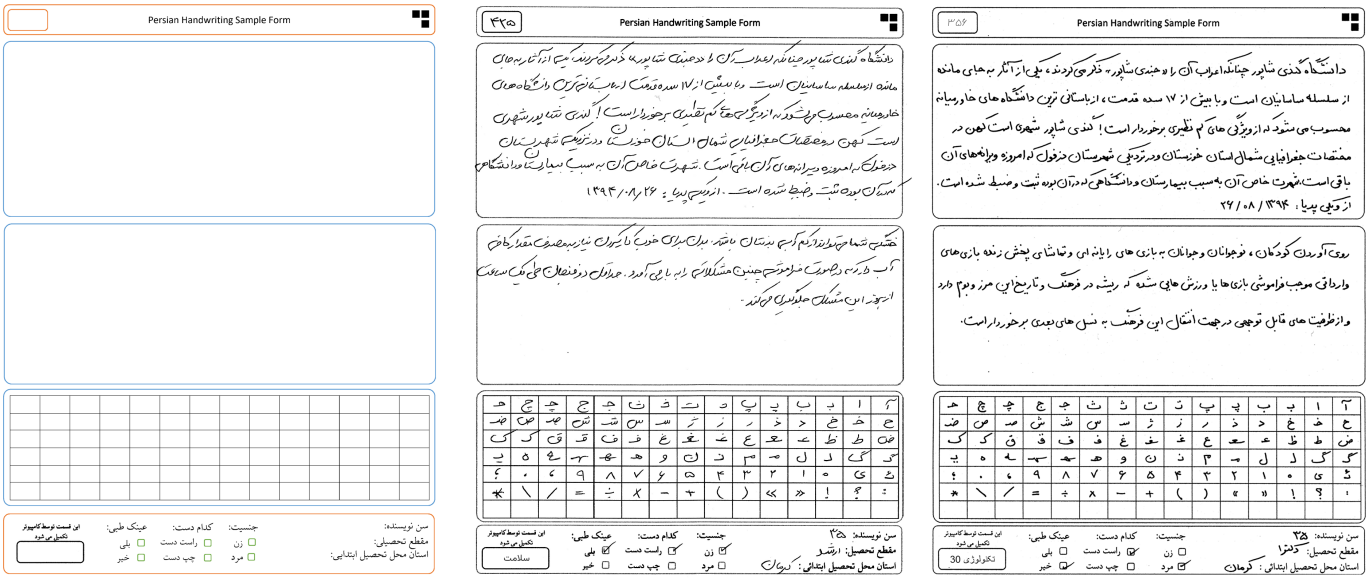
The review of existing datasets reveals several critical gaps that motivate the introduction of MPHD. First, while Arabic and Urdu have benefited from recent large-scale dataset initiatives (KHATT with 1,000 writers, NUST-UHWR with 1,000 contributors), Persian lacks a comparably large and publicly accessible dataset for continuous text recognition. Second, existing Persian datasets are either limited in scope (Hoda: digits only; Sadri: primarily isolated characters and words), distributed in legacy formats with limited modern usability (FHT), or of constrained public availability (Khayyam). Third, no existing Persian dataset provides the multi-level structure, full forms, cropped text regions, and isolated characters that enables comprehensive evaluation across multiple tasks. Fourth, while the Sadri dataset includes demographic metadata, MPHD extends this with a more structured and balanced design, including five thematic text categories. As summarized in Table 3 and discussed in Section 1, MPHD uniquely addresses these gaps through its unified multi-task design and rich demographic metadata.

3 Dataset Description

In this section, we provide a comprehensive description of the Multi-purpose Persian Handwriting Dataset (MPHD). We detail the data collection protocol, writer demographics, form design, data structure, text components, isolated characters, line-level transcriptions, and the accompanying metadata. The dataset is designed to support multiple handwriting analysis tasks.

3.1 Data Collection Protocol

The MPHD dataset was collected through a carefully designed and systematic process to ensure diversity, quality, and representativeness of Persian handwriting. The data collection was conducted over approximately a year and involved 500 native Persian writers from various regions of Iran.



(a) Blank form (b) Filled sample 1 (c) Filled sample 2

Fig. 1: The MPHD data collection form: (a) blank form; (b, c) filled samples by two writers with postgraduate education.

A standardized paper-based form was designed for data collection, comprising three main sections: a fixed text paragraph, a variable text paragraph selected from one of five thematic categories, and a dedicated section for writing 90 isolated characters. This design captures handwriting in both controlled settings, through the fixed text and isolated characters, and unconstrained settings via the variable text, thereby supporting a wide range of downstream tasks. A blank form along with two filled samples is illustrated in Fig. 1.

The forms were distributed to participants across diverse cities and educational institutions in Iran. Writers were asked to transcribe the two texts and the 90 characters in their natural handwriting style, without any constraints on writing speed, slant, or spacing. No specific writing instrument was mandated, allowing participants to use their preferred pens, which further enhances the natural variability of the collected samples. All completed forms were scanned and saved as PNG images of size 1240 × 1748 pixels to preserve quality and facilitate further processing. Each image was manually inspected to ensure clarity and completeness.

For each writer, the handwritten texts were transcribed into plain text files using UTF-8 encoding. The transcription files contain two paragraphs: the first corresponding to the fixed text (Text 1) and the second to the variable text (Text 2). These transcriptions serve as the ground truth for handwriting text recognition (HTR) tasks. In addition, as detailed in Section 3.6, all textual annotations, including paragraph-level and line-level transcriptions, are consolidated into structured JSON files to streamline data access and ensure compatibility with modern HTR pipelines.

All participants were informed about the purpose of the data collection and provided their informed consent. The dataset is fully anonymized: no personally identifiable information, such as names or national IDs, is included. Each writer is identified solely by a unique three-digit ID, ensuring privacy while enabling longitudinal or demographic analysis.

3.2 Writer Demographics

A key strength of the MPHD dataset lies in its rich demographic metadata, which enables a wide range of studies beyond pure handwriting recognition, including line segmentation, writer identification, forensic analysis, and the estimation of age, gender, and education from handwriting. The demographic composition of the 500 writers is summarized in Table 5 and detailed in the following:

The dataset exhibits a nearly balanced gender distribution, comprising 251 male (50.2%) and 249 female (49.8%) writers. Writers range from 11 to 36 years, mean age 18.3, median 17, and std 5.1 years. The majority of writers (41.6%) are in the 16–20 age bracket, reflecting a predominantly young and educationally diverse population. This distribution is valuable for studying handwriting variations across developmental stages, from adolescence to early adulthood.

Education levels are evenly distributed across five categories, each containing approximately 100 writers (roughly 20% per level): Primary, Secondary, and High School, Bachelor, and Postgraduate. This balanced design enables researchers to investigate the correlation between educational background and handwriting style, including aspects such as letter formation, spacing, and overall legibility. In terms of handedness, the dataset reflects global population statistics [28], with 455 right-handed (91%) and 45 left-handed writers (9%). Additionally, 137 writers (27.4%) report using spectacles, while 363 (72.6%) do not, providing a basis for analyzing whether visual correction influences handwriting characteristics such as stroke thickness, slant, or spacing.

In addition to the structured metadata stored in the JSON files, all demographic information is systematically encoded in the folder naming convention. Each folder is named in the format ‘[ID]-[CODE]‘, where the three-digit ID uniquely identifies the writer and the eight-character CODE (format ‘XXXXXX_Y‘) encodes the key demographic attributes. Specifically, the first two digits denote the writer’s age; the

TABLE 5: Demographic summary in the MPHD, including age group, education, handedness, spectacles usage, and gender.

Age	Count(%)	Education	Count(%)	Hand	Count(%)	Glasses	Count(%)	Gender	Count(%)
11–15	115 (23.0%)	Primary	100 (20.0%)	Right	455 (91.0%)	Yes	137 (27.4%)	Male	251 (50.2%)
16–20	208 (41.6%)	Secondary	100 (20.0%)	Left	45 (9%)	No	363 (72.6%)	Female	249 (49.8%)
21–25	114 (22.8%)	High School	100 (20.0%)						
26–30	50 (10.0%)	Bachelor	101 (20.2%)						
31–35	9 (1.8%)	Postgrad.	99 (19.8%)						
36–40	4 (0.8%)								

third digit represents the education level (1: Primary School, 2: Secondary School, 3: High School, 4: Bachelor, 5: Postgraduate); the fourth digit indicates gender (0: Female, 1: Male); the fifth digit specifies handedness (0: Right-handed, 1: Left-handed); the sixth digit denotes spectacles usage (0: No glasses, 1: Uses glasses); and the final digit after the underscore specifies the Text 2 category (1: Sport, 2: Health, 3: Technology, 4: Economy, 5: History). For instance, the folder '050-173010_3' corresponds to writer ID 50, aged 17, with a High School education, who is female, left-handed, does not use glasses, and wrote a Technology text. This encoding scheme not only facilitates efficient data organization but also allows for rapid demographic filtering and subset selection without parsing external metadata files.

3.3 Form Design and Data Structure

The MPHD dataset is organized around a carefully designed paper-based form that captures both constrained and unconstrained handwriting samples from each writer. This section describes the physical form design and the resulting multi-level data structure.

Form Design: The data collection form is divided into three main sections, as illustrated in Fig. 1. The first section contains the fixed text (Text 1), a carefully selected paragraph from Wikipedia designed to include all 32 Persian letters in different positional forms. The second section contains the variable text (Text 2), which is selected from one of five thematic categories. The third section consists of a grid of 90 cells where writers are asked to write isolated characters, including 62 distinct letter forms derived from the 32 Persian letters (covering the most common positional variations), 10 digits, 10 common punctuation marks, and 8 arithmetic symbols. This design enables the collection of handwriting at three levels of granularity: full-form documents, text regions, and isolated characters.

Data Structure: The dataset is organized at three levels of granularity to support a wide range of research tasks, from document-level analysis to character-level recognition. Where Table 6 summarizes the three levels of data organization and their corresponding applications.

Level 1: Full-Form Images Each writer's completed form is scanned and saved as a single image in PNG format. The full-form image provides the complete context of the handwriting, including the layout, margins, and overall writing style. These images are named as '[ID]-[CODE].png' (e.g., '050-173010_3.png') and are essential for document layout analysis and holistic handwriting analysis.

Level 2: Cropped Text Regions For each writer, the two text sections are cropped from the full-form image and saved as separate images. Text 1 (fixed) is saved as '[ID]-T1.png', and Text 2 (variable) is saved as '[ID]-T2-[CAT].png', where '[CAT]' denotes the thematic category (1–5). These cropped images allow researchers to focus on the text content without the distraction of the surrounding form elements, making them ideal for HTR and line segmentation tasks.

Level 3: Isolated Characters Each of the 90 isolated characters written in the grid is segmented and saved as an individual image. These images are stored in a subfolder named 'Chars/' within each writer's directory, with filenames following the pattern '[ID]_Char_XX.png' (e.g., '001_Char_01.png'). This level provides a clean dataset for character and digit recognition, handwriting style analysis, and demographic studies.

Directory Structure: All data for each writer is organized in a single folder with the naming convention '[ID]-[CODE]' (e.g., 050-173010_3). The folder contains the full-form image, the cropped text region images, the transcription file, and the 'Chars/' subfolder containing the isolated character images. This consistent structure ensures easy navigation and compatibility with common data loaders.

Transcription and Annotation: For each writer, all textual ground-truth information is consolidated into a single JSON file named [ID]-[CODE].json, which replaces the legacy plain-text format. This structured file contains the full paragraph-level transcriptions of both Text 1 and Text 2 under dedicated fields, as well as line-level transcriptions stored as arrays of strings, where each string corresponds to a single handwritten line in the corresponding cropped text image. This unified format preserves the complete annotation hierarchy required for both full-paragraph and line-level handwriting text recognition (HTR). The line-level annotations are particularly essential for modern sequence-to-sequence HTR models, which typically operate on segmented lines, and they enable direct compatibility with established evaluation protocols without additional preprocessing.

Metadata: A comprehensive metadata file ('metadata.csv') accompanies the dataset, providing structured access to writer demographics and file paths. Each row in the CSV file corresponds to one writer and includes columns for writer ID, age, education level, text category, gender, handedness, spectacles usage, and paths to all image and transcription files. This metadata enables researchers to easily filter and select subsets of data based on specific criteria.

3.4 Text Components

The MPHD dataset includes two distinct text components for each writer: a fixed text (Text 1) and a variable text (Text 2), whose design rationale and advantages are discussed in the introduction. The following subsections describe each text component in detail.

TABLE 6: Summary of the three-level data structure in the MPHD. Each level provides different granularity of handwriting data, supporting a wide range of research tasks.

Level	Description	File Format	Application
Level 1	Full-form image	`[ID]-[CODE].png`	Document layout analysis, holistic handwriting analysis
Level 2	Cropped text regions	`[ID]-T1.png`, `[ID]-T2-[CAT].png`	HTR, line segmentation, document classification
Level 3	Isolated characters	`Chars/[ID]_Char_XX.png`	Character recognition, handwriting style analysis

3.4.1 Fixed Text (Text 1)

The fixed text, referred to as *Text 1*, is a carefully selected paragraph from the Persian Wikipedia entry on *Gondishapur University (دانشگاه گندی‌شاپور)*. This text was chosen for its linguistic richness and its coverage of all 32 Persian letters, appearing in various positions such as initial, medial, final, or isolated. The text is identical for all 500 writers, providing a controlled reference for evaluating handwriting recognition models and writer identification methods.

Key characteristics: Text 1, serving as the fixed reference text, comprises 80 words and 368 characters (excluding spaces) in its canonical form. Across writers, the average length is 80.0 words and 367.9 characters per writer, nearly identical to the canonical form; the small residual deviation is due to a few writers who omitted or duplicated one or more words. The content is a historical passage on Gondishapur University, one of the oldest academic institutions in the Middle East, and is carefully designed to encompass all 32 Persian letters in their various positional forms of initial, medial, final, or isolated. This design makes Text 1 a controlled benchmark for evaluating both handwriting text recognition (HTR) models and writer identification algorithms, as the fixed content eliminates content variability while retaining natural writing diversity. A sample of Text 1, along with its English translation, is presented below.

> دانشگاه گندی‌شاپور چنانکه اعراب آن را «جندی‌شاپور» ذکر می‌کردند، یکی از آثار به جای مانده از سلسله ساسانیان است و با بیش از ۱۷ سده قدمت، از باستانی‌ترین دانشگاه‌های خاورمیانه محسوب می‌شود که از ویژگی‌های کم نظیری برخوردار است! گندی‌شاپور شهری است کهن در مختصات جغرافیایی شمال استان خوزستان و در نزدیکی شهرستان دزفول که امروزه ویرانه‌های آن باقی است. شهرت خاص آن به سبب بیمارستان و دانشگاهی که در آن بوده ثبت و ضبط شده است. از ویکی‌پدیا: ۱۳۹۴/۰۸/۲۶
>
> *"Gondishapur University, as the Arabs called it 'Jundishapur,' is one of the remaining monuments of the Sassanid dynasty. With a history of over 17 centuries, it is considered one of the oldest universities in the Middle East and possesses unique features! Gondishapur is an ancient city located in the geographical coordinates of northern Khuzestan province, near the city of Dezful, whose ruins still remain today. Its fame is particularly due to the hospital and university that were established there. From Wikipedia: 1394/08/26."*

The fixed text is saved as a separate image for each writer, cropped from the full-form image and named '[ID]-T1.png'. The corresponding transcription along with other related information is provided in the file '[ID]-[CODE].json'.

3.4.2 Variable Text (Text 2)

The variable text, referred to as *Text 2*, comprises 500 texts (one per writer), with 100 samples per category, each containing between 28 and 111 words (excluding spaces). On average, Text 2 contains 70.9 words and 299.2 characters (excluding spaces) per writer, resulting 35,434 words and 149,587 characters across the dataset. Text 2 is selected from a collection of texts covering five thematic categories which are Sport (Code 1), Health (Code 2), Technology (Code 3), Economy (Code 4), and History (Code 5). Sport-related texts primarily focus on football, including match reports, player profiles, and event coverage. The Health category encompasses topics such as disease prevention, wellness advice, and medical research findings. Technology texts cover areas including artificial intelligence, software development, and digital transformation. Economy-related samples discuss market trends, economic policies, and business strategies. Finally, the History category includes texts on Persian history, ancient civilizations, and historical analysis. This component is designed to evaluate HTR models on diverse topics and enable research in document classification and content-based retrieval too. Each writer's variable text is saved as a cropped image named '[ID]-T2-[CAT].png', where '[CAT]' is the category code (1–5). The corresponding transcription is in the JSON file. The following is a sample of Text 2 from the Health category:

> "بیماری تب زرد توسط نوعی پشه که آلوده به ویروس تب زرد است، منتقل می‌شود. این پشه با نیش زدن انسان‌های آلوده به تب زرد، نقش حامل ویروس را ایفا می‌کند و این ویروس را به انسان‌ها و پستانداران سالم منتقل می‌کند. قسمتی از آفریقا و آمریکای جنوبی شایع‌ترین مکان‌ها برای ابتلا به این بیماری می‌باشند."
>
> *"Yellow fever is transmitted by a type of mosquito that is infected with the yellow fever virus. This mosquito, by biting humans infected with yellow fever, acts as a carrier of the virus and transmits it to healthy humans and mammals. Parts of Africa and South America are the most common places for contracting this disease."*

To further characterize the textual properties of the dataset, Fig. 2 presents a detailed linguistic analysis. Panel (a) shows the distribution of characters per line by gender, indicating that female writers tend to produce slightly longer lines on average, though the overlap is substantial. Panel (b) visualizes the consistency of individual writing style by comparing character density between the fixed and variable texts, revealing that each writer maintains a consistent density across different content. Panel (c) reports the average word length across the five thematic categories, confirming that the topic does not introduce significant variation in lexical complexity, which ensures fair evaluation of content-based retrieval tasks.

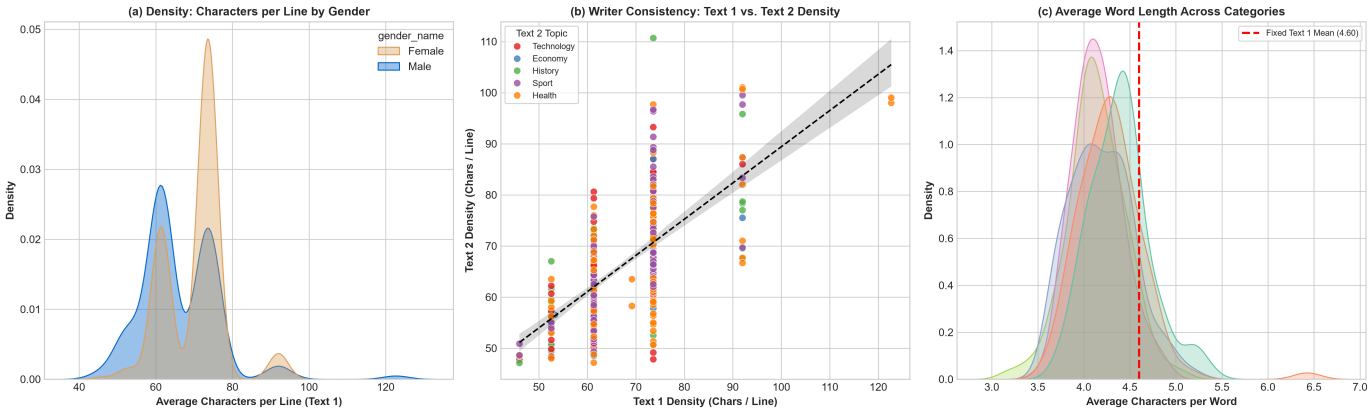


Fig. 2: Detailed linguistic analysis of the MPHD. (a) Density plot of characters per line, split by gender, revealing subtle differences in line length between genders. (b) Consistency plot comparing per-writer character density between Text 1 (fixed) and Text 2 (variable), demonstrating stable individual writing style across content. (c) Average word length across the five thematic categories showing that topic does not significantly affect word complexity, confirming the balanced design of Text 2.

3.5 Line-Level Transcriptions

A distinctive feature of the MPHD dataset is the integration of line-level transcriptions directly within the structured JSON file associated with each writer. While the full text of both the fixed and variable components is stored as continuous paragraphs to support document-level analysis, the dataset additionally preserves line-level annotations as distinct arrays of strings. This hierarchical design ensures that the ground-truth text aligns precisely with each handwritten line in the corresponding cropped image, a prerequisite for training modern handwriting text recognition (HTR) systems.

Contemporary HTR architectures, particularly those built upon recurrent neural networks or Transformer backbones, typically process handwritten content one line at a time. This line-wise formulation simplifies the sequence modeling objective, whether optimized via connectionist temporal classification or attention-based decoding. By embedding the line-level transcriptions into the unified JSON schema, MPHD eliminates the need for researchers to perform manual line segmentation or alignment, thereby enabling direct training and evaluation of such models with minimal preprocessing overhead.

For each writer, the cropped text region images, namely `[ID]-T1.png` and `[ID]-T2-[CAT].png`, were manually annotated at the line level by a human expert. For every image, the annotator delineated the natural line breaks made by the writer and recorded the corresponding text segments. These segments were then encoded as ordered lists within the corresponding JSON file, rather than being stored as separate plain-text files. This approach not only preserves the fidelity of the original writing layout but also ensures that the segmentation reflects the writer's intrinsic line-breaking behavior, rather than arbitrary algorithmic splits.

The JSON file follows a naming convention that mirrors the folder structure: `[ID]-[CODE].json`. Within this file, the line-level transcriptions for Text 1 and Text 2 are accessible under the `text1.lines.text` and `text2.lines.text` fields, respectively, with each entry corresponding sequentially to a handwritten line in the associated image. Additionally, the full paragraph-level transcriptions are retained under the `full_text` fields, accommodating researchers who prefer to work with whole-document inputs for applications such as language modeling or holistic document classification.

The provision of unified JSON annotations substantially lowers the barrier to entry for HTR research on MPHD. Practitioners can directly load the image and its corresponding line-level transcription from a single source, apply standard preprocessing steps (e.g., binarization or deskewing), and train sequence-to-sequence models without the overhead of managing multiple file formats. This streamlined pipeline enhances reproducibility across studies and enables fine-grained evaluation at the line level, facilitating the computation of character error rates (CER) and word error rates (WER), the de facto metrics in the HTR community. Consequently, results reported on MPHD remain directly comparable to those obtained on established benchmarks such as IAM, RIMES, and KHATT, while benefiting from a more coherent and maintainable data structure.

3.6 Isolated Characters

In addition to the two text components, the MPHD dataset incorporates a dedicated section within each form, where writers were asked to transcribe a set of 90 isolated characters in a structured grid. This component provides clean, segmented character-level data that is essential for training and evaluating character recognition models, as well as for studying the influence of demographic factors on the shape of individual glyphs.

While each Persian letter can theoretically appear in up to four positional forms (isolated, initial, medial, and final), not all letters exhibit four distinct shapes in practice. Letters such as د ر و, etc. remain essentially unchanged across positions, so they were requested only once. For هـ, which has four clearly different forms, an additional common everyday variant (without a standard typed counterpart) was also included, resulting in five forms. The character ئـ was likewise added due to its frequent use in everyday writing. Consequently, to capture the actual diversity of Persian handwriting without imposing artificial distinctions, the form includes 62 visually distinct letter forms, representing the most common

Fig. 3: Sample of 90 isolated characters from one writer, comprising 62 distinct letter forms (covering common positional variations of the 32 Persian letters), 10 digits, 10 punctuation marks, and 8 arithmetic symbols. The sample illustrates both the diversity of handwriting styles and the visual similarity among ambiguous letters such as ب، پ، ت، ث or ج، چ، ح، خ.

and meaningful variations, rather than requiring writers to produce all 128 theoretical possibilities.

Each character is written in a designated cell on the form, and the cells are subsequently segmented and saved as individual grayscale images. These images are stored in the `Chars/` subfolder within each writer's directory, following a systematic naming convention of the form `[ID]_Char_XX.png`, where `XX` denotes a two-digit index from 01 to 90. This structured organization ensures straightforward mapping between each character image and its corresponding label, facilitating both automated processing and manual verification. Fig. 3 shows a sample of the 90 isolated characters written by one of the writers, illustrating the variety of handwriting styles and the visual similarity among certain letter groups.

The isolated character component of MPHD serves multiple research purposes. It offers a clean, well-separated dataset for training and benchmarking character-level recognition models, which can underpin more complex HTR systems. Moreover, it enables detailed analysis of handwriting style variation across different demographic groups, such as age, gender, and education level. The isolated characters also provide a valuable resource for transfer learning, where models pre-trained on this clean data can be fine-tuned for more challenging tasks such as line-level or word-level recognition. Finally, the consistent alignment of these isolated forms with the corresponding entries in Table 1 ensures coherence across the dataset, reinforcing both the clarity and usability of MPHD for a wide range of research applications.

3.7 Metadata

To facilitate structured access to the dataset and enable demographic analyses, MPHD is accompanied by a comprehensive metadata file that provides essential information about each writer in a machine-readable format. This metadata serves two primary purposes: it enables researchers to filter and select subsets of the data based on specific criteria such as gender, age group, or education level and it provides direct paths to all associated image and transcription files, eliminating the need for manual file management and ensuring reproducibility across experiments. As described in Section 3.2, folder names follow the [ID]-[CODE] convention encoding age, education, gender, handedness, and spectacles usage.

In addition to the folder naming convention, all metadata is aggregated into a single CSV file (metadata.csv) with 500 rows and 26 columns. Additionally, per-writer JSON files ([ID]-[CODE].json) are provided for researchers preferring structured data, containing metadata, full text transcriptions, line-level annotations, and character information. This dual CSV+JSON release ensures compatibility with both traditional data science workflows and modern deep learning pipelines. This structure simplifies data loading in popular frameworks such as Pandas, TensorFlow, and PyTorch, enabling researchers to quickly integrate MPHD into their existing pipelines. The combination of structured folder naming and a comprehensive CSV metadata file makes MPHD exceptionally easy to use, allowing researchers to filter the dataset based on demographic attributes, generate custom splits for cross-validation, and develop reproducible experiments, all of which align with best practices in machine learning research and ensure that MPHD can serve as a reliable benchmark for future studies.

4 Experimental and Evaluations

To comprehensively evaluate the MPHD dataset and establish robust baselines for future research, we conduct four complementary experiments that span multiple levels of handwriting analysis. First, we evaluate handwriting text recognition (HTR) at both paragraph and line levels to assess the dataset's primary utility for transcription tasks. Second, we benchmark line segmentation, a critical preprocessing step that directly impacts the performance of downstream HTR systems. Third, we conduct writer identification in both closed-set and open-set settings to quantify the discriminative power of individual writing styles. Fourth, we assess fine-grained character and digit recognition to evaluate the dataset's suitability for fundamental visual classification tasks. For each task, we employ standard, reproducible baseline methods, ranging from pre-trained deep learning architectures to off-the-shelf OCR engines, and report a comprehensive set of metrics, including accuracy, F1-score, Intersection over Union (IoU), and Line Count Accuracy (LCA), to facilitate fair and transparent comparisons with future work. Implementation details and hyperparameters specific to each task are provided within the respective subsections.

4.1 Handwriting Text Recognition

To establish a rigorous benchmark for the primary task supported by MPHD, we evaluate handwriting text recognition using two complementary architectures: a standard CRNN baseline [29] and HTR-VT [30], a Vision Transformer-based model specifically designed for data-efficient HTR. This section reports comprehensive results across multiple evaluation protocols, full test-set evaluation, fixed vs. variable text comparison, and fine-grained error analysis that quantitatively confirm the intrinsic challenges of Persian script.

4.1.1 Experimental Setup

Dataset Preparation and Splits: We partition the 5,021 annotated text lines into training, validation, and test sets with a ratio of 70:15:15, strictly separated at the writer level to prevent data leakage. This yields 3,515 lines for training, 767 for validation, and 739 for testing. Both Text 1 (fixed) and Text 2 (variable) lines are included in all splits, preserving the full diversity of writing styles and thematic content. All line images are resized to a fixed height of 64 pixels while

TABLE 7: Handwriting text recognition results on the MPHD test set and performance comparison between Text 1 (fixed) and Text 2 (variable) on the test set for both models.

Model	Text Type	Count	CER (%)	WER (%)
CRNN	Text 1 & Text 2	739	37.65	57.90
	Text 1 (Fixed)	403	11.49	22.57
	Text 2 (Variable)	336	69.80	100.9
HTR-VT	Text 1 & Text 2	739	10.78	31.21
	Text 1 (Fixed)	403	2.17	5.46
	Text 2 (Variable)	336	21.34	62.53

preserving aspect ratio, and pixel values are normalized to the range $[0, 1]$.

Character Set: Our annotation process identifies a vocabulary of 68 unique characters (Unicode code points), encompassing the 32 Persian letters (each letter represented by a single code point irrespective of its positional form), Persian digits (۰–۹), punctuation marks, and special symbols such as the half-space character (ZWNJ, Unicode U+200C). With the CTC blank token, the total number of output classes is 69.

Baseline Model (CRNN): As a lower-bound baseline, we implement the standard CRNN architecture proposed by Shi et al. [29] for sequence recognition. Our implementation follows the original design with a VGG16-inspired CNN backbone, a 2-layer bidirectional LSTM with 256 hidden units, and CTC decoding. To accommodate the longer sequence lengths of Persian handwritten lines (average ∼80 characters per line, significantly longer than the short words in the original paper), we increase the input width from 512 to 1024 pixels to ensure sufficient sequence length for CTC alignment. All training configurations, including data augmentation (random rotation $\pm 5^\circ$, scaling 90–110%, and elastic distortions) and optimizer settings (AdamW, initial LR 10^{-3}), are kept as similar as possible to HTR-VT to ensure a fair comparison. The model is trained for 200 epochs with early stopping based on validation CER.

Primary Model (HTR-VT): We employ HTR-VT, a Vision Transformer tailored for handwritten text recognition that achieves competitive results without requiring large-scale pre-training, an ideal characteristic for moderate-sized datasets like MPHD. The model adopts a patch-based tokenization scheme with a Transformer encoder and employs CTC for sequence alignment. Training is performed using the AdamW optimizer with an initial learning rate of 4×10^{-4}, a batch size of 32, and the same on-the-fly data augmentation as CRNN. The model is trained for 40,000 iterations with early stopping based on validation loss.

Evaluation Metrics: We report Character Error Rate (CER) and Word Error Rate (WER), both computed using the Levenshtein distance. All metrics are reported as percentages, with lower values indicating superior performance.

Statistical Significance Testing: Statistical significance of CER differences was assessed via a non-parametric bootstrap procedure. For each comparison, we performed 20,000 resampling iterations at the line level, computing the micro-averaged CER (i.e., the ratio of total edit distance to total reference length) on each resample. For comparisons involving the same set of lines evaluated by two different models (e.g., CRNN vs. HTR-VT), we used a paired bootstrap, resampling line indices jointly for both models to preserve their correlation. For comparisons across disjoint subsets (e.g., Text 1 vs. Text 2), an unpaired bootstrap was used, resampling each subset independently. Two-sided p-values were estimated as $2\min(\Pr[\Delta^* \leq 0], \Pr[\Delta^* \geq 0])$ over the bootstrap distribution of the difference Δ^*, and 95% confidence intervals were obtained via the percentile method. All results use a fixed random seed of 42 for reproducibility.

4.1.2 Results and Discussion

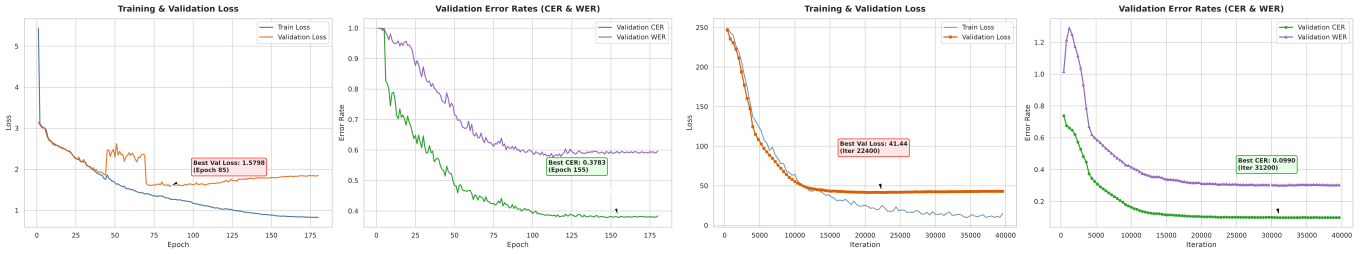
Overall Performance: Table 7 summarizes the performance on the held-out test set. The CRNN baseline achieves a CER of 37.65% and a WER of 57.90%, reflecting the inherent difficulty of Persian handwriting recognition. The HTR-VT model substantially outperforms the baseline, achieving a CER of 10.78% and a WER of 31.21%. This significant improvement underscores the effectiveness of the Vision Transformer architecture in capturing long-range contextual dependencies in cursive scripts, where character shapes are highly context-dependent and diacritical dots are visually subtle.

Performance on Fixed versus Variable Text: To quantify the generalization capability of the models, we separately evaluate on Text 1 (fixed reference text containing repeated vocabulary) and Text 2 (variable thematic text containing unseen word compositions). As shown in Table 7, both models perform substantially better on Text 1 than on Text 2. For CRNN, the CER on Text 1 is 11.49%, while on Text 2 it rises sharply to 69.80%. Moreover, the WER on Text 2 reaches 100.9%, meaning the total number of word-level edit operations (substitutions, deletions, and insertions) exceeds the total word count of the reference text, a consequence of a substantial number of insertion errors compounding the substitution and deletion errors, rather than indicating that every reference word was necessarily misrecognized. HTR-VT exhibits a similar but less severe pattern: 2.17% CER on Text 1 versus 21.34% on Text 2.

The performance gap between fixed and variable text is statistically significant for both architectures. For HTR-VT, the CER increases from 2.17% on Text 1 to 21.34% on Text 2 (ΔCER = 19.17 percentage points, 95% confidence interval (CI) [18.05, 20.31], $p < 0.001$). An even larger gap is observed for the CRNN baseline (11.49% vs. 69.80%, ΔCER = 58.31 percentage points, 95% CI [56.89, 59.66], $p < 0.001$). These results confirm that the dual-text design of MPHD constitutes a rigorous stress test for generalization in cursive Persian HTR.

HTR-VT significantly outperforms the CRNN baseline on both subsets. On the fixed text (Text 1), the absolute CER reduction is 9.32 percentage points (95% CI [8.20, 10.50], $p < 0.001$). On the more challenging variable text (Text 2), the improvement reaches 48.46 percentage points (95% CI [47.45, 49.42], $p < 0.001$).

Training Dynamics and Convergence: Figure 4 presents the training and validation dynamics for both evaluated architectures. For HTR-VT (Fig. 4b), the validation CER decreases rapidly during the first 10,000 iterations, stabilizes around 0.20 by iteration 15,000, and reaches its minimum of 0.0990 at approximately 31,200 iterations. The close alignment between training and validation curves until iteration 40,000 indicates effective regularization; the gradual divergence thereafter suggests the onset of mild overfitting, consistent with the error analysis presented below.



(a) Training dynamics of CRNN on MPHD. (b) Training dynamics of HTR-VT on MPHD.

Fig. 4: Training and validation curves for (a) CRNN and (b) HTR-VT on MPHD. Both architectures converge to stable minima; however, the Transformer-based HTR-VT achieves a lower validation CER (0.1035) compared to the CRNN (0.4093), highlighting the superior capacity of attention-based models for capturing long-range dependencies in cursive Persian text.

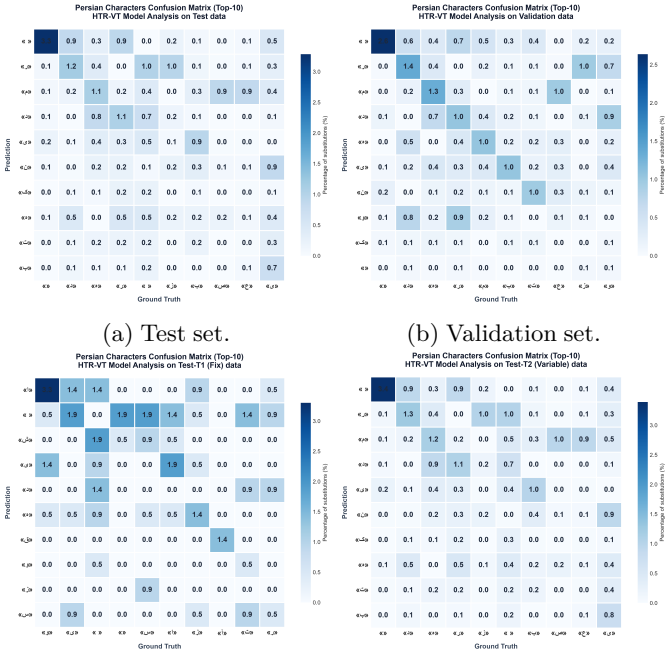
For CRNN (Fig. 4a), the model begins to converge after approximately 100 epochs, with the CER stabilizing around 0.38; the best validation CER of 0.3783 is achieved at epoch 155, with a stable gap between training and validation losses (≈0.85 versus ≈1.80) indicating that the model has reached its capacity limit rather than suffering from severe overfitting. The convergence behavior further confirms that MPHD provides sufficient training signal for both architectures, while the performance gap highlights the superiority of Transformer-based approaches for this challenging task.

Error Analysis: To systematically dissect the failure modes of the HTR-VT model and to isolate the source of generalization error, we perform a fine-grained character-level substitution analysis using Needleman-Wunsch [31] alignment. Critically, we decompose the test set into its two constituent components: Test-T1 (fixed text, comprising repeated vocabulary observed during training) and Test-T2 (variable thematic text, consisting entirely of unseen word compositions). This decomposition reveals a stark and highly instructive disparity in model behavior.

As illustrated in the confusion matrices (Fig. 5), the model achieves remarkably clean predictions on Test-T1, accumulating only 212 total substitutions across 403 lines (approximately 0.52 errors per line). The errors in this subset are sparse and predominantly shallow in nature, such as occasional space insertions or deletions, indicating that the model has successfully memorized the training vocabulary. In stark contrast, when evaluated on Test-T2, the model's performance degrades catastrophically, yielding 2,772 substitutions across 336 lines (approximately 8.25 errors per line), a more than 15-fold increase in error rate relative to the fixed-text subset. This profound degradation is not an artifact of random noise, but rather a systematic failure to generalize to novel lexical compositions, as further corroborated by the combined Test set (739 lines, 2,984 substitutions) and the Validation set (767 lines, 2,768 substitutions), both of which exhibit a nearly identical error distribution dominated by visually ambiguous Persian letter groups.

The confusion matrices for Test-T2 and the combined Test set reveal three dominant and persistent error categories, which directly validate the intrinsic challenges qualitatively discussed in Section 1 and Table 2:

(i) Diacritical Dot Confusions: Pairs such as (ب ↔ ی) in initial/medial positions, (گ ↔ ک), and (ز → ر) consistently rank among the top errors. These substitutions underscore the model's inability to reliably discriminate characters whose



(a) Test set. (b) Validation set.

(c) Test-T1: fixed text. (d) Test-T2: variable text.

Fig. 5: Character-level confusion matrices of HTR-VT on (a) Test, (b) Val, (c) Test-T1 (fixed/seen vocabulary), and (d) Test-T2 (variable/unseen vocabulary). The decomposition reveals a striking disparity: Test-T1 yields only 212 substitutions (0.5/line) versus 2,772 (8.3/line) on Test-T2 confirming overfitting to seen vocabulary and validating Persian ambiguities (Table 2). This gap reinforces MPHD's dual-text design as a rigorous benchmark for generalization in cursive HTR.

base shapes are identical in their contextual forms and whose differentiation hinges solely on the presence, number, and spatial placement of diacritical dots, a challenge that is particularly acute in Persian cursive handwriting, where the same dental shape represents up to four different phonemes depending solely on dot configuration.

(ii) Orthographic Boundary Errors: The substitution of a standard space with the ZWNJ emerges as the single most frequent error, occurring 94 times (3.4% of all Test-T2 substitutions) on Test-T2 alone, for example, predicting a full space where the ground truth requires a ZWNJ (e.g., "می‌شود" → "می شود"). This pattern confirms that the model struggles

TABLE 8: Line segmentation performance comparison across evaluation protocols. Protocol 1 evaluates the baseline (EasyOCR + vertical grouping) on the entire dataset of 500 writers (1,000 text images) to assess the upper-bound performance of off-the-shelf tools. Protocol 2 evaluates the same baseline on the official held-out test set (writers 426–500, 15% of data, 150 images) to provide a reproducible benchmark for future supervised methods. All metrics are averaged across the respective test images.

Protocol	mIoU(all)	mIoU(m)	F1@0.50	F1@0.75	LCA (%)	RelCntErr	MeanSSIM
Protocol 1 (All 500 Writers)	**0.4393**	**0.6724**	**0.6757**	**0.3813**	**18.20**	**0.5376**	**0.9103**
Protocol 2 (Test Set, 15%)	**0.4171**	**0.6455**	**0.6694**	**0.2968**	**20.67**	**0.5510**	**0.9056**

to capture Persian's context-dependent orthographic spacing conventions, a subtle visual cue that is highly variable across writers and writing styles, and is rarely present in Latin-script benchmarks.

(iii) Stroke Curvature and Structural Confusions: High-frequency confusions between structurally similar letters that share no diacritical marks, such as (د ↔ ر) and (ه ↔ م) and (ه → د), dominate the error list. These errors indicate that the model is particularly vulnerable to variations in stroke curvature, loop shapes, and overall body geometry, a challenge that fundamentally distinguishes cursive non-Latin scripts from their Latin counterparts, where such fine-grained structural distinctions are often less critical.

Critically, the near-perfect performance on Test-T1 (which contains vocabulary present in the training set) confirms that the Transformer architecture is, in principle, capable of achieving high recognition accuracy on Persian script when provided with familiar lexical patterns. However, the catastrophic failure on Test-T2 irrefutably demonstrates that this capability does not stem from a robust, generalizable understanding of the script's underlying structure. Instead, the model heavily overfits to training-specific vocabulary, failing to generalize fine-grained diacritic placement and structural stroke rules to unseen character compositions, a phenomenon that is substantially more pronounced than comparable evaluations on Latin benchmarks such as IAM [4].

These findings carry profound implications for the future of HTR research. They empirically establish that Persian cursive script presents a fundamental bottleneck for current Vision Transformer architectures, specifically, a lack of inductive bias for discriminating sub-pixel dot placements and subtle stroke variations in the absence of lexical context. We posit that addressing this challenge will require specialized architectural innovations, including but not limited to: (i) explicit dot-attention modules that localize and reason about diacritical marks independently of the base character; (ii) higher-resolution patch embeddings or multi-scale feature extractors that preserve fine-grained stroke details; and (iii) the integration of linguistic priors (e.g., Persian BERT-based language models for post-processing rescoring). The MPHD dataset, with its adversarial dual-text design (fixed vs. variable), provides a uniquely rigorous stress test for developing, evaluating, and validating such future solutions, thereby establishing itself as a cornerstone benchmark for advancing research on cursive script understanding.

4.2 Line Segmentation

To evaluate the practical utility of the MPHD dataset for document preprocessing tasks, we conducted a line segmentation benchmark. This task is a fundamental prerequisite for many handwriting recognition pipelines, as it involves partitioning a text image into individual text lines, enabling subsequent line-level HTR and layout analysis. Unlike existing datasets that often provide line-level crops directly, MPHD provides only full-form and cropped text-region images, requiring segmentation algorithms to infer line boundaries from raw images. This design makes MPHD a challenging benchmark for evaluating line segmentation methods under realistic, unconstrained conditions.

4.2.1 Experimental Setup and Protocols

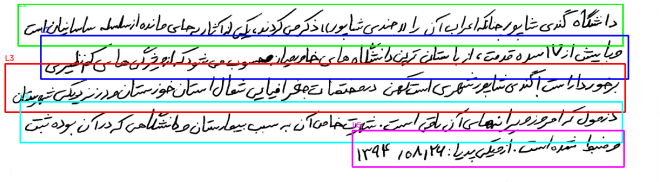
We establish two evaluation protocols to accommodate different types of line segmentation methods and to maximize the utility of the MPHD dataset for future research. Protocol 1 targets unsupervised or pre-trained models such as classical image processing techniques or off-the-shelf OCR engines like EasyOCR [32] and reports results on the entire dataset of 500 writers (1,000 text images) to assess the upper-bound performance of general-purpose tools under ideal conditions. Protocol 2, intended for supervised deep learning methods (e.g., fully-convolutional networks or Transformers), follows the official dataset split of 70% training, 15% validation, and 15% test, strictly separated at the writer level to prevent data leakage; the test set comprises writers 426–500 (75 writers, 150 images) and serves as the primary benchmark for comparing newly developed segmentation algorithms.

Our baseline method, which employs EasyOCR [32] for word detection coupled with a vertical grouping strategy based on the y-coordinates of word bounding boxes, does not require any training on MPHD. We therefore report its performance under both protocols, Protocol 1 to evaluate generalizability and Protocol 2 to provide a reproducible baseline for future comparisons. For Protocol 2, all hyperparameters were kept identical to those used in Protocol 1 and were not adapted to the test set, ensuring a fair and unbiased evaluation.

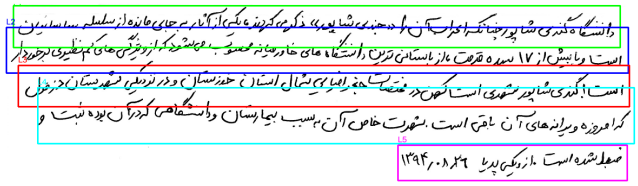
To obtain ground-truth line locations for quantitative evaluation, we employed a feature-based matching approach using SIFT [33]. Since MPHD provides line-level ground truth as cropped image patches rather than bounding box coordinates, SIFT keypoints were extracted from both each cropped patch and the corresponding full text-region image; homography estimation was then used to recover the precise bounding box of each ground-truth line within the original image. This procedure yielded a complete set of ground-truth bounding boxes for every text image, enabling the computation of all reported metrics.

4.2.2 Evaluation Metrics

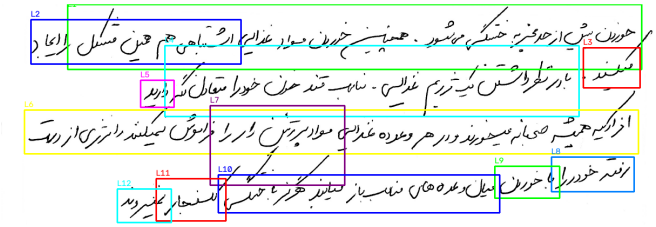
Following standard practice in document analysis [34], [35], we report a comprehensive set of evaluation metrics. The mean Intersection over Union computed over all predicted and



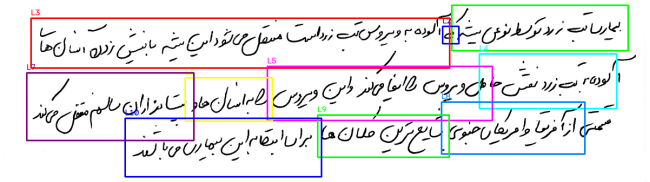
(a) High-quality segmentation: `465-T1` (mIoU = 0.816).



(b) High-quality segmentation: `484-T1` (mIoU = 0.839).



(c) Challenging case: `447-T2` (mIoU = 0.167). The baseline over-segments due to tightly spaced lines.



(d) Challenging case: `499-T2` (mIoU = 0.110). Severe over-segmentation caused by slanted handwriting.

Fig. 6: Qualitative line segmentation on the MPHD test. Each image shows the predicted line bounding boxes overlaid on the original region. (a)–(b) present successful cases where the baseline accurately separates lines with high IoU. (c)–(d) illustrate failure cases where the method over-segments due to large inter-word spacing or slanted writing, show challenges of the dataset.

ground-truth line pairs, where unmatched lines contribute a value of zero, is denoted as mIoU(all), which penalizes both missed and extra lines. The matched mean IoU, mIoU(m), is computed exclusively over optimally matched pairs using a greedy one-to-one assignment, reflecting localization quality when line counts are accurate. We also report F1-scores at IoU thresholds of 0.50 and 0.75 (F1@0.50 and F1@0.75), treating a detection as a true positive only if its IoU with the matched ground-truth exceeds the respective threshold. Line Count Accuracy (LCA) is defined as the percentage of images where the number of predicted lines exactly matches the ground-truth line count, while the relative count error (RelCntErr) measures the average absolute deviation in line counts normalized by the ground-truth count. Finally, we compute the mean Structural Similarity Index Measure (MeanSSIM) [36] between binary masks of the predicted and ground-truth line regions to assess the structural fidelity of the segmentation output.

4.2.3 *Results and Discussion*

Table 8 summarizes the line segmentation performance on the MPHD dataset. Under Protocol 2 (the official test set), the baseline achieves an mIoU(all) of 0.4171 and an mIoU(m) of 0.6455. The notable gap of approximately 0.23 between these two values indicates that while the algorithm effectively localizes the bounding boxes of the lines it does detect, it frequently fails to estimate the correct number of lines in a document. This is further corroborated by the relatively low LCA of 20.67% and the RelCntErr of 0.5510, confirming that the primary challenge for segmentation on this dataset is not the spatial placement of detected lines, but rather the accurate separation of merged lines or the correct splitting of fragmented ones.

The F1@0.50 score of 0.6694 demonstrates that a majority of line detections align reasonably well with the ground-truth at a lenient threshold. However, the sharp decline to 0.2968 at F1@0.75 reveals that the boundaries of the predicted bounding boxes are often imprecise, typically exhibiting deviations of a few pixels in height or horizontal margins, a limitation inherent to using a general-purpose OCR engine optimized primarily for text transcription rather than precise geometric segmentation. Interestingly, despite the moderate IoU and F1 values, the MeanSSIM reaches 0.9056, indicating that the overall shape and structure of the segmented text regions are well preserved, even if exact pixel-perfect boundaries differ slightly from the ground-truth. This confirms that the baseline method captures the global document layout effectively.

Comparing the two protocols, Protocol 1 (all 500 writers) yields consistently higher performance across most metrics: mIoU(all) of 0.4393, mIoU(m) of 0.6724, and F1@0.50 of 0.6757. These improvements reflect the inclusion of more diverse and, in some cases, less challenging samples in the full dataset. Notably, F1@0.75 rises to 0.3813 under Protocol 1, a relative improvement of nearly 29% over Protocol 2. This substantial gap confirms that the test set is enriched with challenging samples, a deliberate design choice that enhances the benchmark's rigor for future evaluations. Conversely, LCA, where higher values indicate better performance, is slightly higher on the test set (20.67% vs. 18.20% for all writers), indicating that line counting errors are not uniformly distributed and may be more pronounced in the broader dataset. Fig. 6 shows qualitative segmentation results. The baseline performs more reliably on denser text lines with relatively consistent spacing (Figs. 6a–6b), while it tends to over-segment when inter-word gaps are larger or the writing exhibits greater slant and stylistic variation (Figs. 6c–6d). This visual evidence aligns closely with the quantitative metrics reported in Table 8.

To further analyze the segmentation quality across protocols, we compare the distribution of IoU values for Protocol 1 and Protocol 2 in Fig. 7. Both distributions exhibit a pronounced peak in the high-IoU region (0.75–0.95), confirming that the baseline method accurately localizes the majority of lines regardless of the evaluation set. However, a notable

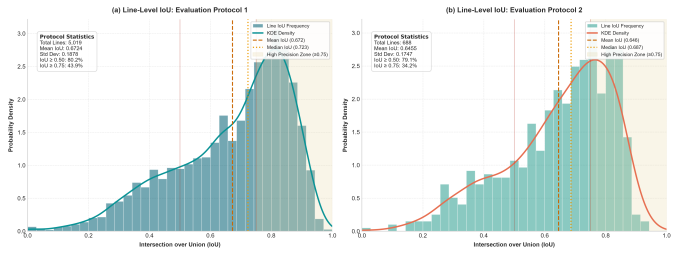


Fig. 7: Distribution of line-level IoU for (a) Protocol 1 (all 500 writers) and (b) Protocol 2 (official 15% test set), shown as histograms with overlaid Kernel Density Estimation (KDE) curves; dashed/dash-dotted lines mark the mean/median IoU, and the shaded region (IoU ≥ 0.75) denotes high-precision segmentations. Both protocols show a pronounced peak in the high-IoU region (0.75–0.95), indicating robust localization, but Protocol 2 exhibits a heavier left tail (IoU < 0.50), reflecting a greater share of edge cases, tightly spaced, slanted, or overlapping lines, in the held-out test set, per Table 8.

difference emerges in the lower tail: Protocol 2 displays a considerably heavier left tail with IoU values below 0.50 compared to Protocol 1. This indicates that the official held-out test set contains a disproportionately larger number of difficult samples such as tightly spaced lines, overlapping ascenders and descenders, or significant writing slant. This observation is consistent with the quantitative results in Table 8, where Protocol 1 consistently outperforms Protocol 2 across all metrics (e.g., mIoU(all) of 0.4393 vs. 0.4171), while Protocol 2 provides a more rigorous evaluation scenario that better reflects real-world variability. The comparative analysis reinforces that the primary challenge in line segmentation on MPHD is not the spatial localization of detected lines, but rather the precise separation of adjacent lines, a difficulty that is deliberately amplified in the test set to ensure a robust benchmark for future research.

These results establish a strong, reproducible baseline for line segmentation on MPHD. The relatively low LCA and mIoU(all) across both protocols indicate that the dataset presents a significant challenge for segmentation algorithms, primarily due to the variability in line spacing, the prevalence of slanted handwriting, and occasional overlapping strokes typical of Persian cursive script. Future work may explore more advanced techniques, such as fully-convolutional networks for pixel-wise line mask prediction or Transformer-based layout analysis models, which could address the over-segmentation errors observed in tightly spaced and slanted lines. The MPHD test set, with its deliberate enrichment of such challenging cases, provides a rigorous benchmark for evaluating these future approaches. In line with this direction, our recent work on heatmap-guided character localization in degraded historical documents [37] suggests that adopting a UNet-style architecture could effectively address the over-segmentation issues observed in challenging cases (e.g., slanted or tightly spaced lines), offering a promising upgrade path for MPHD benchmarks. By providing this comprehensive evaluation, we demonstrate that MPHD not only serves as a valuable resource for HTR and writer identification but also functions as a rigorous benchmark for advancing the state of the art in document layout analysis and line segmentation.

TABLE 9: Closed-set writer identification results on the MPHD test set. Document-level aggregation (softmax probability averaging) significantly improves performance over patch-level predictions.

Level	Top-1 (%)	Top-3 (%)	Top-5 (%)
Patch-level (128)	57.19	74.09	80.59
Document-level	95.19	98.20	99.00

4.3 Writer Identification

We conducted writer identification experiments in both closed-set and open-set settings to evaluate MPHD's ability to capture distinctive handwriting styles, relevant to applications such as forensic document analysis and authentication systems. All experiments follow an offline, text-independent protocol, where the fixed text (Text 1) images are used for training and validation, while the variable text (Text 2) images are reserved exclusively for testing to ensure text-independence.

4.3.1 Closed-Set Writer Identification

We followed a standard patch-based writer identification protocol. For each document, overlapping patches of size 128×128 pixels were extracted with a stride of 64 pixels. Patches with less than 5% ink pixels were discarded to focus on informative regions. The dataset was split at the writer level into training (80%) and validation (20%) sets, resulting in 29,163 training patches and 7,533 validation patches across all 500 writers. The test set consisted of 29,354 patches extracted from the 500 Text 2 images.

We used a ResNet-18 backbone pre-trained on ImageNet as the feature extractor, replacing the final classification layer with a 500-way softmax. The model was trained using cross-entropy loss with an AdamW optimizer (learning rate 10^{-4}, weight decay 10^{-4}) for up to 50 epochs. A learning rate scheduler (ReduceLROnPlateau) and early stopping (patience of 10 epochs) were employed to prevent overfitting. At test time, predictions were aggregated at the document level using softmax probability averaging.

Results and Discussion: Table 9 summarizes the closed-set performance on the test set. The patch-level Top-1 accuracy was 57.19%, which is expected given that individual patches contain limited contextual information. However, when aggregated at the document level, the accuracy improved significantly to 95.19% for Top-1, 98.20% for Top-3, and 99.00% for Top-5, indicating that the correct writer is almost always ranked among the top five candidates.

The substantial gap between patch-level and document-level accuracy confirms that handwriting style is a global property of a document, requiring aggregation of local features to achieve reliable identification. The near-perfect Top-5 accuracy demonstrates that the MPHD dataset captures writer-specific characteristics effectively, even in a text-independent scenario where the content of the training and test texts differs. While we adopt a standard ResNet-18 backbone for this baseline, our prior investigation into CapsNet-ResNet fusion for behavioral biometrics [40] indicates that incorporating capsule-based feature aggregation could further improve the document-level Top-1 accuracy, which we reserve for future extension.

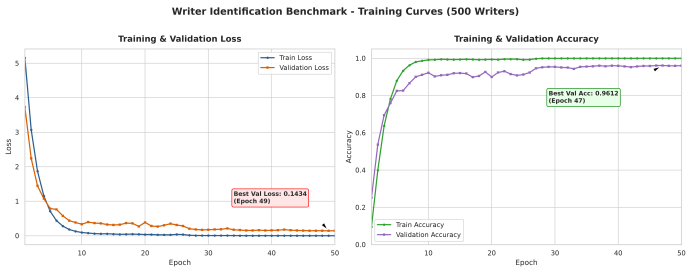


Fig. 8: Training and validation curves for the writer identification benchmark. The left panel presents the cross-entropy loss, and the right panel shows the accuracy over 50 epochs. The best validation loss (0.1434) was achieved at epoch 49, and the highest validation accuracy (96.12%) at epoch 47. The convergence of the training and validation curves indicates that the model generalizes well without overfitting.

The learning dynamics of the writer identification benchmark are illustrated in Fig. 8. The training loss decreases rapidly during the first 10 epochs, from 5.15 to 0.09, while the validation loss follows a similar trend, dropping from 3.72 to 0.33 over the same period. After epoch 20, the model continues to improve gradually, reaching its lowest validation loss of 0.1434 at epoch 49 and its highest validation accuracy of 96.12% at epoch 47. The close alignment between the training and validation curves throughout the training process indicates that the model does not suffer from significant overfitting, despite the large number of parameters in the ResNet-18 backbone. This stable convergence confirms the robustness of the training protocol and the discriminative power of the MPHD dataset for writer identification tasks.

4.3.2 Open-Set Writer Identification

To evaluate the dataset under more realistic and challenging conditions, we conducted an open-set writer identification experiment using a 5-fold rotation protocol. The 500 writers were randomly partitioned into 5 disjoint folds of 100 writers each. In each of the 5 rounds, a classifier was trained from scratch on the Text 1 (fixed-text) documents of the remaining 400 writers only (known identities); the 100 writers assigned to that round's fold were never seen during training and served as unknown impostors. Unlike a closed-set classifier evaluated post hoc, the model therefore has no parametric knowledge of the unknown writers whatsoever.

For each round, the acceptance threshold for a target False Alarm Rate (FAR) was calibrated using the Text 1 documents of that round's 100 unknown writers, by computing the maximum averaged softmax probability per document and setting the threshold at the corresponding percentile of this calibration distribution. Critically, this calibration set is disjoint from the test set used to report results: the Detection & Identification Rate (DIR) was computed on the Text 2 (variable-text) documents of the 400 known writers (a document is counted as a hit if it is accepted above threshold *and* correctly identified), while the actual FAR was computed on the Text 2 documents of that round's 100 unknown writers (accepted above threshold despite being an impostor). Results across all 5 rounds were pooled by summing counts, rather than averaging percentages, before computing the final DIR and FAR, so that every one of the 500 writers is evaluated exactly once as an unknown impostor while every round retains a full 400-writer known training set. A small number of documents (4 known, 1 unknown) yielded no valid ink-containing patches under the extraction criteria and were excluded, leaving $N_{\text{known}} = 1{,}996$ and $N_{\text{unknown}} = 499$ document instances pooled across the 5 folds.

TABLE 10: Open-set writer identification results on the MPHD dataset under 5-fold rotation. For each target FAR, the threshold is calibrated per fold on that fold's unknown writers' Text 1 documents; the pooled DIR and actual FAR are reported on the corresponding Text 2 documents, aggregated by summed counts across 5 folds ($N_{\text{known}} = 1{,}996$, $N_{\text{unknown}} = 499$).

Target(%) FAR (%)	Actual FAR(%)	DIR (%)	Threshold (range across folds)
1	0.60	29.51	0.4385 – 0.8346
5	4.41	61.22	0.3168 – 0.4998
10	7.41	72.80	0.2852 – 0.3749

Results and Discussion: Table 10 presents the pooled open-set results across the 5 folds. At the target FAR of 1%, the pooled DIR reached 29.51%, with an actual (measured) FAR of 0.60%. At FAR=5%, DIR reached 61.22% (actual FAR 4.41%), and at FAR=10%, DIR reached 72.80% (actual FAR 7.41%), with the actual FAR consistently below the target at every level reflecting the coarse resolution of calibrating each fold's threshold from only 100 unknown-writer documents. Compared to the closed-set document-level Top-1 accuracy (95.19%, Table 9), the open-set DIR at a strict 1% target FAR is dramatically lower, underscoring that reliably rejecting previously unseen writers while still correctly identifying known ones is a substantially harder problem than distinguishing among a fixed, previously observed set of identities. Because the model in this protocol is trained exclusively on the 400 known writers of each fold and the threshold is calibrated on a disjoint set of unknown-writer documents, this gap reflects a genuine open-set generalization limitation rather than an artifact of threshold selection on the test set or of the classifier having implicitly seen the unknown writers during training.

We further observe substantial variability across the 5 folds, particularly at the strict 1% target, where per-fold DIR ranges from 8.50% to 55.64% and the calibrated threshold ranges from 0.4385 to 0.8346 (Table 10). This variability arises because each fold's threshold is calibrated from only 100 unknown-writer documents, making percentile estimation at an extreme tail (the 99th percentile) inherently sensitive to sampling noise; the pooled DIR should therefore be interpreted with this fold-to-fold spread in mind. This makes MPHD a valuable and appropriately challenging benchmark for future research in open-set writer identification, e.g., via distance-based methods, OpenMax, or generative approaches to unknown-class rejection that may offer more stable calibration behavior than simple softmax thresholding.

4.4 Character and Digit Recognition

We conducted character (62 classes) and digit (10 classes) recognition experiments using the isolated character samples described in Section 3.5, evaluating MPHD's suitability for the fundamental visual recognition problems that underpin more complex handwriting analysis systems.

TABLE 11: Character (62 classes) and digit (10 classes) recognition results (%, macro-averaged) on MPHD test set.

Task	Model	Acc.	Prec.	Rec.	F1
Character	ResNet-18	96.34	96.45	96.34	96.35
	DenseNet-121	96.47	96.55	96.47	96.48
Digit	ResNet-18	99.07	99.09	99.07	99.07
	DenseNet-121	99.20	99.20	99.20	99.20

4.4.1 Experimental Setup

For both tasks, the dataset was split into training (70%), validation (15%), and test (15%) sets at the writer level to prevent data leakage. For character recognition, this resulted in 21,700 training images, 4,650 validation images, and 4,650 test images. For digit recognition, the corresponding splits yielded 3,500 training, 750 validation, and 750 test images.

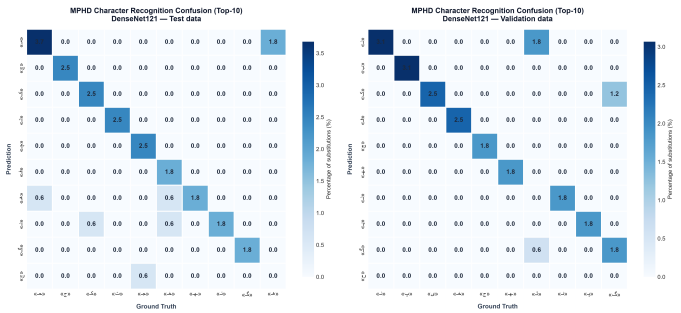
To compensate for the limited number of samples per class, we applied on-the-fly geometric augmentation to the training sets in both tasks, including random rotation ($\pm 7°$), translation ($\pm 6\%$), and scaling (92%–108%). Additionally, each base image was virtually repeated three times per epoch using a dataset multiplier, increasing the effective number of training samples seen per epoch to approximately 65,100 samples per epoch for character recognition and 10,500 samples per epoch for digit recognition.

We evaluated two standard CNN architectures: ResNet-18 and DenseNet-121. Both models were initialized with ImageNet-pretrained weights, with the first convolutional layer adapted to accept single-channel grayscale input (mean and standard deviation normalized to 0.5 and 0.5). Training was performed using the Adam optimizer with an initial learning rate of 10^{-3}, weight decay of 10^{-4}, and a batch size of 64. A learning rate scheduler was applied to reduce the learning rate when the validation loss plateaued, and early stopping with a patience of 10 epochs was used to prevent overfitting. The models were trained for a maximum of 50 epochs, with the same hyperparameters used across both tasks to ensure fair comparison.

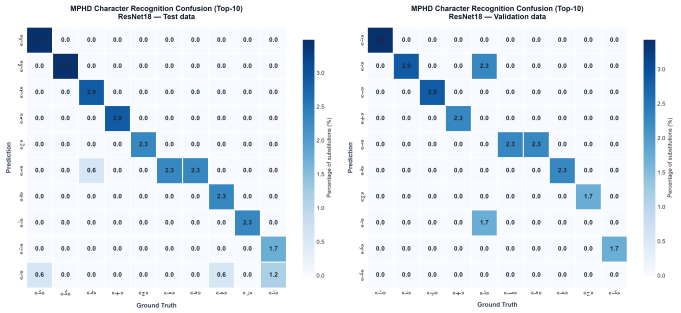
4.4.2 Results and Discussion

Table 11 summarizes the performance of both models on the character and digit recognition test sets. For character recognition, both ResNet-18 and DenseNet-121 achieved high accuracy, with DenseNet-121 slightly outperforming ResNet-18 (96.47% vs. 96.34%). Macro-averaged precision, recall, and F1-scores closely matched the overall accuracy, indicating uniform performance across all 62 classes without bias toward particular character shapes. The marginally better performance of DenseNet-121 can be attributed to its dense connectivity, which facilitates gradient flow and encourages feature reuse, a property particularly beneficial for fine-grained classification tasks with limited data per class.

For digit recognition, both architectures achieved near-perfect accuracy, with ResNet-18 reaching 99.07% and DenseNet-121 reaching 99.20%. Similarly, the macro-averaged precision, recall, and F1-scores closely match the overall accuracy, indicating balanced performance across all 10 digit classes. Unlike character recognition, where visual similarity between letters introduces inherent ambiguity, Persian digits are generally more discriminative, resulting in a simpler classification task. This aligns with other digit benchmarks such as MNIST and Hoda, which report similarly high accuracies using standard CNN architectures.



(a) DenseNet-121 on the test and validation sets.



(b) ResNet-18 on the test and validation sets.

Fig. 9: Normalized confusion matrices for the 62-class isolated character recognition task; strongest residual confusions occur within diacritic-differentiated and isolated/contextual letters.

Both models converged well within 50 epochs, with early stopping triggered sooner for digit recognition (19–28 epochs) than for character recognition (24–33 epochs), reflecting the relative simplicity of the digit task. The high performance on both tasks, combined with the earlier results, demonstrates that MPHD is a well-annotated and versatile dataset suitable for a wide range of handwriting analysis tasks, from digit recognition [39], [38] and fine-grained character-level recognition to holistic document analysis. The near-perfect digit accuracy, while not a challenging benchmark for modern deep learning models, serves as a valuable sanity check and provides a solid baseline for future research on Persian digit recognition, while the character recognition results establish a more demanding benchmark that reflects the inherent complexity of Persian script.

Despite the high overall accuracies achieved by both DenseNet-121 and ResNet-18, a fine-grained inspection of the normalized confusion matrices for both architectures (Fig. 9) reveals that residual errors are far from random. The dominant substitutions consistently occur within the well-known visually ambiguous Persian letter groups that differ solely by the number or placement of diacritical dots (e.g., ب↔پ, ج↔چ, ت↔ث) as well as between isolated and contextual forms of the same letter (e.g., ک↔ـکـ, گ↔ـگـ). Additional recurrent confusions such as ح↔ـمـ and خ↔ـفـ further indicate sensitivity to subtle stroke curvature and loop geometry. These patterns, observed on both the test and validation sets and across both architectures, quantitatively corroborate the intrinsic visual challenges of Persian script that were

qualitatively illustrated in Table 2. Importantly, the fact that such confusions persist even under the idealized conditions of cleanly segmented, isolated characters underscores that the difficulty is not merely an artifact of cursive connectivity or line-level context, but is rooted in the fine-grained visual similarity of the glyphs themselves. This finding strengthens the motivation for specialized architectural components (e.g., explicit diacritic attention) in future Persian HTR systems.

4.5 Discussion and Analysis

Comparison Across Tasks: The four benchmarks established on MPHD reveal a clear hierarchy of task difficulty, reflecting differing levels of granularity and context. Character and digit recognition, operating on cleanly segmented isolated glyphs, achieve near-perfect accuracy, indicating high-quality, well-separated samples suitable for fine-grained recognition. Writer identification, leveraging global style cues across full documents, also achieves strong performance (95.19% Top-1 accuracy), confirming that MPHD captures writer-specific characteristics effectively. In contrast, handwriting text recognition (HTR) and line segmentation, which require holistic understanding of cursive structure and layout, present substantially greater challenges. HTR-VT achieves only 10.78% CER on the combined test set, with performance on variable text (21.34% CER) lagging far behind fixed text (2.17% CER), a gap that highlights the model's reliance on vocabulary memorization rather than genuine script understanding. Similarly, line segmentation achieves an mIoU(all) of only 0.4171 on the test set, with line count accuracy as low as 20.67%, underscoring the difficulty of accurately separating adjacent lines in Persian cursive handwriting. This hierarchy of difficulty confirms that MPHD is not merely a collection of isolated samples, but a multi-level benchmark that challenges models at varying degrees of structural complexity.

Impact of Demographic Factors: The rich demographic metadata in MPHD, including age, gender, education level, handedness, and spectacles usage, enables systematic investigation into the relationship between writer characteristics and handwriting style. While a comprehensive demographic analysis is beyond the scope of this paper, the balanced distribution across these attributes (Table 5) provides a robust foundation for future studies on behavioral biometrics and forensic document analysis. Several concrete directions illustrate this potential: the isolated character samples could serve as a controlled testbed for studying the effects of age and education on glyph formation, while the paired fixed and variable texts enable disentangling content-dependent from content-independent stylistic features. The structured multi-level design (full-form pages, cropped text regions, and line-level annotations) further facilitates research on document layout analysis and multi-modal learning, and the cleanly segmented character images are well-suited for transfer learning and few-shot recognition experiments benefiting low-resource scripts. We released the full metadata to encourage the research community to explore these directions.

Limitations: Despite its comprehensive design, MPHD has several limitations that should be acknowledged. First, the dataset is limited to 500 writers and covers only conventional Persian handwriting styles, using their natural, legible handwriting style. It does not fully include more calligraphic styles such as Nasta'liq or Shekasteh, which pose substantially different challenges. Second, the age range is skewed toward younger writers, which may limit generalization to older adult populations. Third, line-level annotations follow the writers' natural line breaks, including occasional mid-word breaks that, while visually accurate, are linguistically unnatural. Fourth, our open-set writer identification results (DIR of 29.51% at a 1% target FAR, pooled across a 5-fold rotation protocol in which the model was trained exclusively on known writers and the threshold calibrated on a disjoint set) reveal substantial fold-to-fold variability (8.50%–55.64% DIR at this strict target) and confirm that simple softmax thresholding is insufficient for robust, stable open-set recognition; more sophisticated methods (e.g., OpenMax, distance-based classifiers) are needed to handle writers. Finally, while MPHD provides demographic metadata, the sample size within each demographic subgroup (e.g., left-handed writers: 45 individuals) is too small for statistically powered subgroup analyses. We encourage the research community to expand upon MPHD with additional writers and writing styles to further enrich this resource.

5 Conclusion

Persian, with its rich literary and historical legacy, represents a significant domain for computational document analysis. Despite its importance, research in Persian handwriting recognition has remained constrained by the acute scarcity of comprehensive, publicly accessible benchmark datasets that adequately reflect the script's inherent cursive complexity and rich morphological diversity. To bridge this critical and long-standing resource gap, we introduced MPHD, a multi-purpose offline Persian handwriting dataset comprising 500 writers, 75,430 words, and 333,545 characters across 5,021 annotated text lines. Organized at three levels of annotation (full-form, cropped text, and isolated characters) and enriched with detailed demographic metadata, MPHD is, to the best of our knowledge, the first publicly available dual-text design Persian resource that simultaneously supports HTR, writer identification, document classification, line segmentation, and character recognition within a unified framework.

Extensive benchmarking across these tasks reveals a pronounced hierarchy of difficulty. While isolated character/digit recognition achieve near-perfect accuracy, writer identification remains strong, line segmentation and especially handwriting text recognition remain far from solved. The most striking finding is the large and statistically significant performance gap between fixed and variable text (2.17% vs. 21.34% CER for HTR-VT), demonstrating that current Vision Transformer models still rely heavily on lexical memorization rather than robust, script-level visual understanding. This dual-text design therefore constitutes a particularly rigorous stress test for generalization in cursive, right-to-left scripts.

By releasing MPHD with complete annotations, metadata, and evaluation protocols, we provide the community with a challenging, reproducible, and publicly available benchmark. We hope that MPHD will stimulate the development of architectures specifically equipped to handle fine-grained diacritic discrimination, contextual shape variation, and long-range dependencies characteristic of Persian handwriting, while also enabling broader research at the intersection of computer vision, digital humanities and behavioral biometrics.

References

[1] C. Garrido-Munoz, A. Rios-Vila and J. Calvo-Zaragoza, Handwritten Text Recognition: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 12, pp. 4367–4387, 2025, doi: 10.1109/TPAMI.2025.3646002.

[2] R. Li, L. Zhu and H. Huang, "Bridging the Gap in Exam Handwritten Text Recognition: Dataset, Benchmark, and Modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2026.3721731.

[3] Marco Peer, Anna-Scius Bertrand, Patricia Scheurer, Andreas Fischer, BullingerDB: A Dataset for Handwritten Text Recognition and Writer Retrieval, arXiv:2605.30235, 2026.

[4] Urs-Viktor Marti and Horst Bunke, The IAM-database: an English sentence database for offline handwriting recognition, Intl. Journal on Document Analysis and Recognition, 5 (2002) 39–46.

[5] Emmanuel Grosicki, Matthieu Carré, Jean-Marie Brodin, and Edouard Geoffrois, Results of the RIMES Evaluation Campaign for Handwritten Mail Processing, Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), 2009, pp. 941–945.

[6] Ehsan Yarshater, Persian Language and Literature, Encyclopædia Iranica, Online Edition (2000).

[7] H. Khosravi and E. Kabir, Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, Pattern Recognition Letters, vol. 28, no. 10, pp. 1133-1141, 2007.

[8] Saeed Mozaffari, Haikal El Abed, Volker Märgner, Karim Faez, Seyed Ali Amirshahi IFN/Farsi-Database: A Database of Farsi Handwritten City Names, Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2008.

[9] Mario Pechwitz, Samia Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, and Hamid Amiri, IFN/ENIT: A Database of Handwritten Arabic Words, Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), 2003, pp. 819–823.

[10] Javad Sadri, Mohammad Reza Yeganehzad, and Javad Saghi, A novel comprehensive database for offline Persian handwriting recognition, Pattern Recognition, 60 (2016) 378–393.

[11] Pourya Jafarzadeh, Padideh Choobdar, and Vahid Mohammadi Safarzadeh, Khayyam Offline Persian Handwriting Dataset, arXiv preprint, arXiv:2406.01025, 2024.

[12] Majid Ziaratban, Karim Faez, and Fatemeh Bagheri, FHT: An Unconstraint Farsi Handwritten Text Database, Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), 2009, pp. 281–285.

[13] H. Mohammed and M. Jampour, From Detection to Modeling: An End-to-End Paleographic System for Analysing Historical Handwriting Styles, in Document Analysis Systems (DAS), 2024.

[14] M. Javidi and M. Jampour, A deep learning framework for text-independent writer identification, Engineering Applications of Artificial Intelligence, vol. 89, pp. 103466, 2020.

[15] M. Jampour, H. Mohammed, and J. Gippert, Synthetic MSI Images of Georgian Palimpsests (SGP Dataset), Dataset, doi: 10.25592/uhhfdm.13378, 2023.

[16] A. Zanganeh, M. Jampour, and K. Layeghi, IAUFD: A 100k images dataset for automatic football image/video analysis, IET Image Processing, vol. 16, no. 11, pp. 2750–2763, 2022.

[17] Sabri A. Mahmoud, Irfan Ahmad, Wasfi G. Al-Khatib, Mohammad Alshayeb, Mohammad Tanvir Parvez, Volker Märgner, and Gernot A. Fink, KHATT: An open Arabic offline handwritten text database, Pattern Recognition, 47 (2014) 1096–1112.

[18] Anis Ismail, Zena Kamel, and Reem Mahmoud, HICMA: The Handwriting Identification for Calligraphy and Manuscripts in Arabic Dataset, Proceedings of ArabicNLP 2023, 2023, pp. 24–32.

[19] M. Al-Ohali, M. Cheriet, and C. Suen, Databases for Arabic handwritten text recognition research, in Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWHR), pp. 350–354, 2002.

[20] Jonathan Parkes Allen, John Mullan, Lorenz Nigst, et. al Open-ITI MAKHZAN: An Open Annotated Dataset of Arabic, Persian, Ottoman Turkish, and Urdu Print and Manuscript Data, Journal of Open Humanities Data, 12 (2026) Article 69.

[21] ul Sehr Zia, N., Naeem, M.F., Raza, S.M.K. et al. A Convolutional Recursive Deep Architecture for Unconstrained Urdu Handwriting Recognition, Neural Computing and Applications, 34 (2022) 1635–1648

[22] Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig, CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 560–564.

[23] Chawki Djeddi, Abdeljalil Gattal, Labiba Souici-Meslati, Imran Siddiqi, Youcef Chibani, and Haikal El Abed, LAMIS-MSHD: A Multi-script Offline Handwriting Database, Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2014, pp. 93–97.

[24] Puntis Jifroodian Haghighi, Nicola Nobile, Chun Lei He, and Ching Y. Suen, A New Large-Scale Multi-purpose Handwritten Farsi Database, Proceedings of the 6th International Conference on Image Analysis and Recognition (ICIAR), 2009, pp. 278–286.

[25] Shahbaz Hassan, Ahmad Raza Shahid, and Muhammad Asif Naeem, TDA-ViT: A Transformer-Based Framework for Unified Urdu Text Recognition via Topological and Visual Feature Fusion, IEEE Access, 13 (2025) 182940–182959.

[26] Saad Bin Ahmed, Saeeda Naz, Salahuddin Swati, Muhammad Imran Razzak, Arif Iqbal Umar, and Akbar Ali Khan, UCOM Offline Dataset — An Urdu Handwritten Dataset Generation, International Arab Journal of Information Technology, 14 (2017).

[27] A. Raza, A. Siddique, A. Batool, R. Sirhindi, T. Iftikhar, and N. Khan, An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation, Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012, pp. 491–496.

[28] Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R. Munafò, George Ocklenburg, and Silvia Paracchini, Human handedness: A meta-analysis, Psychological Bulletin, 146 (2020) 481–524.

[29] B. Shi, X. Bai, and C. Yao, An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298–2304, 2017.

[30] Y. Li, D. Chen, T. Tang, and X. Shen, HTR-VT: Handwritten text recognition with vision transformer, Pattern Recognition, vol. 158, pp. 110967, 2025.

[31] S.B. Needleman and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology, vol. 48, no. 3, pp. 443-453, 1970.

[32] Jaided AI, EasyOCR: Ready-to-use OCR with 80+ supported languages, GitHub Repository, 2024. https://github.com/JaidedAI/EasyOCR

[33] David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60 (2004) 91–110.

[34] N. Stamatopoulos, G. Louloudis, and B. Gatos, Overview of the ICDAR 2013 Handwriting Segmentation Contest, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1404–1408.

[35] G. Louloudis, N. Stamatopoulos, and B. Gatos, ICDAR 2013 Handwriting Segmentation Contest, Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1402–1403.

[36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, 13 (2004) 600–612.

[37] M. Jampour, Character Localization in Degraded Historical Documents via Heatmap-Guided UNet3+ with Application on Palimpsests, in 2025 12th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2025.

[38] Y. Zamani, Y. Souri, H. Rashidi, and S. Kasaei, Persian handwritten digit recognition by random forest and convolutional neural networks, Proceedings of the 2015 Iranian Conference on Machine Vision and Image Processing (MVIP), 2015.

[39] R. Ebrahimzadeh and M. Jampour, Efficient Handwritten Digit Recognition based on Histogram of Oriented Gradients and SVM, International Journal of Computer Applications, vol. 90, no. 15, pp. 1–6, 2014.

[40] M. Jampour, S. Abbaasi, and M. Javidi, CapsNet Regularization and its Conjugation with ResNet for Signature Identification, Pattern Recognition, vol. 120, pp. 108133, 2021.