

---

# Supplementary file for "HHD-Ethiopic: A Historical Handwritten Dataset for Ethiopic OCR"

---

**Birhanu Hailu Belay\*** **Isabelle Guyon<sup>+,‡</sup>** **Tadele Mengiste<sup>†</sup>** **Bezawork Tilahun<sup>†</sup>**  
**Macus Liwicki<sup>||</sup>** **Tesfa Tegegne<sup>†</sup>** **Romain Egele\*** **Tsiyon Worku<sup>†</sup>**  
\*LISN, Université Paris-Saclay, France + Google Brain, USA ‡ChaLearn, USA  
†Bahir Dar Institute of Technology, Ethiopia ||Luleå University of Technology, Sweden  
birhanu-hailu.belay@upsaclay.fr  
guyon@chalearn.org, macus.liwicki@ltu.se, romain.egele@inria.fr  
{tadele.mengiste, bezawork.tilahun, tesfa.tegegne, tsiyon.worku}@bdu.edu.et

## 1 Appendix

2 This appendix comprises three sections: dataset documentation, Ethiopic writing system, and dataset  
3 preparation and baseline model training details. The dataset documentation outlines its composition,  
4 preprocessing steps, recommended use-case of distribution of the HHD-Ethiopic dataset, and author  
5 statement. The Ethiopic writing system section explores its historical significance and script structure.  
6 Lastly, the dataset and baseline training process section offers insights into dataset preparation and  
7 baseline model training strategies.

## 8 A Dataset documentation for HHD-Ethiopic

9 To prepare this dataset documentation, we use a datasheet [11] for dataset guideline. This docu-  
10 mentation consists of the motivation behind the dataset, its composition, the process of collection,  
11 recommended use cases, as well as information on processing, cleaning, labeling, distribution  
12 (including hosting, licensing), and maintenance. This documentation also includes author statements.

### 13 A.1 Motivation

14 **For what purpose was the dataset created?** Was there a specific task in mind? Was  
15 there a specific gap that needed to be filled? Please provide a description.

16 The dataset targets the challenges of the indigenous Ethiopic script, addressing its scarcity of  
17 resources. It serves as a valuable asset for researchers and developers, facilitating advancements in  
18 OCR technology specifically for historical handwritten Ethiopic recognition. Unlike well-studied  
19 scripts like Latin, it bridges the gap and enables accurate recognition of Ethiopic text in historical  
20 documents using machine learning approaches.

21 **Who created this dataset** (e.g., which team, research group) and on behalf of which entity  
22 (e.g., company, institution, organization)?

23 HHD-Ethiopic dataset is created primarily by the LISN lab at University of Paris-Saclay and ICT4D  
24 research center at Bahir Dar Institute of Technology, in collaboration with other researchers from  
25 Lulea Technology University.

26 **Who funded the creation of the dataset?** If there is an associated grant, please provide  
27 the name of the grantor and the grant name and number.

28 The dataset creation received funding from ChaLearn and the ICT4D research center of Bahir Dar  
29 Institute of Technology. Findings, and/or recommendations expressed in this material are solely those  
30 of the author/s and do not necessarily represent the views of ChaLearn or ICT4D.

31 **Any other comments?** No.

## 32 **A.2 Composition**

33 **What do the instances that comprise the dataset represent (e.g., documents, photos,**  
34 **people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings;  
35 people and interactions between them; nodes and edges)? Please provide a description.

36 The HHD-Ethiopic dataset is an OCR dataset that consists of text-line images extracted from  
37 historical handwritten Ethiopic manuscript and there corresponding ground truths text, sample images  
38 and their corresponding ground truth texts are shown Figure 14

39 **How many instances are there in total (of each type, if appropriate)?**

40 The HHD-Ethiopic dataset comprises 79,684 text-line images accompanied by their respective ground-  
41 truth texts. These images are extracted from a collection of 1,746 pages of Ethiopic manuscripts  
42 dating from the 18<sup>th</sup> to the 20<sup>th</sup> centuries. The dataset is divided into a training set, containing 57,374  
43 text-line images, and two distinct Test sets. One Test set, that consists 6,375 images, is randomly  
44 sampled from the training set, while the other is exclusively prepared from the 18<sup>th</sup> century Ethiopic  
45 manuscripts and includes about 15,935 text-line images along with their corresponding ground-truth  
46 texts ( details are provided in the main paper).

47 **Does the dataset contain all possible instances or is it a sample (not necessarily**  
48 **random) of instances from a larger set?** If the dataset is a sample, then what is the  
49 larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so,  
50 please describe how this representativeness was validated/verified. If it is not representative  
51 of the larger set, please describe why not (e.g., to cover a more diverse range of instances,  
52 because instances were withheld or unavailable).

53 HHD-Ethiopic, is a historical handwritten dataset between the 18<sup>th</sup> and 20<sup>th</sup> centuries. It is a sample  
54 of instances from that time period and includes 306 out of 317 frequently used characters in the  
55 Ethiopian writing system.

56 **What data does each instance consist of? “Raw” data (e.g., unprocessed text or**  
57 **images) or features?** In either case, please provide a description.

58 Each instance in the training set consists of text-line images and their corresponding ground-truth  
59 text. The test set, on the other hand, includes raw human-level prediction texts from 13 independent  
60 annotators which we use as a baseline to compare the human-level performance with OCR models in  
61 this paper

62 **Is there a label or target associated with each instance?** If so, please provide a  
63 description.

64 Yes, there is a ground-truth text for each text-line image.

65 **Is any information missing from individual instances?** If so, please provide a description,  
66 explaining why this information is missing (e.g., because it was unavailable). This does not  
67 include intentionally removed information, but might include, e.g., redacted text.

68 No, everything is included.

69 **Are relationships between individual instances made explicit (e.g., users' movie**  
70 **ratings, social network links)?** If so, please describe how these relationships are made  
71 explicit.

72 The relationships between individual instances in the text-line image dataset are not explicitly  
73 defined, as each image is formed from a sampled set of 306 Ethiopic characters rather it may have  
74 indirect/inferred connection.

75 **Are there recommended data splits (e.g., training, development/validation, testing)?** If  
76 so, please provide a description of these splits, explaining the rationale behind them.

77 The HHD-Ethiopic dataset is split into first into training, and testing. The training set includes  
78 text-line images from the 19<sup>th</sup> and 20<sup>th</sup> centuries. A validation set is then randomly sampled as 10%  
79 of the training set. Two test sets are propose: the first testing set consists of 6,375 images randomly  
80 selected from a similar distribution as the training set. The second testing set contains 15,935 images  
81 from a different distribution, representing 18<sup>th</sup> century manuscripts. The first test evaluates baseline  
82 performance in an IID setting, while the second test assesses performance in an OOD scenario. The  
83 detail statistic is provided in section 3 of the main paper.

84 **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please  
85 provide a description.

86 While the ground-truth text was double-checked by a supervisor for each annotator, we recommend  
87 additional revision of the the ground-truth texts by multiple historical document experts to minimize  
88 annotation errors.

89 **Is the dataset self-contained, or does it link to or otherwise rely on external resources**  
90 **(e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a)  
91 are there guarantees that they will exist, and remain constant, over time; b) are there official  
92 archival versions of the complete dataset (i.e., including the external resources as they  
93 existed at the time the dataset was created); c) are there any restrictions (e.g., licenses,  
94 fees) associated with any of the external resources that might apply to a future user? Please  
95 provide descriptions of all external resources and any restrictions associated with them, as  
96 well as links or other access points, as appropriate.

97 The dataset is entirely self-contained. It will exist, and remain constant, over time once we release it.  
98

99 **Does the dataset contain data that might be considered confidential (e.g., data that is**  
100 **protected by legal privilege or by doctor-patient confidentiality, data that includes the**  
101 **content of individuals non-public communications)?** If so, please provide a description.

102 No.

103 **Does the dataset contain data that, if viewed directly, might be offensive, insulting,**  
104 **threatening, or might otherwise cause anxiety?** If so, please describe why.

105 No.

106 **Does the dataset relate to people?** If not, you may skip the remaining questions in this  
107 section.

108 No.

109 **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please  
110 describe how these subpopulations are identified and provide a description of their respective  
111 distributions within the dataset.

112 No.

113 **Is it possible to identify individuals (i.e., one or more natural persons), either directly**  
114 **or indirectly (i.e., in combination with other data) from the dataset?** If so, please  
115 describe how.

116 No.

117 **Does the dataset contain data that might be considered sensitive in any way (e.g., data**  
118 **that reveals racial or ethnic origins, sexual orientations, religious beliefs, political**  
119 **opinions or union memberships, or locations; financial or health data; biometric or**  
120 **genetic data; forms of government identification, such as social security numbers;**  
121 **criminal history)?** If so, please provide a description.

122 No.

123 **Any other comments?** No.

### 124 **A.3 Collection Process**

125 **How was the data associated with each instance acquired?** Was the data directly  
126 observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or  
127 indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses  
128 for age or language)? If data was reported by subjects or indirectly inferred/derived from  
129 other data, was the data validated/verified? If so, please describe how.

130 The historical Ethiopic manuscripts were solely collected from Ethiopian national Archive and  
131 Library Agency (ENALA). Each instance is an image/scanned version of documents and is directly  
132 observable (see the main paper from section 3).

133 **What mechanisms or procedures were used to collect the data (e.g., hardware appa-**  
134  **ratus or sensor, manual human curation, software program, software API)?** How were  
135 these mechanisms or procedures validated?

136 After obtaining the scanned copy of the manuscript from ENALA and extracting the text-image lines,  
137 we hire individuals to annotate each text-line image. During the annotation process, all annotators  
138 have the freedom to refer to any external sources. for annotation purpose, annotation, we develop an  
139 offline tool that can be easily installed on each user's machine (see Figure 13).

140 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g.,**  
141 **deterministic, probabilistic with specific sampling probabilities)?**

142 The historical documents were collected from ENALA. While we did not have the authority to select  
143 specific documents, the workers randomly select pages, taking into account our request and the need  
144 to maintain the confidentiality of the book's information.

145 **Who was involved in the data collection process (e.g., students, crowdworkers,**  
146 **contractors) and how were they compensated (e.g., how much were crowdworkers**  
147 **paid)?**

148 the participants were students and staff members and for the raw manuscript collection and digitiza-  
149 tion we paid money as a compensation.

150 **Over what timeframe was the data collected? Does this timeframe match the creation**  
151 **timeframe of the data associated with the instances (e.g., recent crawl of old news**  
152 **articles)?** If not, please describe the timeframe in which the data associated with the  
153 instances was created.

154 The dataset was collected in March-May 2022 and the complete data creation (including preprocess-  
155 ing, annotation and verification were done from September 2022-February 2023.

156 **Were any ethical review processes conducted (e.g., by an institutional review board)?**

157 If so, please provide a description of these review processes, including the outcomes, as  
158 well as a link or other access point to any supporting documentation.

159 No.

160 **Does the dataset relate to people?** If not, you may skip the remaining questions in this  
161 section.

162 No.

163 **Did you collect the data from the individuals in question directly, or obtain it via third  
164 parties or other sources (e.g., websites)?**

165 As described section 3 of the main paper, the data was collected from ENALA directly.

166 **Were the individuals in question notified about the data collection?** If so, please  
167 describe (or show with screenshots or other information) how notice was provided, and  
168 provide a link or other access point to, or otherwise reproduce, the exact language of the  
169 notification itself.

170 Yes, the scanned copies of document images were collected directly from ENALA. This request was  
171 made in person along with a letter, which also explained the objectives, goals, and the need for data  
172 in our work.

173 **Did the individuals in question consent to the collection and use of their data?** If so,  
174 please describe (or show with screenshots or other information) how consent was requested  
175 and provided, and provide a link or other access point to, or otherwise reproduce, the exact  
176 language to which the individuals consented.

177 Yes, once we met with the staff at ENALA and explained the goals of our project, they agreed to  
178 provide the data and arranged a way for delivering the documents.

179 **If consent was obtained, were the consenting individuals provided with a mechanism  
180 to revoke their consent in the future or for certain uses?** If so, please provide a  
181 description, as well as a link or other access point to the mechanism (if appropriate).

182 No.

183 **Has an analysis of the potential impact of the dataset and its use on data subjects  
184 (e.g., a data protection impact analysis) been conducted?** If so, please provide a  
185 description of this analysis, including the outcomes, as well as a link or other access point  
186 to any supporting documentation.

187 No.

#### 188 **A.4 Preprocessing/cleaning/labeling**

189 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or  
190 bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of  
191 instances, processing of missing values)?** If so, please provide a description. If not, you  
192 may skip the remainder of the questions in this section.

193 Yes, preprocessing tasks such as image segmentation and the removal of non-Ethiopic characters  
194 were performed. Furthermore, alignments between the images and their corresponding text-line  
195 images were double-checked for each submission by the annotators and verified by a reviewer.

196 **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g.,**  
197 **to support unanticipated future uses)?** If so, please provide a link or other access point  
198 to the “raw” data.

199 No.

200 **Is the software used to preprocess/clean/label the instances available?** If so, please  
201 provide a link or other access point.

202 Yes, here is the link for the labeling tool that we developed with the aim of fitting and making it easier  
203 for the target annotators. It is designed to accommodate their operating systems and internet service  
204 settings, allowing them to work offline when there is no internet connection. You can access the  
205 tool at this link: [https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/labeling\\_](https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/labeling_tool)  
206 [tool](https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/labeling_tool). For preprocessing tasks, including column detection, binarization, and text-line segmentation,  
207 we utilize the OCRopus framework. You can find more information about the framework and its  
208 functionalities on their GitHub page: <https://github.com/ocropus/ocropy>

## 209 **A.5 Uses**

210 **Has the dataset been used for any tasks already?** If so, please provide a description.

211 HHD-Ethiopic is a new historical handwritten Ethiopic OCR dataset for a text-line image recognition.  
212 In this work we evaluate several state-of-the-art deep learning models and an independent human-level  
213 recognition performance on a dataset, which involves comparing the performance of several human  
214 annotators with the performance of machine models. The human-level performance serves as a  
215 benchmark and in turn it also contribute to the uniqueness and quality of the dataset.

216 **Is there a repository that links to any or all papers or systems that use the dataset?** If  
217 so, please provide a link or other access point.

218 Yes, we release our dataset, code, baseline models and human-level performances at [https://](https://github.com/bdu-birhanu/HHD-Ethiopic)  
219 [github.com/bdu-birhanu/HHD-Ethiopic](https://github.com/bdu-birhanu/HHD-Ethiopic).

220 **What (other) tasks could the dataset be used for?**

221 The HHD-Ethiopic dataset was specifically created to address the gap in Historical handwritten  
222 Ethiopic manuscript recognition. However, it can also be utilized to benchmark the performance of  
223 machine learning models for other scripts.

224 **Is there anything about the composition of the dataset or the way it was collected**  
225 **and preprocessed/cleaned/labeled that might impact future uses?** For example, is  
226 there anything that a future user might need to know to avoid uses that could result in unfair  
227 treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other  
228 undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is  
229 there anything a future user could do to mitigate these undesirable harms?

230 The datasets can be used without further considerations.

231 **Are there tasks for which the dataset should not be used?** If so, please provide a  
232 description.

233 No.

234 **Any other comments?** No.

235 **A.6 Distribution**

236 **Will the dataset be distributed to third parties outside of the entity (e.g., company,**  
237 **institution, organization) on behalf of which the dataset was created?** If so, please  
238 provide a description.

239 Yes, both the dataset and baseline results will be made available to the public research community for  
240 experimentation and further work on historical handwritten recognition.

241 **How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the**  
242 **dataset have a digital object identifier (DOI)?**

243 The HHD-Ethiopic dataset can be downloaded from [https://github.com/bdu-birhanu/](https://github.com/bdu-birhanu/HHD-Ethiopic)  
244 [HHD-Ethiopic](https://github.com/bdu-birhanu/HHD-Ethiopic) or directly for the Huggingface [https://huggingface.co/datasets/](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic)  
245 [OCR-Ethiopic/HHD-Ethiopic](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic). The images can be downloaded as a zipped file. The digital  
246 object identifie (DOI) of the dataset is: doi:10.57967/hf/0691. Our dataset has also been made public  
247 on Zenodo.org. However, we have chosen to provide it on Hugging Face and GitHub as well, as  
248 we believe these platforms are commonly used within the document image analysis and machine  
249 learning community.

250 **When will the dataset be distributed?**

251 The dataset is currently available for use in our repository.

252 **Will the dataset be distributed under a copyright or other intellectual property (IP)**  
253 **license, and/or under applicable terms of use (ToU)?** If so, please describe this license  
254 and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant  
255 licensing terms or ToU, as well as any fees associated with these restrictions. This  
256 work is licensed under a [CC-BY-4.0 International License](https://creativecommons.org/licenses/by/4.0/) and available at: [https://github.com/](https://github.com/bdu-birhanu/HHD-Ethiopic)  
257 [bdu-birhanu/HHD-Ethiopic](https://github.com/bdu-birhanu/HHD-Ethiopic) or can be directly downloaded from [https://huggingface.co/](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic)  
258 [datasets/OCR-Ethiopic/HHD-Ethiopic](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic)

259 **Have any third parties imposed IP-based or other restrictions on the data associated**  
260 **with the instances?** If so, please describe these restrictions, and provide a link or other  
261 access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees  
262 associated with these restrictions.

263 No.

264 **Do any export controls or other regulatory restrictions apply to the dataset or to**  
265 **individual instances?** If so, please describe these restrictions, and provide a link or other  
266 access point to, or otherwise reproduce, any supporting documentation.

267 **A.7 Maintenance**

268 **Who will be supporting/hosting/maintaining the dataset?**

269 The authors of this paper are responsible for supporting the datasets.

270 **How can the owner/curator/manager of the dataset be contacted (e.g., email ad-**  
271 **dress)?**

272 The curators of the dataset can be contacted via email and we provide it in the repository <https://github.com/bdu-birhanu/HHD-Ethiopic>  
273 [/bdu-birhanu/HHD-Ethiopic](https://github.com/bdu-birhanu/HHD-Ethiopic)

274 **Is there an erratum?** If so, please provide a link or other access point.

275 There is no an explicit erratum.

276 **Will the dataset be updated (e.g., to correct labeling errors, add new instances,**  
277 **delete instances)?** If so, please describe how often, by whom, and how updates will be  
278 communicated to users (e.g., mailing list, GitHub)?

279 Yes, we have plans to add more data to the dataset. As updates are made, we will ensure that both the  
280 documentation and our repository are updated accordingly.

281 **If the dataset relates to people, are there applicable limits on the retention of the data**  
282 **associated with the instances (e.g., were individuals in question told that their data**  
283 **would be retained for a fixed period of time and then deleted)?** If so, please describe  
284 these limits and explain how they will be enforced.

285 No.

286 **Will older versions of the dataset continue to be supported/hosted/maintained?** If so,  
287 please describe how. If not, please describe how its obsolescence will be communicated to  
288 users.

289 Any changes made to the dataset will ensure that the original version remains available, and  
290 subsequent versions, such as HHD-Ethiopic-1.1, will be released with documentation.

291 **If others want to extend/augment/build on/contribute to the dataset, is there a mech-**  
292 **anism for them to do so?** If so, please provide a description. Will these contributions  
293 be validated/verified? If so, please describe how. If not, why not? Is there a process  
294 for communicating/distributing these contributions to other users? If so, please provide a  
295 description.

296 Yes, users can contribute to the dataset and can contact the original authors about incorporating  
297 fixes/extensions. This is encouraged. Users are free to extend or augment the dataset for their  
298 purposes.

299 **Any other comments?** None.

## 300 **A.8 Accessibility**

- 301 1. Links to access the **dataset** and its **metadata** and **code and simulation environment**.  
302 <https://github.com/bdu-birhanu/HHD-Ethiopic>
- 303 2. **Data format:** we follow widely used data formats in OCR dataset. The actual text-line  
304 images are stored in .png format while ground-truth texts are in .txt. the image-ground truth  
305 pair are given in .CSV formats, in addition, the images and their corresponding ground-truth  
306 are also stored in numpy format. An example of the dataset structure can be found in the  
307 README.md file of our dataset repository.
- 308 3. **Long-term preservation:** we the authors are responsible to maintain and ensure consistency  
309 of the data and it will be in our GitHub repository.
- 310 4. **Explicit license:** The dataset is licensed under a [CC-BY-4.0](#) and the source code is under  
311 MIT license <https://github.com/bdu-bf/HHD-Ethiopic>
- 312 5. **A persistent dereferenceable identifier:** A DOI from Hugging Face, doi:10.57967/hf/0691

## 313 **A.9 Author statement**

314 The authors have conducted a thorough review of the information presented in this document. To the  
315 best of our knowledge, the datasets included in HHD-Ethiopic are intended for research purposes  
316 and should be used in accordance with the described methodology and licenses outlined in the  
317 Accessibility section. It is important to note that the authors assume full responsibility in the event of  
318 any violation of rights.



319 **B Ethiopic writing systems**

320 Ethiopic script is an ancient writing system used primarily in Ethiopia and Eritrea. With its origins  
321 dating back to the 4<sup>th</sup> century AD [13]. The script is characterised by its unique syllabic structure,  
322 which combines consonants and vowels to form complex characters. In literature the Ethiopic writing  
323 system also named with various names including "Abugida", "Amharic", "Ge'ez", and "Fidel".

324 Ethiopic script has been a significant cultural and linguistic heritage of the region, playing a vital  
325 role in preserving the rich history and traditions of Ethiopia. It is primarily used for writing over 27  
326 languages including the Amharic and Tigrinya languages, among others. As depicted in Figure 7, the  
327 script has a distinct visual appearance, characterized by its curved and geometric shapes, making it  
visually distinctive and is written and read, as English, from left to right and top to down [5].



Figure 7: Sample historical handwritten Ethiopic manuscripts

328

329 Despite the long history of the Ethiopic script, it has encountered numerous challenges in the digital  
330 world due to its low-resource nature [7, 16]. Issues such as limited digitized fonts, linguistic tools,  
331 and datasets have posed obstacles in the fields of natural language processing and document image  
332 analysis technologies.

333 The Ethiopic script poses unique challenges for machine learning due to the scarcity of available  
334 resources. This script is characterized by its complex orthographic identities and visually similar char-  
335 acters. Comprising over 317 distinct characters, including approximately 280 characters organized in  
336 a 2D matrix format known as Fidel-Gebeta (Figure 8), along with 20 digits and 8 punctuation marks  
337 (Figure 9).

338 As depicted in Figure 8, the Ethiopic script consists of 34 consonant characters, which serve as  
339 the base for deriving additional characters using diacritics. These diacritics can be found as small  
340 marks placed on the top, bottom, left, or right sides of the base character. Furthermore, specific  
341 vowel characters are formed by shortening either the left or right leg of consonant characters, as  
342 demonstrated in columns 4 (shortening left leg) and 7 (shortening right leg) of the fidel-Gebeta. The  
343 vowels, derived from these consonants, span from 1 to 12 and correspond to the respective columns.

344 For example, in the second row of the fidel-Gebeta, the consonant character  $\Lambda$  represents the sound  
345 "le" in Ethiopic. From this base character, various vowel characters emerge, such as:

- 346 •  $\Lambda$  is formed by adding a horizontal diacritic at the middle left side of the base character and  
347 represents the sound "lu".
- 348 •  $\Lambda$  is formed by adding a horizontal diacritic at the bottom left leg of the base character and  
349 represents the sound "li".
- 350 •  $\Lambda$  is formed by shortening the left leg of the base character and represents the sound "la".

351 These examples showcase the versatility of the Ethiopic script, where modifying the diacritics or leg  
352 lengths of consonant characters allows for the representation of different vowel sounds.

		1	2	3	4	5	6	7	8	9	10	11	12
		ä/e	u	i	a	ē	ə	o	ʷä/ue	ʷi/u	ʷa/ua	ʷē/uē	ʷə
1	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
2	l	ለ	ሉ	ሊ	ላ	ሌ	ሎ	ሎ			ሊ		
3	h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ			ሐ		
4	m	ሙ	ሚ	ሚ	ሚ	ሚ	ሚ	ሚ			ሚ		
5	s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ			ሠ		
6	r	ረ	ሩ	ሪ	ራ	ራ	ሪ	ሪ			ረ		
7	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሰ		
8	s	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሸ		
9	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቇ	ቈ	቉	ቊ	ቋ
10	b	በ	ቡ	ቢ	ባ	ቤ	ቦ	ቧ			በ		
11	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቯ			ቨ		
12	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ተ		
13	č	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቸ		
14	ḥ	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	ሑ	ሕ	ሐ	ሑ	ሕ
15	n	ነ	ቡ	ኒ	ና	ኔ	ነ	ኖ			ነ		
16	n̄	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኘ		
17	ʔ	አ	ኡ	ኢ	ኣ	ኤ	ኦ	ኰ			አ		
18	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	኱	ኲ	ኳ	ኴ
19	x	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	኿	ኻ	ኼ	ኽ	ኾ
20	w	ወ	ዐ	ዑ	ዓ	ዔ	ዕ	ዖ					
21	ʔ	ዐ	ዑ	ዓ	ዔ	ዕ	ዖ	ዘ					
22	z	ዘ	የ	ዪ	ዩ	ያ	ዴ	ድ			ዘ		
23	ž	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዠ		
24	y	የ	ይ	ይ	የ	ይ	ይ	የ					
25	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ደ		
26	ḡ	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ			ጀ		
27	g	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ
28	ḥ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጠ		
29	č	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጠ		
30	p	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ			ጸ		
31	s	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ			ጸ		
32	š	ፀ	ፀ	ፀ	ፀ	ፀ	ፀ	ፀ					
33	f	ፈ	ፈ	ፈ	ፈ	ፈ	ፈ	ፈ			ፈ		
34	ḥ	ፒ	ፒ	ፒ	ፒ	ፒ	ፒ	ፒ			ፒ		

Figure 8: Fidel-Gebeta: the row-column matrix structure of Ethiopic characters. The first column shows the consonants, while the following columns (1-12) illustrate syllabic variations (obtained by adding diacritics or modifying parts of the consonant).

353 Ethiopic numerals also called Ge’ez numerals, are a numeric system traditionally used in Ethiopic  
 354 writing. These numeral system has its own distinct symbols for representing numbers, which are  
 355 different from the Arabic or Roman numerals commonly used in many other parts of the world. The  
 356 system has a base of 10, with unique characters for each digit from 1 to 9, as well as special symbols  
 357 for tens, hundreds, and thousands (Figure 9). For example:

- 358 • Ethiopic symbol ፩ is similar to the Arabic numeral 1.
- 359 • symbol ፪፫ is similar to the Arabic numeral 44.
- 360 • symbol ፬፭፮፯፰፱፺፻፰፱ similar to the Arabic numeral 99999.
- 361 • symbol ፷፲፱ is similar to the Arabic numeral 10002.
- 362 • symbol ፲፯፱፻፷፫፱ similar to the Arabic numeral 1233.

363 Though modern Arabic numerals dominate daily life and official documents, understanding Ethiopic  
 364 numerals is vital for deciphering historical texts and preserving cultural heritage.

365 In the Ethiopic writing system, punctuation marks convey meaning and guide text interpretation  
 366 (see Figure 9). Understanding their usage is vital for clear and effective written communication in  
 367 Ethiopic script.

C	፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
1	2	3	4	5	6	7	8	9	10	
፳	፴	፵	፶	፷	፸	፹	፺	፻	፼	፽
20	30	40	50	60	70	80	90	100	10000	

a

⌘	፥	፦	፧	፨	፩	፪	፫
section mark	word separator	full stop (period)	comma	semicolon	colon	question mark	paragraph separator

b

Figure 9: Numbering system (a) and punctuation marks (b) in Ethiopic script

368 The complexities of symbols within the Ethiopic script present significant challenges for machine  
 369 learning tasks, requiring attentive approaches to achieve accurate recognition and analysis. An  
 370 example of these challenges is the non-standardized usage of punctuation marks 10 and variations in  
 371 writing styles, as depicted in Figure 7. These factors contribute to the difficulties encountered in the  
 372 development of Ethiopic OCR systems.

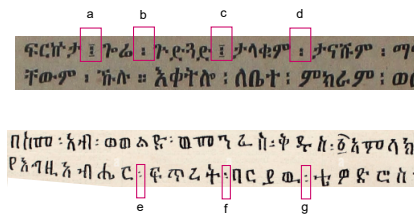


Figure 10: Examples of punctuation usage and writing Styles: As shown by the red rectangle and labeled by [a, b, c, d], there is typically a space before and after the punctuation mark. In contrast, the punctuation marks labeled by [e, f, g] do not have any space before or after them. The punctuation marks labeled by a and c serve as list separators and are distinct from the other punctuation marks, which are used as word separators.

## 373 C Methods and implementation details

374 In this section, we provide additional details of models implemented and evaluated on our HHD-  
 375 Ethiopic OCR dataset. We evaluate several state-of-the-art methods, which can be broadly grouped  
 376 as CTC-based, Attention, and Transformer-based. However, our primary focus in this section is on  
 377 the CTC-based model, which is designed to operate effectively in lower resource settings. This is  
 378 because the other CTC, attention and Transformer-based model (evaluated on this new datasets) are  
 379 validated from previous works [9, 10, 14, 17, 18] and involves extensive hyperparameters, making  
 380 it more suited for higher-resource environments. These SOTA methods are implemented using the  
 381 open-source toolbox, mmocr: <https://github.com/open-mmlab/mmocr>.

### 382 C.1 Baseline models

383 The implementation of the CTC-based model follows a typical pipeline depicted in Figure 11. In case  
 384 of Plain-CTC, initially, the preprocessed images are passed through a convolutional neural network  
 385 (CNN) backbone, which extracts relevant image features using a series of convolutional and pooling  
 386 layers.

387 The output features from the CNN backbone are reshaped and subsequently fed into a Long Short-  
 388 term Memory (LSTM) network with connectionist temporal classification (CTC) network. This  
 389 combination enables the model to effectively capture the temporal dependencies between the image  
 390 features and the corresponding text labels. The RNN layer incorporates two Bi-directional LSTM  
 391 units to learn sequential patterns and generate a  $[(c + 1) \times T]$  matrix of Softmax probabilities for

each character at each time-step, where  $c$  and  $T$  denote the number of characters and the length of maximum time-step. Finally, a the CTC converts the intermediate representations into the final output text predictions.

The alternative CTC-based approach, referred to as Attn-CTC within this paper and previously introduced for Amharic text recognition[6], extends the Plain-CTC methodology by incorporating an attention mechanism into the CTC layers. The rationale behind incorporating the attention layer lies in leveraging its capacity to derive a more potent hidden representation through a weighted contextual vector. This model comprises a combination of CNN and LSTM as the encoding module. The output of this module feeds into the attention module, and subsequently, the decoded output string is obtained through the CTC layer.

During training, the CTC algorithm calculates the likelihood of the output sequence given the input sequence and uses it as the objective function [12, 15]. The training process maximizes this likelihood, which, in turn, maximizes the probability of the correct output sequence. The loss that is minimized during training is the negative of this likelihood, which can be defined as:

$$CTC_{loss} = -\log \sum_{(y,x) \in S} p(y/x) \quad (1)$$

where  $x$  and  $y$  denote pair of input and output sequences in sample dataset  $S$  respectively and the probability of label sequence for a single pair  $p(y/x)$  is computed by multiplying the probability of labels along a specific path  $\pi$  for the overall time steps  $T$  and it can be defined as:

$$P(y/x) = \prod_{t=1}^T p(a_t, \pi) \quad (2)$$

where  $a$  is a character in the specified path and  $p(a)$  is its probability on each time-step on that path.

Once training and evaluating the OCR model with network settings proposed in [4, 6], we employed Bayesian optimization for the selection of hyperparameters, with the CTC validation loss serving as the criteria for optimization. Bayesian optimization captures the relationship between the hyperparameters and the CTC validation loss, iteratively updating and refining the model as it explores different hyperparameter configurations ( see ref [3]for details) that yields lower CTC validation loss values. This approach allowed us to effectively tune our model and enhance its performance, contributing to the overall success of our text-image recognition model.

The source code for hyperparameter selection and training procedures are provided at <https://github.com/bdu-bf/HHD-Ethiopic>.

The recognition performance of all human-level and baseline models evaluated in this work is reported using the character error rate (CER) and Normalized Edit Distance (NED) metrics. All results reported with these two metrics are converted to 100%. The CER metric can be computed as follows,

$$CER(T, P) = \left( \frac{1}{c} \sum_{m \in T, n \in P} ED(m, n) \right) \times 100, \quad (3)$$

where  $c$  denotes the total number of characters in the ground-truth,  $t$  and  $p$  denote the ground-truth labels and predicted respectively, and  $ED(m, n)$  is the Levenshtein edit-distance between sequences  $m$  and  $n$ .

while the NED metric is computed as:

$$NED = \left( \frac{1}{N} \sum_{i=1}^N \frac{ED(m_i, n_i)}{\max(l_i, \hat{l}_i)} \right) \times 100 \quad (4)$$

<sup>6</sup><https://deephper.readthedocs.io/en/latest/index.html>

427 where  $N$  is the maximum number of paired ground truth and prediction strings,  $ED$  is the Levenshtein  
 428 edit distance,  $m_i$  and  $n_i$  denote the predicted text and the corresponding ground truth (GT) string,  
 429 respectively, and  $l_i$  and  $\hat{l}_i$  are their respective text lengths.

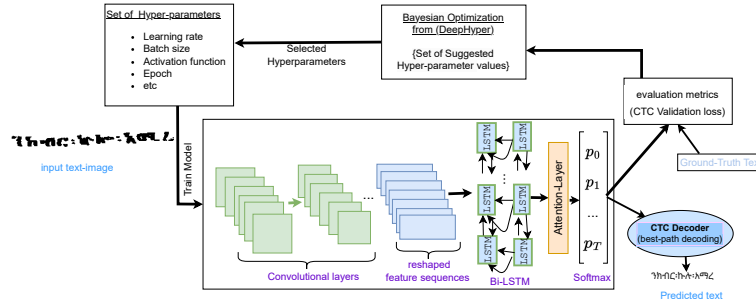


Figure 11: A typical view of the proposed model and set of best hyper-parameters value selection using Bayesian optimization from DeepHyper<sup>6</sup>. The output denoted by  $p_0, p_1 \dots p_T$ , is a matrix of Softmax probabilities with dimensions  $[(c + 1) \times T]$ , where  $c$  is the number of unique characters in the ground-truth text and  $T$  is the length of the input time-step to the LSTM layers. The validation loss was utilized as the metric for tuning the hyperparameters. To obtain the final output sequence from the predicted probabilities produced by the model, we use the best-path decoding strategy.

## 430 C.2 Training details and configurations

431 During our experiments, we employed various hyperparameter settings, including those selected by  
 432 Bayesian Optimization [3] specifically for the CTC-based models. Training and evaluation were  
 433 performed on a single NVIDIA RTX A6000 GPU for all the baseline models. Except for the TrOCR  
 434 transformer-based models, the training process for each individual model required a wall time of  
 435 less than 2.5 hours. However, when considering that there were 10 experiments conducted, with  
 436 each model trained 10 times, the cumulative wall training time is going to be 25 hours each (i.e 10  
 437 experiments \* 10 runs \* 2.5 = 250 hours in total). Additional details regarding the training can be  
 438 found in the provided at <https://github.com/bdu-bf/HHD-Ethiopic>.

439 For the CTC-based baseline models, we trained them multiple times with different hyperparameter  
 440 values, including epochs ranging from 10 to 100, employing a trial-and-error approach and utilizing  
 441 the hyperparameters suggested by Bayesian Optimization. In this paper, we report the results obtained  
 442 from the two CTC-based models (without attention) achieving better CER in 15 epochs. Additionally,  
 443 the attention-CTC models showed improved performance as we trained them for more epochs. The  
 444 reported results, for attention-CTC models, in the main paper were trained for 100 epochs.

445 Despite the TrOCR [14] model has been reported to achieve state-of-the-art performance in the  
 446 original paper, it has a significant drawback due to its large number of parameters. Our attempts to  
 447 fine-tune the TrOCR model using our HHD-Ethiopic dataset, following the provided tutorial, faced  
 448 substantial computational challenges. Training the model for just 3 epochs on a single NVIDIA  
 449 RTX A600 GPU took over 24 hours, resulting in comparatively lower performance compared to the  
 450 CTC-based baseline models. Considering our focus on low-resource settings, we prioritize optimizing  
 451 our time and resources effectively. Hence, as it is not suitable for training in resource-constrained  
 452 environments, we do not recommend utilizing the TrOCR model for Ethiopic text recognition. **Instead,**  
 453 **we prioritize exploring alternative models ( such as the smaller CTC-based methods discussed in the**  
 454 **main paper) which balance between computational efficiency and performance to ensure the feasibil-**  
 455 **ity of the OCR system in limited resources. However, if you possess significant computing resources,**  
 456 **using synthetic data and conducting more extensive training iterations on those models could lead to**  
 457 **an improvement in recognition performance for historical handwritten Ethiopic manuscripts.**

458 We also evaluated various other models [9, 10, 17, 18] **using our HHD-Ethiopic dataset. Although**  
 459 **these models still have a relatively high number of parameters in comparison to the CTC-based**

460 models ( the plain and Attn-CTC), they remain more manageable in low-resource settings. Despite  
 461 the increased parameter count, we run these models for 25 epochs using limited computational  
 462 resources. We achieved an improved recognition performance compared to the results presented in the  
 463 TrOCR paper. By balancing performance and resource demands, the models [9, 10, 17, 18] present a  
 464 viable option for practical deployment and utilization, especially in situations where computational  
 465 resources are constrained.

466 Due to the limited number of experimental runs conducted for [9, 10, 14, 17, 18] baseline models, we  
 467 decided not to include box plots for all baseline models in the main paper. Box plots are commonly  
 468 used to visualize results distribution across multiple runs, allowing for the assessment of variations  
 469 and identification of outliers. Since a box plot is not suitable for representing a single experiment, we  
 470 have illustrated the learning curve of the four models (ABINet, ASTER, SVTR and CRNN) in Figure  
 471 12. This learning curve illustrates the recognition performance on both IID and OOD test sets using  
 472 the CER and metric across 25 epochs. For detailed configurations of each baseline OCR model and  
 473 the implementation of Bayesian optimization, please refer to our GitHub repository at the following  
 474 link at <https://github.com/bdu-bf/HHD-Ethiopic>.

475

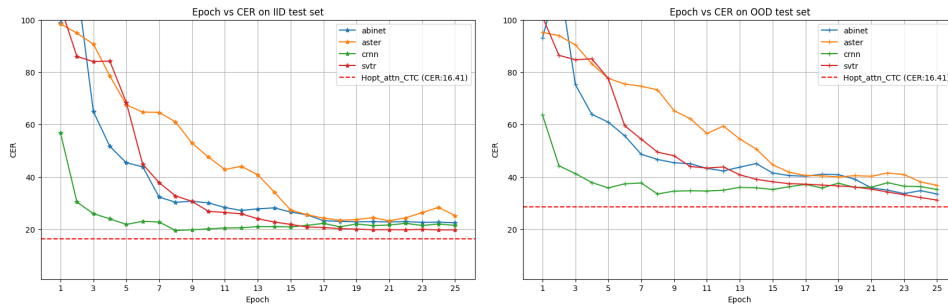


Figure 12: Learning curve on IID and OOD test data. CER<sup>1</sup> on IID test set (left), CER on OOD test set (right) across 25 epochs for ASTER, ABINet, SVTR, and CRNN models. In all plots, the red horizontal line represents the CER value of the Hopt-attn-CTC network on IID and OOD data respectively.

476 Based on learning curve depicted in Figure 12, we can conclude that all models would perform better  
 477 as we train for longer epochs. Within the first 25 epochs, SVTR outperforms the others, while ASTER  
 478 is the least performer. We are limited to running for 25 epochs due to time and computational re-  
 479 sources. The red horizontal line in both the right and left plots represents the CER for Hopt-attn-CTC  
 480 model. This line serves as our benchmark, as it represents the best-performing model.

### 481 C.3 Data collection and annotation process

482 The Ethiopic script, one of the oldest in the world, is underrepresented in the fields of document  
 483 image analysis (DIA) and natural language processing (NLP). This is due to the lack of attention  
 484 from researchers in these fields and the absence of annotated datasets suitable for machine learning.  
 485 However, in recent times, there has been a significant increase in interest from individuals involved in  
 486 computing and digital humanities. As part of this growing attention, we have contributed by preparing  
 487 this first sizable historical handwritten dataset for Ethiopic text-image recognition. The primary  
 488 source of these documents is the Ethiopian National Archive and Library Agency (ENALA), spanning  
 489 from the 18<sup>th</sup> to the 20<sup>th</sup> century. To ensure privacy, each page is randomly sampled from about seven  
 490 different books covering cultural and religious related contents. After obtaining scanned copies of

<sup>1</sup>Please note that the CER can exceed 100% when the predicted text is much longer than the ground truth. Excessive length leads to an edit distance surpassing the ground truth's character count. For instance, if the ground truth is 'ab' and the prediction is 'abced' the edit distance is 3 compared to the ground truth's 2 characters. This results in a ratio of  $1.5 \times 100 = 150$  (see equations 3). In contrast, NED ranges from 0 to 100%, where values close to 0 are better, while values closer to 100% are indicative of poorer performances in both metrics.

491 the documents from ENALA, we utilize the OCRopus<sup>2</sup> OCR framework and the ground-truth text  
492 annotation process is described as follows:

493 The annotation process can be grouped in three phase:

494 • **Phase-I:** In this phase, we hired 14 individuals who are familiar with the Ethiopic script.  
495 Out of the 14, 12 were assigned the task of annotation, while the remaining two served as  
496 supervisors responsible for follow-up the annotation process and ensuring the completeness  
497 of each annotation submission. Additionally, the supervisors were responsible for multiple  
498 tasks, including monitoring the progress of each annotator, providing assistance when issues  
499 arose, making decisions to address any problems encountered during the annotation process,  
500 checking alignment consistency between images and ground-truth at each phase of the  
501 annotator’s submission, and making necessary corrections in case of errors. Throughout the  
502 annotation process, all annotators and supervisors had the freedom to refer to any necessary  
503 references.

504 • **Phase-II:** Once we have all the annotated text-line images from phase-I, we divide the text-  
505 image into training and test sets. For the training set, we reserve all text line images from  
506 the 19<sup>th</sup> and 20<sup>th</sup> centuries, as well as a few documents with unknown publication dates.  
507 The test set is exclusively composed of text line images from the 18<sup>th</sup> century. Additionally,  
508 we randomly sample another test set, which constitutes 10% of the training set. We call this  
509 randomly selected set as **Test-set-I**, which allows us to evaluate the baseline performance in  
510 the classical IID (Independently and Identically Distributed) setting.

511 On the other hand, the test set that is drawn from a different distribution than the training set,  
512 known as Out-Of-Distribution (OOD), is called **Test-set-II**. This setup enables us to assess  
513 the performance in real scenarios where the test set differs from the training distribution.

514 • **Phase-III:** In this phase, we hired approximately 20 individuals who are familiar with  
515 the Ethiopic script, along with one historical expert for the second round of annotation  
516 and request them to submit within 5 weeks. This annotation phase has the following two  
517 objectives:

- 518 – to ensure the quality of the test set.
- 519 – to evaluate the human-level performance in historical Ethiopic script recognition, which  
520 serves as a baseline for comparison with machine learning models.

521 Out of the 20 individuals hired, only 13 annotators successfully completed the annotation  
522 task within the specified submission deadline, while the remaining individuals failed and  
523 resigned from the task. Among the 13 successful annotators, the first group comprised 9  
524 people who transcribed text-line images from the first test set, which consisted of 6,375  
525 randomly selected images from the training set. The second group consisted of 4 people  
526 who transcribed the second test set, consisting of 15,935 images from the 18<sup>th</sup> century.

527 With the exception of the expert reviewer, who was allowed to use external references, all  
528 annotators in this phase were instructed to perform the task without the use of references.  
529 Detailed data from each annotator was documented as metadata for future reference and can  
530 be accessed from our GitHub repository. One observation we made during this annotation  
531 process was that some annotators anonymously shared information, despite our efforts to  
532 ensure data confidentiality. However, despite this limitation, we have successfully compute  
533 the human-level performance for each annotator and have reported the results accordingly.

534 Considering the resources available to the annotators, including computing infrastructure and internet  
535 access, we developed a simple user-friendly tool with a easy to use Graphical User Interface (GUI)  
536 for the annotation process. The tool is depicted in Figure 13.

537 Each annotator’s machine was equipped with this tool, enabling them to work offline when internet  
538 access was unavailable. Additionally, we provided them with a comprehensive *README* file and  
539 instructed them on how to utilize the annotation tool.

---

<sup>2</sup><https://github.com/ocropus/ocropy>

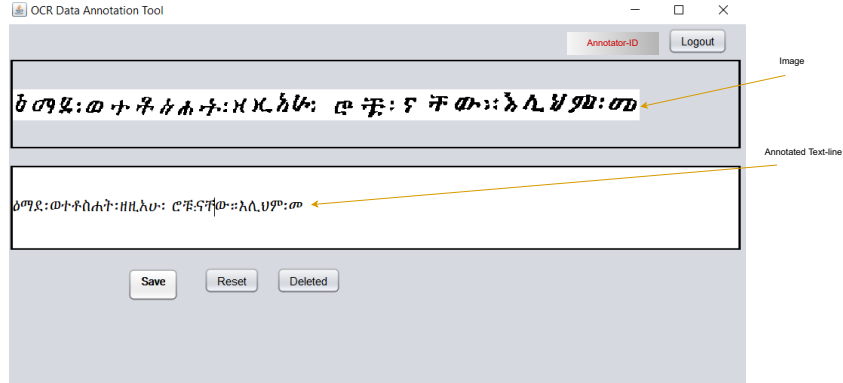


Figure 13: Text-line image annotation tool

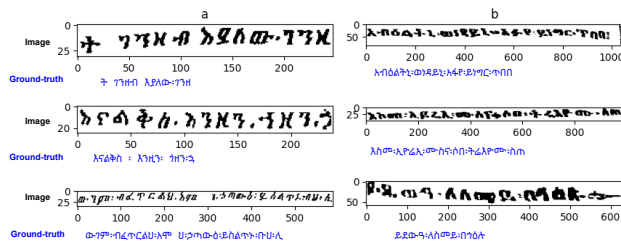


Figure 14: Sample text-line images and ground-truth for HHD-Ethiopic: a) Training-set. b) Test-set.

#### 540 C.4 Dataset statistical overview and comparisons

541 This section provides a detailed description of the characteristics of the HHD-Ethiopic dataset. These  
 542 characteristics include the diversity of content, variations in image quality, distribution of image sizes  
 543 in the trainin and test sets, the number of samples per class, and a comparison with related datasets.

544 Examples of sample page images are illustrated in Figure 15, showcasing pages from various  
 545 publication years (categorized as 18<sup>th</sup>, 19<sup>th</sup>, 20<sup>th</sup>, and unknown date of publication). In addition,  
 546 Figure 16 displays page images categorized by image quality, which ranges from bad to medium and  
 547 good. It's important to note that documents of insufficient quality, falling below the "bad" threshold,  
 548 are excluded during the process of text line extraction.

549 The histogram in Figure 18 illustrates the distribution of text-line image sizes (width and height)  
 550 across the training set and two test sets. Additionally, access to the distribution of characters for each  
 551 class (i.e., the frequency of characters within the 306 unique characters) in both the training and test  
 552 sets is available at [https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/Dataset/](https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/Dataset/distribution_of_characters)  
 553 [distribution\\_of\\_characters](https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/Dataset/distribution_of_characters).

554 To better represent characters that are infrequent or absent in the training set, we have employed a  
 555 solution involving the generation of synthetic images. Each character is incorporated into synthetic im-  
 556 ages approximately 200 times on average. In our scenario, we have identified characters that occur 20  
 557 times or less. About 1200 newly generated synthetic text-line images featuring these underrepresented  
 558 characters are provided on Hugging Face: [https://huggingface.co/datasets/OCR-Ethiopic/](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic/tree/main/train/train_raw/under_represented_char_synth)  
 559 [HHD-Ethiopic/tree/main/train/train\\_raw/under\\_represented\\_char\\_synth](https://huggingface.co/datasets/OCR-Ethiopic/HHD-Ethiopic/tree/main/train/train_raw/under_represented_char_synth). Figure 17  
 560 depicts these characters along with their corresponding frequencies in the training set.

561 Though it may not be fair to directly compare datasets from distinct settings, we provide a com-  
 562 parisons between our historical handwritten (HHD-Ethiopic) dataset and the existing collections of  
 563 modern printed, modern handwritten, and scene text datasets for the task of Ethiopic script recognition.  
 564 The summary of comparisons is is given in Table C.4.





Figure 15: Sample page images ranging from 18<sup>th</sup>, 19<sup>th</sup>, 20<sup>th</sup> centuries, as well as images of unknown publication dates, arranged from top left, top right, bottom left and bottom right respectively.

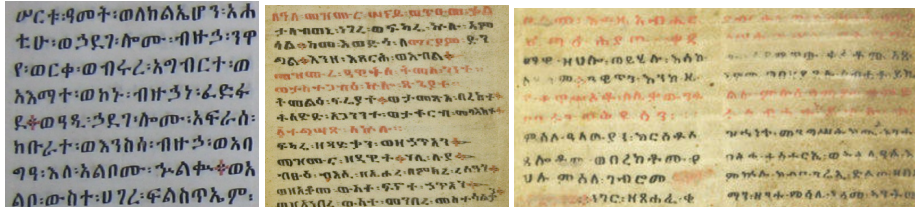


Figure 16: Sample page image images with good(left), medium(middle) and bad (right) quality.

Table 4: Summary of publicly available datasets for Ethiopic script

Dataset-type	image-type	# images	# uniq-chars	# test-sample	annotations
Printed[5]	real	40,929	280	2,907	line-level
	synthetic	296,408	280	15724	line-level
Scene[8]	real	15,39	302	9,257	word-level
	synthetic	2.8M	302	-	word-level
Handwritten[1]	real/modern	12,064	300	1,2064	word-level
	Augmented	33,672	-	*	word-level
Handwritten[2]	real/modern	10,932	265	-	word-level
<b>Our</b> (HHD-Ethiopic)	real/historical	79,684	306	22,310	line-level
	synthetic	100,000	306	*	line-level

- denotes information that is unavailable/ not given  
 \* denotes data that has not been utilized for testing

565 **C.5 Sample predicted texts**

566 Sample images with the corresponding ground truth, model prediction and the edit distance between  
 567 the ground truth and the prediction at line level is shown in Figure19

568 In text lines where characters with low occurrence rates appear in the ground truth of the training set  
 569 often leads to an increased edit distance between the ground truth and the predicted texts during test  
 570 time. This pattern is demonstrated by sample examples depicted in Figure.20

፩	20	፪	14	፫	7	፬	5	፭	4	፮	3	፯	1	፰	1
፱	18	፳	14	፴	7	፵	5	፶	4	፷	3	፸	1	፹	1
፻	17	፺	13	፻	6	፼	5	፽	4	፿	2	፾	1	፿	0
፿	17	፽	12	፼	6	፻	5	፺	4	፻	2	፺	1	፻	0
፻	17	፻	12	፻	6	፺	5	፻	3	፻	2	፻	1		
፻	17	፻	10	፻	6	፻	5	፻	3	፻	2	፻	1		
፻	16	፻	9	፻	6	፻	4	፻	3	፻	1	፻	1		
፻	16	፻	9	፻	6	፻	4	፻	3	፻	1	፻	1		
፻	15	፻	8	፻	6	፻	4	፻	3	፻	1	፻	1		
፻	14	፻	7	፻	6	፻	4	፻	3	፻	1	፻	1		

Figure 17: Frequency distribution of underrepresented characters occurring 20 times or less in the training set. zero in the frequency column refers to the characters that exit in the test set but not in the training set.

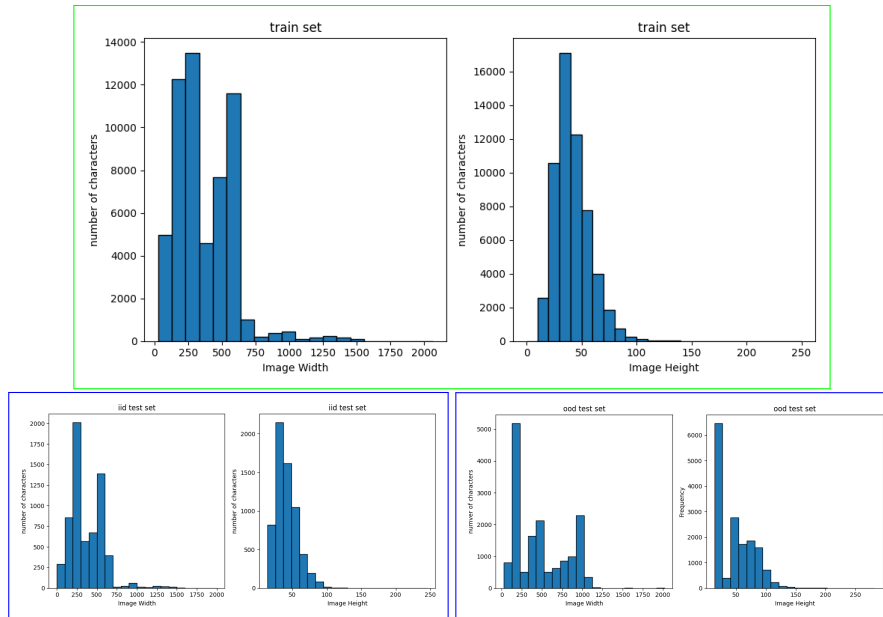


Figure 18: A histogram for distribution of image sizes in the HHD-Ethiopic dataset: a) Training-set (top), b) IID test-set (bottom left), c) OOD test set(bottom right).

571 Finally, we (the authors) believe that this supplementary material serves as an invaluable resource  
572 for reproducing the reported results and conducting further research on historical Ethiopic OCR.  
573 It encompasses crucial contents, including detailed information on the dataset and its preparation,  
574 strategies employed for training the baseline model, and additional essential information required for  
575 replicating the findings. This supplementary material also includes access links to the dataset and  
576 source code, enabling researchers to easily access and utilize these resources. By making use of this  
577 comprehensive supplementary material, researchers can gain deep insights into the HHD-Ethiopic  
578 dataset, the training process of the baseline OCR model, and other necessary details for accurately  
579 reproducing the results and to use this new OCR dataset. This comprehensive resource significantly  
580 supports individuals interested in working on Ethiopic OCR, providing a benchmark for their machine  
581 learning models and contributing to the advancement of research in these field.



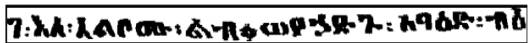
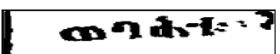
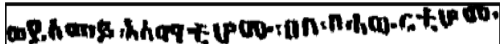
<p>GT Text: እስሙ፡ኢዪሬኢ፡ሙስና፡ሰባ፡ትሬእዮሙ፡ለባ        Pred Text: እስሙ፡ኢዪሬኢ፡ሙስና፡ሰባ፡ትሬእዮሙ፡ሰሙ        Edit Distance: 9</p>	
<p>GT Text: ቢባን፡ይመው፡ቱ፡ወከማሁ፡ይትጎሎ፡አብዳ        Pred Text: ቢባን፡ይመው፡ቱ፡ወከማሁ፡ይትጎሎ፡አብዳ        Edit Distance: 4</p>	
<p>GT Text: ገ፡እለ፡አልባሙ፡ልብ፡ወዮኃድገ፡ለባዕድ፡ብዕ        Pred Text: ገ፡እለ፡አልባሙ፡ልብ፡ወዮኃቡገ፡ዘግዕድ፡ብስ        Edit Distance: 6</p>	
<p>GT Text: ወባሕ፡ቱ፡        Pred Text: ወባሕ፡ቱ፡        Edit Distance: 2</p>	
<p>GT Text: ወደሰሙይ፡አስማቲሆሙ፡ባብ፡ባሐውርቲሆሙ        Pred Text: ወደሰሙይ፡አስማቲሆሙ፡ባብ፡ባሐውርቲሆሙ        Edit Distance: 7</p>	

Figure 19: Sample text-line images with their corresponding ground-truth and prediction texts

<p>GT Text: እስከ፡ማዕዘኑ፡ትኳን፡ጎሳመዓ፡        Pred Text: እስከ፡ማዕዘኑ፡ትፍገ፡ጎሳመዓ        Edit Distance: 5</p>	<p>GT Text: መጿ፡እባክ፡ሃይማኖት        Pred Text: መጻኢ፡እንከ፡ሃይማኖት        Edit Distance: 6</p>
<p>GT Text: ዠሞ፡መዓጅሞ፡መዓጅ        Pred Text: ገርም፡መዓድ፡ም፡መዓድ        Edit Distance: 6</p>	<p>GT Text: ጳ፡ፓፒሮስ፡ኢየሐክሮስ፡        Pred Text: ክ፡ምጥሮስ፡ኢየሐክሮስ        Edit Distance: 7</p>

Figure 20: Examples of prediction errors for underrepresented characters. The characters marked in red within the ground-truth text are less frequent characters and are wrongly predicted.



This dataset is licensed under a [CC-BY-4.0 International License](https://creativecommons.org/licenses/by/4.0/).

# Bibliography

- 583 [1] Fetulhak Abdurahman, Eyob Sisay, and Kinde Anlay Fante. Ahwr-net: offline handwritten amharic word  
584 recognition using convolutional recurrent neural network. *SN Applied Sciences*, 3:1–11, 2021.
- 585 [2] Yaregal Assabie and Josef Bigun. Offline handwritten amharic word recognition. *Pattern Recognition  
586 Letters*, 32(8):1089–1099, 2011.
- 587 [3] Prasanna Balaprakash, Michael Salim, Thomas D Uram, Venkat Vishwanath, and Stefan M Wild. Deep-  
588 hyper: Asynchronous hyperparameter search for deep neural networks. In *2018 IEEE 25th international  
589 conference on high performance computing (HiPC)*, pages 42–51. IEEE, 2018.
- 590 [4] Birhanu Belay, Tewodros Habtegebrial, Million Meshesha, Marcus Liwicki, Gebeyehu Belay, and Didier  
591 Stricker. Amharic ocr: An end-to-end learning. *Applied Sciences*, 10(3):1117, 2020.
- 592 [5] Birhanu Hailu Belay, Tewodros Habtegebrial, Marcus Liwicki, Gebeyehu Belay, and Didier Stricker.  
593 Amharic text image recognition: database, algorithm, and analysis. In *2019 International conference on  
594 document analysis and recognition (ICDAR)*, pages 1268–1273. IEEE, 2019.
- 595 [6] Birhanu Hailu Belay, Tewodros Habtegebrial, Marcus Liwicki, Gebeyehu Belay, and Didier Stricker. A  
596 blended attention-ctc network architecture for amharic text-image recognition. In *ICPRAM*, pages 435–441,  
597 2021.
- 598 [7] Tadesse Destaw, Seid Muhie Yimam, Abinew Ayele, and Chris Biemann. Question answering classification  
599 for amharic social media community based questions. In *Proceedings of the 1st Annual Meeting of the  
600 ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 137–145, 2022.
- 601 [8] Wondimu Dikubab, Dingkan Liang, Minghui Liao, and Xiang Bai. Comprehensive benchmark datasets  
602 for amharic scene text detection and recognition. *arXiv preprint arXiv:2203.12165*, 2022.
- 603 [9] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-  
604 Gang Jiang. Svtr: Scene text recognition with a single visual model. In Lud De Raedt, editor, *Proceedings  
605 of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 884–890.  
606 International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.
- 607 [10] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans:  
608 Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of  
609 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- 610 [11] Timit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,  
611 Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92,  
612 2021.
- 613 [12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal  
614 classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the  
615 23rd international conference on Machine learning(ICML)*, pages 369–376, 2006.
- 616 [13] James Jeffrey. The 4th century art that died out across the world and the ethiopian scribes trying to preserve  
617 it. <https://www.globalissues.org/news/2014/05/08/18652>, 2004.
- 618 [14] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and  
619 Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI 2023*,  
620 February 2023.

- 621 [15] Marcus Liwicki, Alex Graves, Santiago Fernández, Horst Bunke, and Jürgen Schmidhuber. A novel  
622 approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In  
623 *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*,  
624 2007.
- 625 [16] Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. Morpho-syntactically annotated  
626 amharic treebank. In *CLiF*, pages 48–57, 2016.
- 627 [17] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based  
628 sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis  
629 and machine intelligence*, 39(11):2298–2304, 2016.
- 630 [18] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An  
631 attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and  
632 machine intelligence*, 41(9):2035–2048, 2018.