GrEma: an HTR model for automated transcriptions of the Girifalco asylum's medical records

Grazia Serratore[1,2], Emanuela Nicole Donato[2], Erika Pasceri[1], Antonietta Folino[1], Maria Teresa Chiaravalloti[2]

[1]University of Calabria, Rende(CS), Italy

[2] Institute of Informatics and Telematics of the National Research Council, Rende (CS), Italy

1. CONTEXT

The first asylum in Calabria opened in 1881 in Girifalco (CZ). It remained in function until 1978, when the Basaglia Law has ordered the closure of psychiatric hospitals in Italy. Today, the historical archive - still stored within the original building - preserves **15,794 handwritten medical records**.

2. DIGITIZATION

Digitization encompassed the complete conversion of all medical records into digital format. It was carried out in the premises where the archive is stored, which required the use of equipment suitable for transport. **ScanSnap SV600** flatbed scanner was employed, offering a maximum resolution of 600 dpi. **PDF** (Portable Document Format) was selected as the long-term preservation format because it minimizes risks of alteration while maintaining high quality.

3. TRANSCRIPTION

Different transcription methods were evaluated to balance efficiency, accuracy, and reliability.

- **Manual** transcription ensures high accuracy but is extremely time-consuming;
- **Manually assisted** transcription, including the use of voice dictation software (e.g., Microsoft Dictation, Dragon v5), can speed up the process but still requires human revision;
- **Automated transcription**, particularly Handwritten Text Recognition (**HTR**), proved most suitable for handling the large volume of records.

Transkribus, a leading software for automated image-to-text recognition platform, was ultimately adopted.

4. METHODOLOGY

4.1 Corpus

Only clinical documents including **nosological tables**, **informative forms**, **clinical diaries**, and **patient correspondence** were transcribed not considering personal sensitive data. Effective HTR requires a substantial amount of training data - approximately 15,000 transcribed words (≈75 pages), so we created a representative ground truth consisting of images of clinical documents and their transcriptions, with 10% of the dataset reserved for validation.

4.2 GrEma

The GrEma model was trained using M1 as a base model through transfer learning. Four training sessions of the GrEma model were carried out, as shown in **Table 1**. To measure a model's accuracy, the **Character Error Rate** (CER) metric is considered. It expresses the proportion of misrecognized characters compared to the ground truth, with lower values indicating better performance.

Table 1: Training phases and results of the GrEma model.

Training phase ID	Training set size (pages)	Number of words	Character Error Rate
ID1	129	20,776	19.90%
ID2	245	34,512	16.92%
ID3	347	46,621	14.70%
ID4	566	94,624	14.04%

The learning curve in **Figure 1** shows the trend of the CER throughout the different training phases. The CER decreases progressively over the course of the training process, reflecting diminishing returns in performance improvement as the dataset size increases. The CER value is also influenced by the progressive increase in the number of training epochs: as the number of epochs increases, the CER decreases. as illustrated in **Figure 2**.

Figure 1: Learning curve of GrEma during the training phases (x-axis: training phase ID, y-axis CER%)

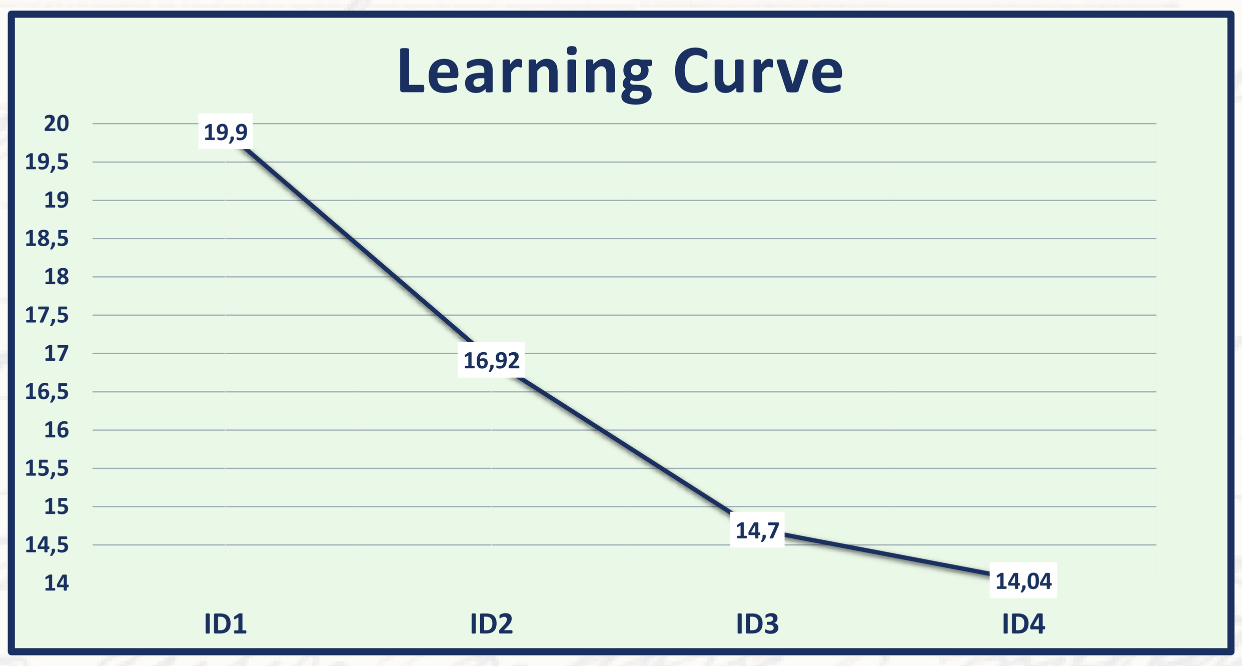
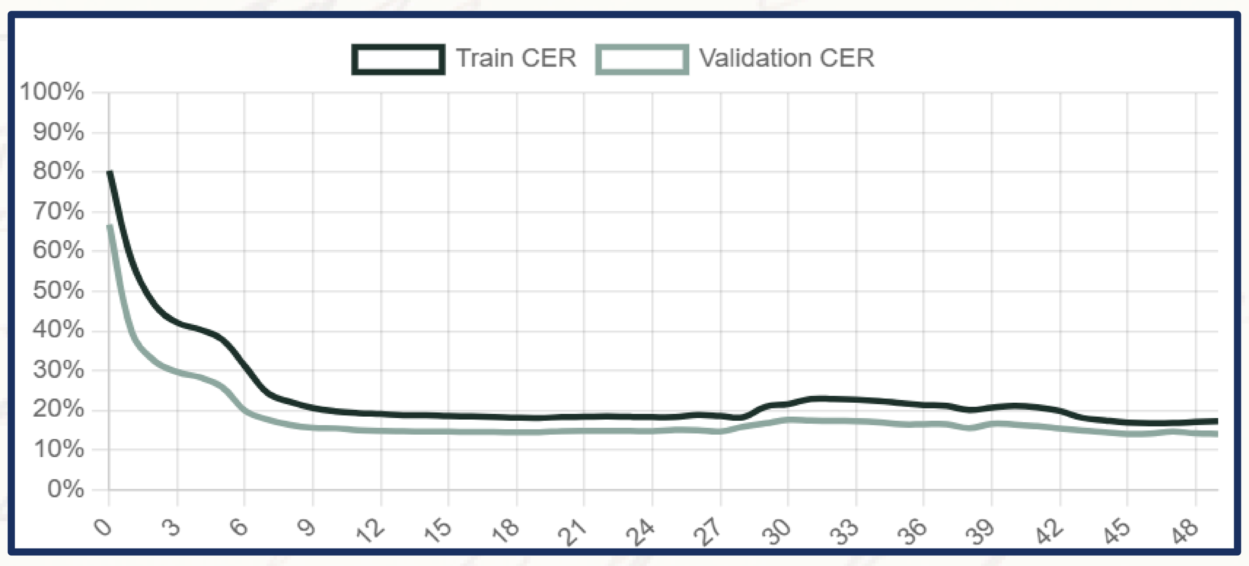


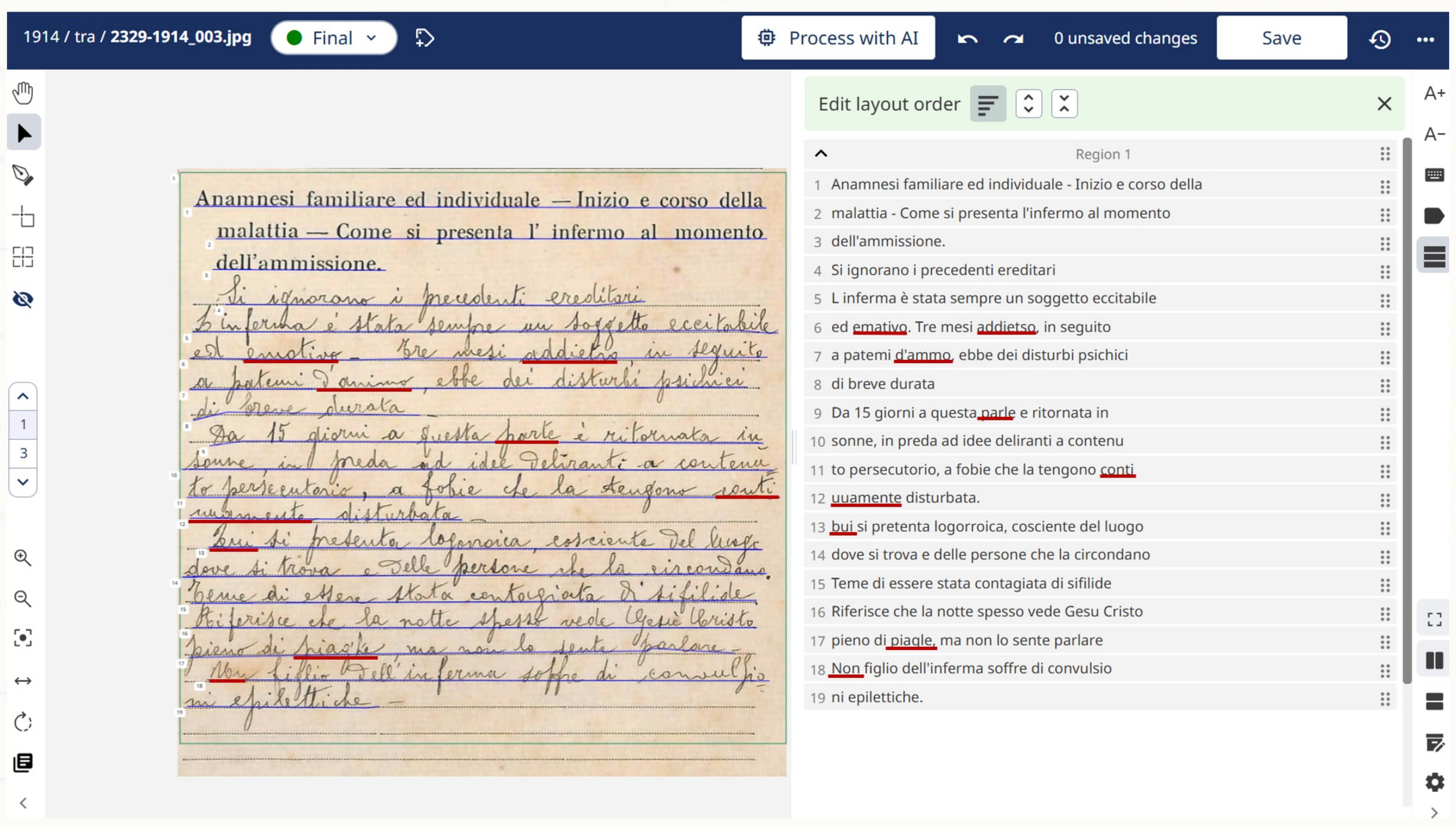
Figure 2: The training chart show how CER changes with the number of GrEma training epochs (x-axis: number of epochs, y-axis: CER%)



4.3 Results

The model produces a confidence matrix rather than a direct transcription, assigning probabilities to each character without considering grammar or syntax rules. As a result, it often confuses similar letters (e.g., **u/n**, **o/a**, **s/r**), as shown in **Figure 3**. Introducing new handwriting styles remains a key adaptation challenge, as medical records were written by multiple hands over the years. Underrepresented handwriting in the training dataset can raise the CER value, even though the model may still provide effective transcriptions.

Figure 3: An example of transcription made with GrEma, with CER at 14.04%



5. FUTURE STEPS

While human revision remains essential for ensuring transcription accuracy, automation could support validation by targeting error-prone contexts. **Natural Language Processing** techniques could be used to automatically correct transcription errors by identifying recurring patterns and applying rules based on Italian grammar. Furthermore, the development of a **digital platform** - modeled on the Cambridge Digital Collection Platform (**CDCP**) - could provide authorized users with access to digitized medical records and corrected transcriptions. Built on international standards, such as IIIF, XML-TEI, EAD and ISAD(G), the platform would combine faithful document reproduction with structured, navigable transcriptions and proper archival contextualization.