

**MRRM AI/ML**

**Page Clustering for Grouping Medical Records**

# Business Requirements

## Problem Statement:

Given a medical history of patient in single document in PDF format contains thousands of pages, the objective is to develop a system that can accurately segment and groups the pages based on their content and its structure. The document contains different Electronic Health Records (EHRs) with treatment summaries, physician notes, doctor notes, clinical notes, correspondence, and other related medical information, spanning an individual's lifetime. **Assume it is completely unstructured data.**

The system must be capable of identifying and group pages that belong to the same report, even if they are interspersed with other information. This will ensure related medical records are properly organized and segmented and helps in carrying forwarding the same **date of service (DOS), header, provider, and facility** for all pages belonging to a single record, ensuring data continuity and consistency.

## Scenarios:

### Scenario 1: Laboratory Report Split across Pages:

In the attached file “**Sample Document .pdf”,** the Laboratory Report starts at page 1, continues on pages 1 to 6, and then concludes on page 6. The system should group pages 1-6 together as a single record. At page 7 there may be chance of encountering is reviewable false page or other category page etc. Please refer to the attached Excel file **Sample Data 1.csv** where we carry forward the same DNP (dos, header, provider, facility) for the relevant pages (1 to 6).

### Scenario 2: Progress Note Split across Pages:

In the attached file “**Sample Document .pdf”**,a physician's progress note starts on page 7, continues on pages 7 to 10, and then concludes on page 10. The system should group pages 7-10 together as a single record, at 11th page there may be chance of encountering is reviewable false record etc. Please refer to the attached Excel file **Sample Data 1.csv** where we carry forward the same DNP (dos, header, provider, facility) for the relevant pages (7 to 10).

### Scenario 3: Interspersed Records:

Generally, a document contains a discharge summary on pages 5-7, a medication list on page 8, and then a follow-up note on pages 9-11. The system should correctly segment and group these records separately.

### Scenario 4: Consistent Headers and Footers:

Generally, a series of lab reports have the same header and footer information on each page, but the page numbers indicate that they are separate reports. The system should recognize this pattern and group the pages accordingly.

### Scenario 5: Varying Record Lengths:

Generally, a document contains records of different lengths, ranging from single-page documents to multi-page reports. The system should be able to handle this variability and accurately group pages regardless of record length.

### Scenario 6: Missing Information across Multiple Pages:

Generally, Many pages in the document may have missing information, such as the Date of Service (DOS), header, provider (doctor names), or facility details. However, each of these fields may be on different pages across the document, not necessarily on the starting of page or same page.

* The system must be robust enough to handle these gaps and correctly group pages using clues such as:
  + **Headers**: Use consistent headers across pages as indicators that the pages belong to the same record, even if the data fields are incomplete.
  + **Continuity across pages**: Track context such as previous pages’ information (like the dos, header, provider (doctor names) and facility (hospital names)) and carry them forward to the next pages where the data might be missing.

This will allow the system to intelligently infer and group pages together, ensuring continuity even in the presence of missing data on individual pages.