10.48047/jocaaa.2026.35.01.08

A Hybrid Approach to Prior Authorization Data Mining from Fax PDFs Using Altova MapForce and Natural Language Processing

Balasubramanian Rengasamy
Humana Inc., USA

Abstract

Another major challenge requiring solutions by healthcare organizations is how to process prior authorization requests that are sent over fax and presented in PDF files, which constitute unstructured or semi-structured data. The continuous use of manual processing incurs the creation of bottlenecks in operations, the escalation of administrative expenses, and delays in the access of patients to the needed medical care. The technologies of Natural Language Processing have developed quite a way, but standalone NLP systems are not capable of identifying essential information with high reliability in the faxed documents, which are usually full of handwritten notes, low-quality images, uneven structure, and complicated medical terms. This article introduces an amalgamation of NLP functionality and Altova MapForce data conversion tools, which can be used to deal with the complexities of automated prior authorization processing. The suggested architecture uses NLP, such as named entity recognition, document classification, and contextual inference, to understand unstructured text in scanned fax PDFs, and MapForce to process extracted data into structured formats that can be processed by the downstream healthcare information systems. The combination of these complementary technologies in the hybrid solution saves a significant volume of manual processing, shortens authorization turnaround time, ensures the quality and accuracy of the data, and provides scalability of the process to address the varying volumes of requests. Implementation issues cover training requirements of NLP models, MapForce configuration plans, quality assurance, and integration issues. The resultant structure provides quantifiable operational efficiencies without compromising the accuracy requirements needed by healthcare applications, and eventually aids in the delivery of patient care more promptly and decreases administrative overhead throughout the healthcare ecosystem.

Index terms: Prior authorization automation, Natural Language Processing, Altova MapForce, Healthcare Document Extraction, Medical Data Transformation.

1. Introduction

Healthcare administrative processes remain burdened by inefficiencies that seem almost anachronistic in an age of digital transformation. Prior authorization workflows, in particular, continue to rely heavily on fax machines and PDF documents, a reality that might surprise those outside the industry. What arrives through these fax lines rarely resembles the clean, structured data that modern systems prefer. Instead, authorization requests come as a chaotic mix of scanned forms, handwritten notes, stamped approvals, and clinical documentation that defies easy categorization.

The burden has grown heavier over recent years. Medical practices find themselves drowning in paperwork, with authorization tasks consuming time that could otherwise go toward patient care [1]. Every scanned request is a riddle: how to read low-quality scans, how to understand abbreviations that differ between providers, how to dig the key information out of the papers that do not follow any common template. Conventional solutions to this issue have been heavily based on human crowdsourcing, where people are

10.48047/jocaaa.2026.35.01.08

required to hand-write information into authorization systems, a practice that is not only timeconsuming but also very likely to create mistakes.

Current trends in Natural Language Processing indicate that the technology is a way to go, but single solutions in NLP do not cope with the specific problems that the faxed medical records are currently facing. The financial toll of maintaining manual processing infrastructure affects every corner of the healthcare ecosystem, from small practices to large payer organizations [2]. What remains needed is something different: an approach that combines the interpretive capabilities of modern NLP with specialized tools designed for data transformation and system integration. This article explores exactly such a hybrid solution, one that pairs NLP's document understanding with Altova MapForce's transformation capabilities to create something more effective than either technology could achieve alone.

2. Problem Statement and Current Challenges

At this point, fax machines are likely to be artifacts of the museum, but they are still widespread in healthcare facilities. Authorization requests still pour through fax lines daily, bringing with them a host of technical problems that complicate any attempt at automation. The documents themselves often look terrible blurred text, crooked scans, coffee stains, and handwriting that ranges from merely difficult to genuinely illegible. Medical practices send requests using whatever forms seem convenient, resulting in hundreds of different templates with varying layouts and information hierarchies [3].

Consider what typically appears on these faxed pages: typed patient demographics sitting alongside handwritten clinical notes, checkboxes marked with uncertain Xs, provider signatures overlapping critical information, and diagnostic images squeezed into margins. Each element demands different handling. Standard OCR software, designed for clean document scans, often chokes on fax quality. Failure to recognize the characters causes failure of all the downstreams with it. Health jargon further complicates the already perplexing abbreviations that have different meanings in different situations, medical codes of diagnosis in the tens of thousands, and short formats of clinician records that presuppose expertise. Even sophisticated NLP models hit walls when processing these documents. General-purpose language models lack the medical domain knowledge necessary to interpret clinical content accurately. Specialized medical NLP systems perform better but still struggle with handwritten notes, where poor penmanship meets medical abbreviations in ways that confound automated interpretation. The stakes matter here, too. Unlike processing restaurant reviews or social media posts, healthcare document extraction errors can cascade into authorization delays that genuinely affect patient outcomes.

Manual processing persists because automation keeps failing. Staff members spend hours each day routing faxes to appropriate queues, squinting at unclear handwriting, typing information into authorization systems field by field, and double-checking entries against source documents. This workflow creates bottlenecks that stretch authorization timelines from hours to days. Physicians report frustration with these delays, particularly when time-sensitive treatments hang in limbo awaiting approval [4]. The human cost shows up in staff burnout rates, while the financial cost appears in operational budgets strained by the need to maintain large teams dedicated to what should be automatable work. Organizations continue searching for solutions that can actually handle the messy reality of faxed authorization requests rather than just working well in controlled test environments.

Challenge Category	Key Issues
Document Quality	Poor fax resolution, skewed scans, background noise, degraded images

10.48047/jocaaa.2026.35.01.08

Content Complexity	Handwritten notes, mixed typed and handwritten text, checkboxes, stamps, embedded images
Format Variability	Hundreds of different templates, inconsistent layouts, and diverse organizational structures
Medical Terminology	Specialized vocabulary, clinical abbreviations, extensive diagnostic codes, and procedure classifications
NLP Limitations	Handwriting recognition difficulties, context dependency, hallucination risks, and accuracy constraints
Operational Impact	Manual routing queues, extended processing times, increased costs, staff burnout, and patient care delays

Table 1: Problem Statement and Current Challenges [3, 4]

Methodology

3. Proposed Hybrid Solution Architecture

Breaking the manual processing cycle requires combining technologies in ways that leverage complementary strengths. Pure NLP approaches struggle with document variability and quality issues. Pure rules-based systems lack the flexibility to handle natural language and contextual interpretation. A hybrid architecture that brings both together offers something more robust. Evidence from medical document processing research supports this direction. Studies show that pairing OCR with NLP techniques can successfully extract structured data from documents that resist conventional processing methods [5].

The architecture begins with document reception and preprocessing. Incoming faxes first undergo enhancement routines that attempt to salvage whatever quality the transmission process degraded. Contrast adjustment, noise filtering, and geometric correction help prepare images for text extraction. Modern OCR engines then tackle the processed images, ideally ones trained on medical document characteristics rather than generic text. What OCR produces is unprocessed text that is partly correct, partly incoherent, and is at least readable by the computer.

This text is extracted and then feeds NLP models, which are trained on healthcare text. Named entity recognition algorithms hunt through the text looking for patterns that signal patient names, identification numbers, dates, diagnosis codes, procedure requests, and provider information. These models have learned from thousands of similar documents what these entities typically look like and where they tend to appear. Document classification algorithms simultaneously assess what type of authorization request has arrived surgery, medication, imaging, or durable medical equipment, because different types need different handling. Relationship extraction adds another dimension, identifying connections between elements: which diagnosis justifies which procedure, which provider is requesting what service, which clinical history supports which medical necessity argument [5].

What NLP produces, however, remains somewhat messy: extracted entities with varying confidence levels, relationships that may or may not be certain, and structured data that still needs transformation into formats that downstream systems can consume. Here, Altova MapForce enters the picture. MapForce specializes in exactly this kind of transformation work, taking data in one format and reliably converting it to another while applying business rules and validation logic along the way [6].

MapForce configurations define precise mappings between NLP extraction outputs and target healthcare data standards. HL7 messages, FHIR resources, X12 transactions, and proprietary payer formats each require different field structures and content expectations. MapForce handles these conversions through

10.48047/jocaaa.2026.35.01.08

visual mapping interfaces where transformation logic becomes diagrams rather than code. Validation rules ensure extracted data meets format requirements before transmission. Enrichment processes draw new information in reference databases that identify the provider NPIs with national registries, diagnoses codes with current ICD-10 versions, and member eligibility with a real-time API call.

The complete workflow moves documents through this pipeline with minimal human touchpoints for straightforward cases. Faxes arrive, preprocessing cleans them up, OCR extracts text, NLP identifies entities and relationships, MapForce transforms everything into target formats, and authorization systems receive properly structured data ready for processing. Confidence scoring throughout the pipeline flags uncertain extractions, routing low-confidence cases to human reviewers who validate and correct before final submission. This human-in-the-loop design maintains quality standards while still achieving significant automation gains on the majority of requests that fall within system capabilities.

Component	Function
Document Preprocessing	Image enhancement, contrast adjustment, noise filtering, geometric correction
OCR Engine	Text extraction from scanned images, medical terminology recognition
Named Entity Recognition	Identification of patient demographics, provider details, diagnosis codes, and procedure requests
Document Classification	Authorization type categorization, routing determination, and priority assessment
Contextual Inference	Clinical justification extraction, medical necessity analysis, and relationship identification
Altova MapForce Validation	Business rules application, format verification, and logical consistency checking
Schema Mapping	Transformation to HL7, FHIR, X12, proprietary formats
Data Enrichment	Provider validation, code verification, eligibility checking, and reference database integration
Exception Handling	Confidence scoring, human review routing, and quality assurance flagging

Table 2: Proposed Hybrid Solution Architecture [5, 6]

4. Technical Implementation Considerations

Getting this hybrid approach working in production environments involves solving numerous technical challenges. Model training alone represents substantial work. Healthcare NLP cannot simply borrow general-purpose language models and expect good results. Medical language has peculiarities, abbreviations, terminology, and contextual patterns that general models miss. Training effective healthcare NLP requires curating datasets from real authorization documents, annotating them with entity labels and relationships, then using these datasets to fine-tune models until performance reaches acceptable levels [7].

Dataset creation becomes a project unto itself. Thousands of documents need manual annotation by people who understand both medical terminology and the extraction task requirements. Clinical staff or medical coders typically handle annotation, marking patient identifiers here, diagnosis codes there, clinical justifications in these paragraphs, and provider details in those sections. This annotation work consumes

10.48047/jocaaa.2026.35.01.08

significant time; each document might take fifteen or twenty minutes to properly label. Organizations need annotated datasets covering diverse document types, provider practices, and authorization categories to build models that generalize well beyond training examples.

Model selection involves trade-offs between accuracy and computational efficiency. Larger transformer models achieve better accuracy but demand more processing power and introduce higher latency. Smaller models run faster but may sacrifice some precision. Healthcare applications tolerate less error than many other domains, pushing requirements toward higher accuracy even at the cost of speed or computational expense. Continuous improvement mechanisms help models learn from mistakes, human reviewers correct errors, corrections feed back into retraining pipelines, and models progressively improve on problematic cases. Implementing these feedback loops requires careful pipeline design to capture corrections systematically and retrain models periodically without disrupting production operations. Quality assurance extends beyond model accuracy to encompass the entire extraction and transformation process. Confidence scoring gives probability estimates to every element extracted, allowing a thresholdbased decision on which cases will automatically be processed, and which require human inspection. The determination of these thresholds is a matter of balancing between automation and accuracy targets, where lower thresholds would imply increased automation and increased error levels, and higher thresholds would imply less automation and reduced errors. Cross-validation methods compare the results of many extraction methods or variants of different models, indicating inconsistencies that indicate extraction uncertainty.

MapForce configuration requires translating business requirements into transformation logic. Healthcare organizations have specific rules about data formats, required fields, valid value ranges, and logical relationships between elements. MapForce mappings encode these rules visually, allowing business analysts familiar with authorization requirements to participate in configuration design without necessarily writing code [8]. Complex scenarios arise frequently: conditional logic that routes different authorization types through different transformation paths, lookup functions that enrich extracted data with reference information, and validation rules that reject submissions with incomplete or inconsistent data before they reach target systems.

Integration work connects MapForce outputs to diverse healthcare information systems. HL7 interfaces, FHIR endpoints, file drops, and database connections. Each integration point has a particular need for the structure of messages, authentication, error reporting, and transaction confirmation. Comprehensive testing of these integrations before deployment into production will help to avoid problems that might corrupt data or cause interruptions to the processes of authorization. A set of test suites subject transformation logic to realistic variations of documents, edge cases, and error conditions to ensure that the system copes well with the messiness of the real world, as opposed to crashing in the face of odd inputs.

Implementation Area	Key Requirements
Dataset Creation	Thousands of annotated documents, diverse format representation, entity labeling, and relationship mapping
Model Training	Domain-specific fine-tuning, medical terminology adaptation, and transformer architecture selection
Confidence Thresholds	Automated processing criteria, human review triggers, and accuracy balancing
MapForce Configuration	Visual mapping design, transformation logic encoding, and validation rules implementation

10.48047/jocaaa.2026.35.01.08

Integration Setup	HL7 interfaces, FHIR endpoints, database connections, and authentication protocols
Quality Assurance	Cross-validation mechanisms, error detection, correction feedback loops
Testing Protocols	Edge case coverage, realistic document variations, integration validation
Continuous Improvement	Retraining pipelines, performance monitoring, and model optimization cycles

Table 4: Technical Implementation Considerations [7, 8] **Discussion**

5. Benefits and Expected Outcomes

5.1 Scalability and Adaptability

Scalability benefits become apparent when authorization volumes fluctuate or grow over time. Manual processing scales linearly at best; doubling volume requires roughly doubling staff. Automated systems scale more efficiently through incremental infrastructure capacity rather than proportional staffing increases. This efficiency proves particularly valuable during seasonal spikes like annual enrollment periods when authorization requests surge temporarily. Organizations handle volume peaks without emergency hiring or service degradation, maintaining consistent processing speeds regardless of volume fluctuations [10].

Adaptability ensures that implementation investments remain valuable despite changing requirements. Healthcare regulations evolve, payer policies change, and new document formats appear as provider practices adopt different systems. Hybrid architectures accommodate these changes through configuration updates rather than fundamental redesign. NLP models retrain on new document types, MapForce mappings adjust to revised data requirements, and integration interfaces update to support new target system versions. This flexibility protects against obsolescence while ensuring systems continue meeting operational needs as healthcare environments shift around them.

Results

6. Benefits and Expected Outcomes

6.1 Operational Efficiency Gains

Companies that have deployed automated prior authorization systems mention the benefits that are far deeper than mere cost savings. Processing speed improvements stand out most obviously; authorization requests that previously took days for manual processing can be completed in hours when automation handles routine cases effectively [10]. Staff experience profound relief when repetitive data entry tasks disappear, freeing time for work that requires human judgment and expertise rather than mechanical transcription.

Financial impacts show up across multiple dimensions. Direct labor costs decrease as automation handles volume that previously required human processing. Organizations processing tens of thousands of authorizations annually realize savings substantial enough to justify implementation investments within reasonable timeframes. Accuracy improvements yield additional savings by reducing downstream correction costs, fewer appeals from providers contesting erroneous denials, less staff time spent investigating discrepancies between authorization records and actual services rendered, and reduced member service interventions when authorization problems disrupt care access [9].

Data quality improvements enable analytics that were previously difficult or impossible. When authorization data exists only in scanned PDF files and scattered notes, extracting insights requires heroic

10.48047/jocaaa.2026.35.01.08

manual efforts. The automated flow of structured and consistent data into data warehouses allows reporting and analysis to be easily done. Organizations become more aware of patterns of authorization, bottlenecks in the processing, denial causes, provider practices, and utilization patterns. This visibility supports evidence-based refinement of authorization criteria and processes, identifying requirements that create administrative burden without proportional clinical value, spotting utilization patterns that warrant policy updates, and recognizing provider education opportunities that could reduce unnecessary authorization requests.

6.2 Enhanced Care Delivery

Patient care quality improves when authorization delays shrink. Time-sensitive treatments proceed without administrative bottlenecks intervening between clinical decision and treatment initiation. Surgical cases get scheduled promptly rather than waiting for authorization confirmation. Medication therapies begin when medically indicated rather than days later after paperwork clears. These improvements matter considerably for conditions where treatment timing affects outcomes. Physicians notice the difference too less time spent following up on pending authorizations, fewer frustrated conversations with staff about administrative delays, and more predictable workflows that align authorization decisions with clinical schedules [10].

Benefit Category	Realized Improvements
Processing Speed	Reduced authorization turnaround times, eliminated manual entry delays
Staff Experience	Elimination of repetitive tasks, focus on high-value activities, and reduced burnout
Patient Care	Timely treatment initiation, reduced appointment cancellations, improved access
Provider Satisfaction	Predictable workflows, faster responses, reduced administrative friction
Cost Reduction	Lower labor expenses, decreased correction costs, and reduced appeal processing
Data Quality	Higher accuracy rates, consistent structured data, and enhanced completeness
Analytics Capability	Authorization pattern visibility, utilization trend identification, policy refinement support
Scalability	Volume handling without proportional staffing, seasonal spike accommodation
Adaptability	Configuration updates for new requirements, model retraining for format changes

Table 5: Benefits and Expected Outcomes [9, 10]

Conclusion

The hybrid system that uses Natural Language Processing and Altova MapForce is a realistic approach to the long-standing problem of faxed prior authorization requests in healthcare facilities. The old manual way of processing is costly in terms of operations, it also introduces so much delay in access to patient care, and it has become a cause of administrative load to both the healthcare givers and the payer organization. In spite of the fact that NLP technologies have grown to a considerable extent, standalone applications cannot cope with the specifics of poor image quality, irregular formatting, handwritten text, and specialized medical lingo that do not lend themselves to the usual automated methods of extraction. A combination of the interpretive functionality of NLP and the powerful transformation and validation capabilities of MapForce forms a solution that meets the document understanding problem as well as the downstream data integration issues that health care settings necessitate. This mixed design recognizes that no one technology can

10.48047/jocaaa.2026.35.01.08

contribute to an organization to address the entire space of problems, but rather to engage in synergetic advantages to produce something that is better than either of the two entities. The operational advantages of the organizations that adopt this architecture are extensive, such as a reduction in processing time by a significant factor, administrative cost reduction, increased accuracy in data handling, and improved scalability, which enables the organization to handle the change in volume without corresponding staffing changes. In addition to operational measures, the solution enables patients to have improved results, removing delays in authorization that may delay needed medical interventions, and at the same time makes the work of delivering the authorization process more seamless and predictable through improved provider satisfaction. The flexibility of the hybrid architecture is necessary to support the long-term value of adaptability, where the organization can add new NLP models, OCR engines, or increased transformation features as technologies change and not fully replace the entire system. The success relies on a considered implementation that involves creating quality training data, setting up correct confidence levels of automated processing, creating a quality assurance system, and ensuring the ability to continue human review of complicated situations beyond the capability of the system. This hybrid solution will provide a feasible solution to healthcare organizations under pressure to cut administrative expenses and enhance the quality of their services by offering a long-term solution to the long-standing resistance of automation efforts and authorization, which will bring long-term benefits in the form of reduced costs, enhanced accuracy, and timeliness of patient care.

References

[1] American Medical Association, "2024 AMA prior authorization physician survey". [Online]. Available: https://www.ama-assn.org/system/files/prior-authorization-survey.pdf

[2] CAQH, "2023 CAQH Index Report," Council for Affordable Quality Healthcare. [Online]. Available: https://www.caqh.org/hubfs/43908627/drupal/2024-01/2023_CAQH_Index_Report.pdf

[3] Jose K. James et al., "Experience With an Optical Character Recognition Search Application for Review of Outside Medical Records", 2024. [Online]. Available: https://www.mcpdigitalhealth.org/article/S2949-7612(24)00075-0/fulltext

[4] OncoSpark, "2024 in Review: Key Prior Authorization Trends and What's Next". [Online]. Available: https://oncospark.com/2024-in-review-key-prior-authorization-trends-and-whats-next/

[5] Enshuo Hsu et al., "Deep learning-based NLP data pipeline for EHR-scanned document information extraction," PMC, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9188320/ [6] Simbo.AI, "The Impact of Prior Authorization Processes on Healthcare Claim Denials and Solutions for Improvement". [Online]. Available: https://www.simbo.ai/blog/the-impact-of-prior-authorizationprocesses-on-healthcare-claim-denials-and-solutions-for-improvement-641450/

[7] Shiyong Wang and Hong Luo, "The application of natural language processing technology in hospital network information management systems: Potential for improving diagnostic accuracy and efficiency," SLAS Technology, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2472630325000457

[8] Shengyu Liu et al., "Evaluating Medical Entity Recognition in Health Care: Entity Model Quantitative Study," JMIR Med Inform, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11528166/

[9] Altova, "Altova MapForce offers unparalleled power and flexibility for advanced data mapping, conversion, and transformation". [Online]. Available: https://www.altova.com/mapforce

10.48047/jocaaa.2026.35.01.08

[10] Dustin Cragun et al., "Counting down to 2025 KLAS Emerging Solutions Top 20," KLAS Research, 2023. [Online]. Available: https://klasresearch.com/report/automated-prior-authorization-2023-whatimpact-do-automated-prior-authorization-solutions-provide/1894