AI-Driven Reimbursement Code Discovery for HealthTech Start-Ups

Xingyin Xu xx943@nyu.edu

Tongyu Zhao tz2658@nyu.edu

Albert Kong lk3189@nyu.edu

Erchi Zhang ez806@nyu.edu

Abstract

Accurate identification of CPT codes is critical for health tech startup reimbursement, but remains labor intensive and costly due to the reliance on manual methods and third-party consultants. In this project, we solve this problem by providing a fast, efficient and reliable tool to streamline the matching of CPT codes, enabling users to focus on providing quality care and innovation. We developed a web application[1] that helps healthcare start-ups match the appropriate CPT codes. The application processes the uploaded documents and returns the top 5 most relevant CPT codes with their descriptions. Our application received qualitative positive feedback and is capable of providing the CPT codes accurately and effectively.

1 Introduction

In the ever-evolving healthcare landscape, securing accurate reimbursement is crucial for the sustainability and growth of HealthTech startups. Central to this process is the identification of relevant Current Procedural Terminology (CPT) codes, which are essential for billing accuracy and compliance. However, traditional methods of matching medical services with CPT codes are often labor intensive, error-prone, and costly, relying heavily on third-party consultants or manual effort.

This project seeks to address these challenges by developing an AI-driven solution that automates and enhances the CPT code identification process. By leveraging the power of natural language processing (NLP) and machine learning, we designed a system to analyze diverse HealthTech startup materials, such as pitch decks and go-to-market (GTM) plans, and recommend the most relevant reimbursement codes. By focusing on key data sources, including open-source reimbursement databases and HealthTech industry materials, our solution provides a comprehensive and scalable approach to facilitate the operations of startups.

The outcome of our project is a user-friendly web application that empowers healthcare startups to efficiently match their services with appropriate CPT codes. Users can upload documents in PDF format and choose between two analytical methods: a keyword-based approach utilizing TF-IDF, or an advanced AI-driven analysis powered by GPT-4o. The system returns the top five CPT codes and their descriptions, simplifying the reimbursement process and reducing the reliance on manual work.

Through this project, we not only contribute to the operational efficiency of HealthTech startups but also offer a real-world application of advanced data science techniques. By streamlining a traditionally complex process, we enable startups to focus on what truly matters: delivering innovative healthcare solutions to improve patient outcomes.

[1] Source code available at `https://github.com/Archertakesitez/deck-to-CPT`

Preprint. Under review.

2 Datasets

The project relies on two main data sources: a curated CPT (Current Procedural Terminology) dataset and user-uploaded PDF documents containing healthcare-related text.

2.1 CPT Dataset

The CPT dataset, sourced from the Codify by AAPC healthcare platform, includes over 10,000 CPT codes along with their detailed descriptions. This dataset undergoes extensive preprocessing to ensure it is ready for efficient use. Raw data inconsistencies are addressed through cleaning and imputation processes, and the refined data is converted into a serialized pickle file format. This conversion significantly reduces loading times and enhances runtime efficiency during analysis.

2.2 User-uploaded PDF

The second data source consists of user-uploaded PDF files containing descriptions of healthcare services. These documents require specialized preprocessing steps to extract relevant information while maintaining data security and privacy. Text extraction is followed by transformations designed to anonymize sensitive information. Using regular expression patterns and the spaCy's Named Entity Recognition (NER) package, personal and organizational identifiers such as email addresses, phone numbers, and company names are systematically removed. These privacy-preserving measures ensure that the data used in downstream processes, including analysis by generative AI, adheres to stringent compliance standards. This dual dataset strategy—combining a rich repository of CPT codes with real-world healthcare text—creates a comprehensive foundation for the project's goals.

3 Methodology

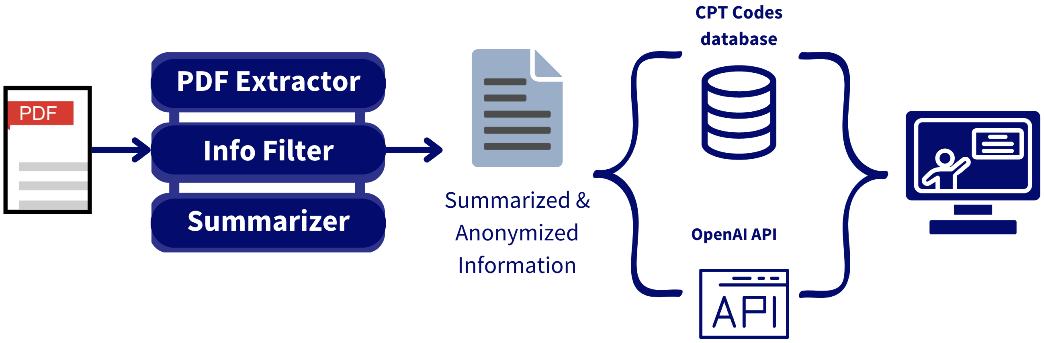


Figure 1: Workflow for parsing the uploaded PDF

The methodology employed in this project focuses on developing a robust tool for matching healthcare services to the appropriate CPT codes. Two distinct approaches are implemented to achieve this objective, providing users with flexibility in how they utilized the system.

3.1 Similarity Based Retrieval

Similarity-based retrieval utilizes traditional text processing and matching techniques. After extracting text from user-uploaded PDFs, the content is preprocessed using a pipeline designed to standardize and clean the data. This pipeline involves removing English stop words, converting all text to lowercase, and so on. These representations allow for effective comparisons across documents, enabling similarity analysis. We examined various algorithms to enhance matching accuracy, including fuzzy text matching, Word2Vec embeddings for semantic similarity, BERT embeddings analyzed through cosine similarity, named entity matching for domain-specific terms, Latent Dirichlet Allocation

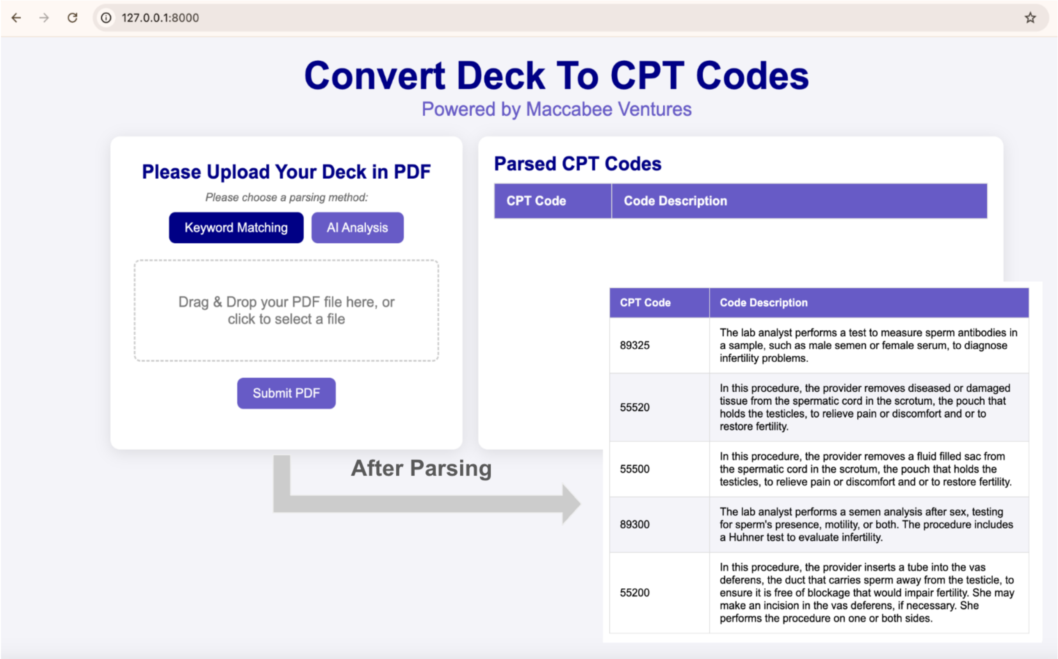
(LDA), and term-frequency inverse-document-frequency (TF-IDF) vectorization. After multiple trials and evaluation, we decided to employ TF-IDF vectorization to generate a numerical representation of the text and generate a similarity score for each comparison between the extracted text from PDFs and the description of each CPT code. The top 5 CPT codes ranked by similarity scores are the results of this methodology. The top 5 CPT codes and their descriptions are stored in JSON format and made accessible through a web interface that provides visual summaries and insights.

3.2 AI Analysis

The second approach relies on AI-powered analysis using GPT-4o, a cutting-edge generative large language model(LLM). Unlike the similarity-based method, this approach requires no prior preparation of CPT dataset, as the generative model has domain knowledge for CPT code in healthcare industry. The extracted and anonymized text from user-uploaded PDF is fed into the GPT-4o API alongside a carefully constructed prompt. This prompt is designed to guide the model in identifying critical service information within the documents and summarizing it into coherent sentences and paragraphs. The summarized information is then used as an intermediary representation to match the services with relevant CPT codes. The model generates a list of the top 5 most relevant CPT codes, ranked by their relevance to the services described in the document. The output, including both the codes and their descriptions, is formatted as JSON and integrated into the web application for user-friendly visualization.

These two complementary approaches are incorporated into a web application that enables users to upload PDF files, select an analysis method, and view results in an intuitive interface. Similarity-based retrieval offers a faster, keyword-driven option for users seeking quick results, while the AI-powered approach provides a deeper analysis with improved accuracy and flexibility. Together, these methods underscore the project's commitment to addressing the challenges of healthcare service classification with innovative yet accessible solutions. By combining traditional text retrieval techniques with state-of-the-art AI, the tool streamlines CPT code identification for healthcare startups and professionals, ultimately reducing the time and effort required for medical billing tasks.

4 Web Interface


127.0.0.1:8000
Convert Deck To CPT Codes
Powered by Maccabee Ventures
Please Upload Your Deck in PDF
Please choose a parsing method:
Keyword Matching
AI Analysis
Drag & Drop your PDF file here, or click to select a file
Submit PDF
Parsed CPT Codes
CPT Code
Code Description
After Parsing

CPT Code	Code Description
89325	The lab analyst performs a test to measure sperm antibodies in a sample, such as male semen or female serum, to diagnose infertility problems.
55520	In this procedure, the provider removes diseased or damaged tissue from the spermatic cord in the scrotum, the pouch that holds the testicles, to relieve pain or discomfort and or to restore fertility.
55500	In this procedure, the provider removes a fluid filled sac from the spermatic cord in the scrotum, the pouch that holds the testicles, to relieve pain or discomfort and or to restore fertility.
89300	The lab analyst performs a semen analysis after sex, testing for sperm's presence, motility, or both. The procedure includes a Huhner test to evaluate infertility.
55200	In this procedure, the provider inserts a tube into the vas deferens, the duct that carries sperm away from the testicle, to ensure it is free of blockage that would impair fertility. She may make an incision in the vas deferens, if necessary. She performs the procedure on one or both sides.

Figure 2: Web Interface, before and after parsing

We have built a web application in which the users can upload their PDF documents, and they can use the "Keyword Matching" tab and the "AI Analysis" tab located on the left side of the screen to choose to parse their documents via the Similarity Based Retrieval approach or the AI Analysis Approach, respectively.

When selecting the 'Keyword Matching' tab, users can upload their document for parsing via drag-and-drop or file selection and then click 'Submit PDF' to process it using the TF-IDF function. The results will appear on the right side.

When selecting the 'AI Analysis' tab, users must input a company name in a designated field before uploading the file. This step helps prevent the users' company information from being leaked to OpenAI's servers. Clicking 'Submit PDF' sanitizes the PDF file, passes the cleaned information to the LLMs, and displays the results on the right.

5 Results

Due to the proprietary nature of CPT codes and their connection to companies' earnings, we were able to collect only limited ground truth data. To evaluate our methods quantitatively, we used precision@5 and recall@5 metrics. Additionally, even though some companies could not disclose the actual CPT codes they use, we gathered qualitative feedback, which further provided valuable insights and context to supplement the evaluation.

5.1 Quantitative Results

To evaluate the performance of our CPT code retrieval tool, we use three key metrics: *Covered Percentage*, *Precision@5*, and *Recall@5*. These metrics are defined as follows:

- **Precision@5**: Precision@5 evaluates the proportion of relevant CPT codes among the top-5 retrieved results.

$$Precision@5 = \frac{Number of Relevant CPT Codes in Top 5}{5} \tag{1}$$

- **Recall@5**: Recall@5 measures the proportion of relevant CPT codes retrieved within the top-5 results compared to the total relevant CPT codes.

$$Recall@5 = \frac{Number of Relevant CPT Codes in Top 5}{Total Number of Relevant CPT Codes} \tag{2}$$

Precision offers an overall measure of the proportion of relevant codes within the top-5 results, while recall emphasizes the presence of relevant codes, regardless of the inclusion of irrelevant ones. By combining these metrics, we gain a balanced evaluation that accounts for both relevance and completeness of the retrieval method.

From Table 1, it is evident that the *AI-analysis* method significantly outperforms the *Similarity-Based Retrieval* method across all metrics. The result indicates that AI-analysis method, on average, is able to generate more than half relevant results.

Table 1: Comparison of retrieval methods on CPT code identification

Metric	Similarity based retrieval	AI-analysis
Precision@5	0.12	**0.52**
Recall@5	0.18	**0.80**

5.2 Qualitative Results

The qualitative feedback received from stakeholders further validates the effectiveness of our application. Users provided positive remarks such as, "Yes 100% these codes are appropriate or relevant to our offerings!" and "This is a great solution that would provide value to digital health companies." Such comments underscore the practical applicability and relevance of the retrieved CPT codes.

Additionally, feedback like “I absolutely love this idea” and “I think this (digital solution) is great...” highlights the enthusiasm and confidence stakeholders have in this solution, even when ground truth data could not be fully shared. These qualitative insights complement the quantitative results, reinforcing that our application delivers meaningful and actionable outcomes for real-world applications.

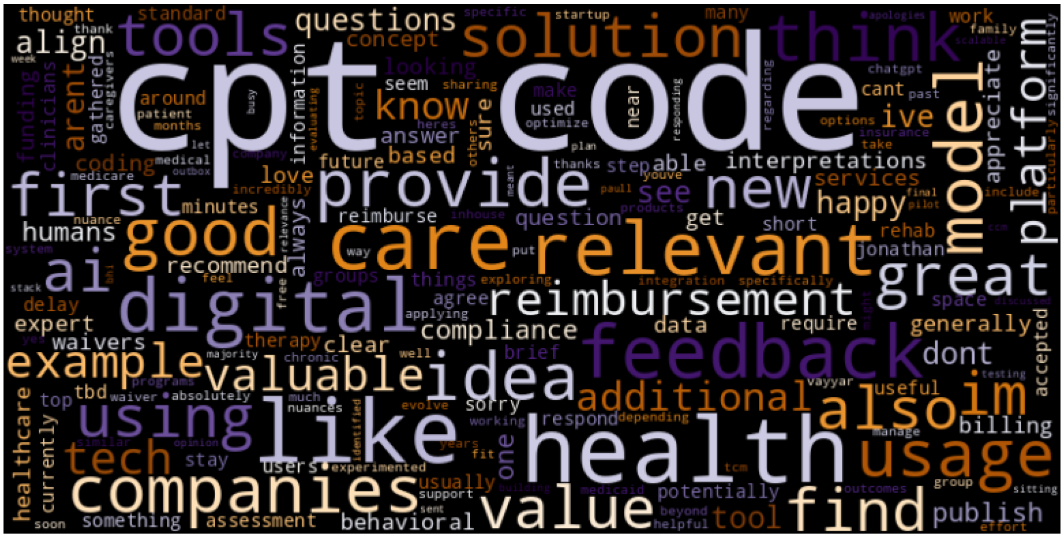


Figure 3: Word cloud of user feedback

As shown in Figure 3, the generated word cloud effectively highlights the most significant words such as "like", "relevant", "good", "value", indicating a strong positive sentiment towards our application. The users seem pleased with the visually appealing result and its relevance to their analysis.

6 Future Work

One key priority is to expand the pitch deck repository by collaborating with healthcare startups, medical organizations, and digital health associations to collect a broader and more diverse set of materials. By engaging with billing consultants and HealthTech venture capital firms, we aim to gather real-world pitch decks, go-to-market (GTM) plans, and solution descriptions that reflect the unique documentation styles and language used across the healthcare industry. Ensuring this data represents various healthcare specialties and service models will allow us to generalize the solution for a wider range of use cases.

Another critical step involves curating an annotated ground truth dataset to improve the accuracy and evaluation of our system. We plan to work closely with domain experts, such as medical coders and billing specialists, to create a benchmark dataset with accurately annotated CPT codes. This will serve as a gold standard for model evaluation and continuous improvement. Incorporating diverse healthcare scenarios, such as collaborative care, telemedicine, and specialty services, will further enhance the robustness and adaptability of the AI models.

We also plan to release new versions of the web application with additional functionalities aimed at enhancing the user experience. These upgrades will include user sign-up and login features, enabling users to access personalized experiences and retrieve previous CPT code matching histories. Another improvement will be the integration of detailed explanations for the recommended CPT codes, offering transparency into the AI’s decision-making process and building trust among users. Additionally, the user interface will be refined to create a more intuitive, responsive, and seamless experience for individuals with varying levels of technical expertise.

To ensure ongoing improvement, we will introduce an interactive feedback mechanism that allows users to validate, reject, or suggest CPT codes. User feedback will be collected in real time and leveraged to retrain and refine both the keyword-based TF-IDF system and the AI-powered GPT-4 model. This iterative approach will help identify patterns in mismatches or edge cases, enabling continuous enhancements to the system. Furthermore, tracking user interactions will allow us to address recurring challenges, while providing incentives to encourage high-quality feedback from domain experts and end-users.

By addressing these future directions, we aim to transform the current tool into a scalable, adaptable, and user-driven platform. These advancements will not only improve the accuracy and efficiency of CPT code identification but also empower HealthTech startups to navigate the reimbursement landscape more effectively, allowing them to focus on delivering innovative solutions in healthcare.