MataAI: AI Assistant for Medical Research and Diagnosis
Project Scope Document

Student Team, COMSATS University Islamabad

May 2025

1 Introduction

- MataAI is an AI-powered application designed to assist doctors in accessing relevant medical research papers, addressing the challenge of information overload in the medical field.
- With over 1 million research papers published annually [2] and medical knowledge doubling every 73 days [1], staying updated is increasingly difficult for physicians.
- The platform prioritizes delivering research papers tailored to doctors' specialties, with secondary features including personalized recommendations, evidence-based Q&A, diagnostic practice, and professional collaboration.
- Developed by a team of 2-3 Computer Science undergraduate students at COMSATS University Islamabad, MataAI aims to enhance research accessibility and improve patient care through advanced AI technologies.

2 Project Overview

- MataAI is a research-first platform focused on collecting, processing, and delivering medical research papers from reputable sources like PubMed, EMBASE, and Cochrane Library.
- Key features include a research paper scraper, personalized feed generator, RAG-based Q&A chatbot, diagnostic chat simulator, and a collaboration platform for professional discussions.
- The platform will offer subscription-based access to premium features like the Q&A chatbot, diagnostic chat, and analytics dashboard, ensuring scalability and user engagement.
- Web and mobile interfaces will provide seamless access, with development completed over 12 months, targeting deployment by May 2026.

3 Problem Statement

The rapid growth of medical literature, with over 1 million papers published annually [2], creates significant challenges for doctors in accessing relevant research tailored to their specialties. This information overload hinders their ability to stay updated, potentially impacting patient care. MataAI aims to address this by providing a system to deliver research papers directly to doctors based on their professional needs, with secondary goals of offering personalized recommendations, evidence-based Q&A, and diagnostic practice tools to enhance efficiency and knowledge application.

4 Objectives

- Develop a system to collect and process medical research papers from sources such as PubMed, EMBASE, Cochrane Library, and others.
- Deliver research papers tailored to doctors' specialties through a personalized feed.
- Implement a RAG-based Q&A chatbot with evidence-based answers and citations, available via subscription.
- Build a diagnostic chat simulator for practicing diagnosis with AI-driven virtual patients, accessible through subscription.
- Foster professional collaboration through a tag-based discussion platform with voting mechanisms.
- Provide a standalone bookmarking system for saving and organizing research papers.
- Offer an analytics dashboard for user insights, available under subscription plans.
- Design user-friendly web and mobile interfaces for seamless access.

5 Scope

5.1 In-Scope

- Development of 17 core modules: user authentication, profile management, research paper scraper, data processing, vector database integration, personalized feed generator, Q&A chatbot, diagnostic chat simulator, notification system, learning tracker, advanced search, bookmarking and saving, collaboration platform, analytics and insights dashboard, subscription and payment processing, web interface, and mobile interface.
- Integration of the PICO framework into search, Q&A, and learning tracker modules for clinical relevance.
- Real-time paper extraction using APIs from PubMed, EMBASE, and other sources, with data pipelines for processing and storage.
- Implementation of data encryption, secure authentication, and compliance with data protection standards.
- Testing and quality assurance for functionality and reliability.

5.2 Out-of-Scope

- Integration with proprietary electronic health record (EHR) systems.
- Support for non-English research papers in the initial phase.
- Development of telemedicine or patient management features.
- Real-time integration with external physician networks (e.g., Sermo API) in the initial phase; considered for future scalability.

6 Features and Modules

The following modules form the core of MataAI, designed to prioritize research paper delivery while supporting secondary features. Each module includes a brief description.

6.1 User Authentication and Management

Enables secure user registration, login, and account management using JSON Web Tokens (JWT) for authentication and role-based access control.

6.2 Profile Management and Personalization

Allows users to set specialties and interests, dynamically updated based on interaction patterns to personalize research feeds.

6.3 Research Paper Scraper

Automates data collection from PubMed, EMBASE, Cochrane Library, Web of Science, Scopus, JAMA Network, The Lancet, BMJ Journals, WHO Global Research Database, ClinicalTrials.gov, SciELO, and PubMed Central using APIs (e.g., PubMed E-utilities [3]) and Scrapy.

6.4 Data Processing and Embedding Generation

Cleans text (e.g., removing HTML tags, stop words) using BeautifulSoup and generates embeddings with BioBERT [4] for medical context, enabling efficient retrieval.

6.5 Vector Database Integration

Stores embeddings in Pinecone for fast similarity searches, supporting recommendations and RAG-based Q&A [5].

6.6 Personalized Feed Generator

Delivers tailored research feeds using a multi-factor algorithm:

- Content-based filtering: Matches papers to specialties and interests.
- Collaborative filtering: Recommends papers read by similar users.

- Behavioral analysis: Tracks reading time and saves.
- Recency weighting: Prioritizes recent papers while preserving seminal works.

Includes "Journal Club" functionality to highlight peer-discussed papers.

6.7 Q&A Chatbot

Provides evidence-based answers via RAG, retrieving papers from Pinecone and generating cited responses. Integrates PICO framework for clinical question structuring. Available under subscription.

6.8 Diagnostic Chat Simulator

Simulates virtual patient interactions for diagnostic practice using rule-based systems and machine learning models like AMIE [6]. Includes real-time guidance and post-case analysis. Available under subscription.

6.9 Notification System

Sends real-time alerts for new papers and updates using AWS SNS for scalability.

6.10 Learning Tracker

Monitors progress based on research papers studied and correct diagnoses in the diagnostic chat simulator, integrating PICO to identify knowledge gaps.

6.11 Advanced Search Functionality

Enables paper discovery with filters (e.g., publication date, journal), boolean operators, and PICO-structured queries for clinical relevance [7].

6.12 Bookmarking and Saving System

A standalone, premium module allowing users to save, organize, and annotate research papers, with features like tagging and offline access for enhanced usability.

6.13 Collaboration Platform

Facilitates professional discussions with a hybrid model inspired by StackOverflow, Quora, and Reddit:

- Tag-based system for topic organization (e.g., cardiology, oncology).
- Upvote/downvote system for prioritizing high-quality content.
- Specialty-specific channels for real-time discussions.

Future scalability may include API integration with platforms like Sermo.

6.14 Analytics and Insights Dashboard

Provides usage statistics (e.g., papers read, Q&A interactions) and personalized insights, available under subscription plans to enhance user engagement.

6.15 Subscription and Payment Processing

Manages access to premium features (Q&A chatbot, diagnostic chat simulator, analytics dashboard) using a subscription model, implemented with Stripe for payment processing.

6.16 Web Interface

Develops a responsive web interface using Next.js, ensuring accessibility and seamless navigation.

6.17 Mobile Interface

Develops a cross-platform mobile app using Flutter, providing consistent functionality with the web interface.

7 Methodology

1. **Data Collection**: Fetch papers using APIs and Scrapy, with Apache Airflow for scheduled real-time updates.
2. **Data Processing**: Clean text and generate BioBERT embeddings, storing them in Pinecone.
3. **RAG Implementation**: Develop RAG-based Q&A combining Pinecone retrieval with Hugging Face Transformers.
4. **Personalization**: Implement multi-factor algorithms for feed generation, using PICO-structured queries.
5. **Interface Development**: Design web (Next.js) and mobile (Flutter) interfaces for responsiveness.
6. **Testing**: Conduct unit, integration, and user acceptance testing with pytest and Selenium.
7. **Deployment**: Deploy on AWS using containerized services for scalability.

8 Timeline

9 Technologies

- **Backend**: Python with FastAPI for scalable RESTful APIs.
- **Web Frontend**: Next.js for responsive, SEO-friendly interfaces.
- **Mobile Frontend**: Flutter for cross-platform iOS and Android apps.

Phase	Duration
Requirements Gathering and Design	Months 1-2
Data Collection and Processing	Months 3-4
Core Feature Development	Months 5-6
Diagnostic Chat and Learning Tracker	Months 7-8
Notification, Collaboration, and Subscription	Months 9-10
Testing and Deployment	Months 11-12

Table 1: Project Timeline

- **Relational Database**: PostgreSQL for user data and metadata.
- **Vector Database**: Pinecone for storing embeddings.
- **AI/ML**: PyTorch and Hugging Face Transformers for NLP and diagnostic models.
- **Data Scraping**: BeautifulSoup and Scrapy for paper extraction.
- **Cloud Services**: AWS for hosting, storage, and notifications.
- **Workflow Management**: Apache Airflow for data pipelines.
- **Payment Processing**: Stripe for subscription management.

10 Resources

- **Team**: 2-3 Computer Science undergraduate students.
- **Tools**: Git for version control, Jira for project management, Slack for communication.
- **Budget**: Free-tier cloud services and open-source tools to minimize costs.

11 Risks and Mitigation

- **Data Quality**: Implement validation and cleaning protocols for reliable data.
- **Model Performance**: Use pre-trained BioBERT with fine-tuning for accuracy.
- **User Adoption**: Gather feedback from medical students and professionals.
- **Scalability**: Design with AWS to handle increased data and users.

12 Data Privacy and Security

- Data encryption at rest and in transit using AES-256 and TLS 1.3.
- Secure authentication with JWT and OAuth 2.0.
- Compliance with data protection regulations (e.g., GDPR, HIPAA where applicable).
- Regular security audits and penetration testing.

13 Deliverables

- Functional web and mobile applications with 17 modules.
- Documentation including user manuals, API specifications, and system architecture.
- Test reports validating functionality and security.
- Final project presentation and demonstration.

14 Constraints

- Team size of 2-3 members requires prioritization of core features.
- 12-month academic timeline limits scope to essential functionalities.
- Budget constraints favor open-source and free-tier tools.

15 Assumptions

- Free-tier cloud services and open-source tools will suffice.
- Research database APIs will remain accessible and stable.
- Feedback from medical students and professionals will guide improvements.

16 References

References

[1] Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, 122, 48–58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116346/

[2] National Library of Medicine. (2023). MEDLINE Citation Counts by Year of Publication. https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

[3] National Library of Medicine. (2023). PubMed E-utilities. https://www.ncbi.nlm.nih.gov/books/NBK25500/

[4] Lee, J., et al. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[5] Pinecone. (2023). Vector database use cases. https://www.pinecone.io/learn/vector-database/

[6] Google Research. (2024). AMIE: A research AI system for diagnostic medical reasoning and conversations. https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-convers

[7] Richardson, W. S., et al. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123(3), A12–A13. https://doi.org/10.7326/ACPJC-1995-123-3-A12