International Journal of Engineering Research & Informatics (IJERI) ISSN: 2348-6481
Vol.4, Issue No 1, 2024

Agentic AI framework for End-to-End Medical Data Inference

[1]G.Prabhakar, Associate Professor , CSE, gantelaprabhakar@gmail.com
Swarna Bharathi institute of science and technology,
Khammam

[2] N.Savitha, Assistant Professor, CSE(DS), savitha.natuva@gmail.com
Swarna Bharathi institute of science and technology,
Khammam

[3]B.Yugandhara Chary ,Assistant Professor, CSE(AIML), yugandhar.bandla@gmail,com
Swarna Bharathi institute of science and technology,
Khammam

Abstract:

Due to fragmented preprocessing procedures, model compatibility challenges, and tight data privacy limitations, building and implementing machine learning solutions in healthcare remains costly and labor-intensive. Through a system of task-specific agents, we automate the full clinical data pipeline in this study, from intake to inference. This architecture is introduced as an Agentic AI system. No human interaction is required for feature selection, model selection, or preprocessing recommendation since these agents can handle both structured and unstructured data. Using geriatric, palliative care, and colonoscopy imaging datasets that are publicly accessible, we assess the system's performance. As an example, the pipeline starts with file-type detection by the "Ingestion Identifier Agent" to ensure privacy compliance. Then, the "Data Anonymizer Agent" identifies the type of data and anonymizes it. This process is repeated for both structured and unstructured data, such as data from colonoscopies or anxiety disorders. With tabular data, the "Feature Extraction Agent" uses an embedding-based technique to identify features, yielding all the column names; with picture data, it uses a multi-stage MedGemma-based approach, yielding the modality and illness name. Following these characteristics, the "Model- Data Feature Matcher Agent" may choose the most suitable model from a pre-selected database. After that, depending on the kind of data and the needs of the model, the "Preprocessing Implementor Agent" and the "Preprocessing Recommender Agent" conduct individualized pre-processing. In the end, the "Model Inference Agent" applies the chosen model to the user-uploaded data and produces understandable results by making use of techniques such as DETR attention maps, SHAP, and LIME. The suggested methodology provides a scalable and cost-effective solution to operationalize AI in healthcare settings by automating these high-friction phases of the ML lifecycle, thereby reducing the requirement for recurring expert involvement. Terms for the Index: AI pipeline, Agentic AI, Medical AI.

Introduction:

Healthcare might undergo a radical shift with the introduction of AI into clinical processes, which would allow for data-driven decisions to be made in real-time regarding diagnosis and treatment planning [1]. Deploying ML models from raw clinical data is still quite laborious and fragmented, even if AI infrastructure has advanced. Instead of developing and evaluating models, data scientists spend as much as 80% of their time on preprocessing, model selection, and pipeline setup [2]. Large multidisciplinary teams consisting of physicians, data engineers, ML specialists, and privacy officers are often needed for these operations. , which results in yearly costs for healthcare facilities ranging from $850,000 to $1.5 million [4, 5], 6]. Institutions dealing with massive amounts of patient data are especially vulnerable to the delays, human error, and financial pressure brought on by this dependence [7]. Adoption of AI in clinical settings is fraught with difficulties relating to privacy, data heterogeneity, model-data alignment, and scalability, in addition to the aforementioned issues [8, 9, 10, 11]. Anonymization and explainability must be built into AI systems as default capabilities rather than add-ons [14] to comply with legal frameworks that require stringent protection of personal health information, such as HIPAA [12] and the General Data Protection Regulation (GDPR) [13]. Similarly, if the models used do not reflect the features of the data that is

available, the performance and dependability of the system might be negatively affected [15]. Even still, model selection is still a domain-specific, labor-intensive process that struggles to keep up with demands in high-volume or time-sensitive settings. To make things even more complicated, clinical data is often multimodal, high-dimensional, and poorly annotated. Still unsolved, despite advancements in model topologies, is the need for infrastructure capable of automatically cleaning, interpreting, and standardizing varied data formats for use in real-world deployments [16]. Agentic AI, a paradigm that describes AI systems as collections of autonomous, modular "agents" each with defined functions and purposes [17], offers a viable solution for tackling these systemic challenges. The system increases its adaptability, interoperability, and efficiency in managing complicated clinical processes by delegating tasks to specialist agents. These agents are capable of seeing, reasoning, acting, and communicating on their own, which gives them the flexibility to work in teams or independently as needed [17].

Literature Review

More and more, fields that need interpretability and flexibility are looking to agentic and modular architectures as potential substitutes for inflexible end-to-end systems [18]. Because medical data pipelines are complicated and variable, agentic architectures are particularly significant in clinical AI. When it comes to enhancing single-model performance, traditional pipelines are often vertically integrated. On the other hand, agentic frameworks facilitate horizontal integration by coordinating various operations that cover the whole data-to-deployment lifecycle. In areas like medical question answering, language models that can reason, maintain context, and perform chain tasks have shown promising outcomes [19], [20]. However, there has been little use of these models in operational clinical systems. An intelligent, specialized agent can carry out every crucial subtask in clinical AI in an agentic framework. Let me give you a few instances. Agents for Preprocessing. Imputation, normalization, encoding, temporal alignment, and feature engineering are all part of the preprocessing stage in clinical machine learning pipelines. These transformations have a crucial influence on the performance and interpretability of the models [21]. Although AutoML systems automate basic preprocessing, they often fail to provide the contextual sensitivity needed for clinical data. Examples of such frameworks include AutoGluon [22] and AutoKeras [23]. Specifically, they might fail to implement recommended procedures like imputation or scaling inside cross-validation folds. Agentic AI systems provide a more flexible answer by coordinating preprocessing operations in real time according to job specifications, data properties, and domain limitations. Bel Esprit [25] and ELT-Bench [24] are two recent systems that demonstrate how LLM-based agents may build end-to-end pipelines autonomously, including logic for feature processing and transformation. When it comes to evaluating agents' abilities to autonomously create Extract-Load-Transform (ELT) pipelines utilizing LLMs that parse metadata and data structure, ELT-Bench is a great tool [24]. In a similar vein, Bel Esprit provides a conversational framework for several agents to work together, whereby they modify a pipeline that goes from user intent to model deployment, often selecting preprocessing depending on context [25]. When it comes to distributed and scalable data processing, frameworks like Intelligent Spark Agents show how LLMs may modularly and adaptively conduct transformation and preprocessing tasks by orchestrating Spark SQL and DataFrame APIs [26]. These agents may adapt their actions in real-time to different kinds of data and environmental input, allowing for better processing of streaming or batch clinical data as well as context-aware feature modifications. For the purpose of scaling stateless preprocessing operations (such as imputation and normalization) across clusters, some systems, like GoldMiner, use a distributed "data worker" approach [27]. Agents for Privacy and Compliance. A dynamic and context-aware approach is provided by agent-based privacy frameworks: they can detect, redact, or anonymize sensitive data at multiple stages of the pipeline—from ingestion through model output. Agentic AI frameworks that include hybrid PHI sanitization, immutable audit logging, and attribute-based access control (ABAC) are one example of what is required to be HIPAA compliant. This method uses a BERT-based model that has been fine-tuned using clinical corpora to identify contextual PHI in unstructured notes [28]. It also makes use of structured rule-based detection, such as regex for SSNs or medical record numbers. In accordance with legislative requirements like as HIPAA and GDPR, these agents preserve audit trails, enforce standards like "minimum necessary" access, and make redaction choices based on context. These days, agents use LLMs to precisely anonymize both structured and unstructured data. With a recall of over 97% and a precision of over 99%—on par with GPT-4 [29]—the LLM-Anonymizer pipeline assesses Llama-2/3 based models on real-world clinical records. This application is great for document-level PII redaction since it lets you use configurable entity definitions, processes scanned documents using OCR, and refines prompts in-context, all without relying on the cloud or losing data. Federated learning (FL), differential privacy (DP), and homomorphic encryption (HE) are cryptographic and decentralized approaches that sophisticated privacy agents may use in addition to redaction. With the use of differential privacy,

a multisite federated learning framework was able to train predictive models on one million actual EHR records without transmitting raw data, all while maintaining accuracy and providing verifiable privacy assurances [30]. Additionally, commercial-grade PII detection programs like NuExtract (numind/NuExtract-v1.5) outperform traditional NER pipelines and display good accuracy across varied document formats [31]. Through system-level output checks and optional human review workflows, these agents intelligently separate PII from other content, use context-aware replacement (e.g., gender consistent synthetic names) to preserve format and context, and maintain entity consistency across documents via deterministic mapping. Assistants for Feature Matching and Model Selection. Model selection agents strive to intelligently automate the difficult and error-prone process of aligning models with the statistical structure and semantic content of clinical data. These bots examine incoming datasets according to domain-specific semantics, dimensionality, schema structure, and modality (e.g., text, tabular, imaging). To make sure they're compatible before deploying, they compare them to databases of pretrained or fine-tunable models.

Agent	Purpose and Functionality
Feature Identifier Agent `[Ingestion_Classifier]`	Detects and classifies file types (e.g., CSV, Excel, ZIP) using Magika, so data-specific downstream workflows can be used.
Data Anonymization Agent `[Ingestion Anonymizer]`	Automatically detects and redacts PII from both structured(tabular) and unstructured(image) data using Google Cloud DLP.
Feature Extraction Agent `[Ingestion_Selector]`	Extracts semantic "headers" for both tabular (column names) and image data (Modality, Disease Type).
Model-Data Matcher Agent `[Ingestion Feature Matcher]`	Matches user data to the most suitable AI model, using the "headers" based on the model database Figure 2. Ensures model-data compatibility.
Preprocessing Recommender Agent `[Preprocessing_Recommender]`	Recommends preprocessing operations for both structured(tabular) and unstructured(image) data based on the identified "headers". Supports automated or custom user-defined preprocessing based on the data size.
Preprocessing Implementor Agent `[Preprocessing Implementor]`	Executes the preprocessing pipeline suggested by the recommender agent.
Model Inference Agent `[Model_Inferencer]`	Runs selected models for final prediction. Supports interpretability via SHAP, LIME, and attention visualizations for respective modalities.

TABLE I: Autonomous AI Agents in the Clinical Data Pipeline and Their Roles

II.INFORMATION

The data used in this research came from places where the data is already accessible to the public. We utilized the GSTRIDE [40] dataset, which includes multimodal sensor data including foot angle, foot pressure, and ankle related variables, for models that were aimed at geriatric applications, including fall prediction. The models are able to accurately represent the nitty-gritty of a person's gait and movement dynamics thanks to these inputs. In the geriatrics scenario, we used a variety of numerical and other data points to determine if the patient was at risk of falling. We utilized data from an open-access dataset for the palliative care-focused hope prediction model [41], [42]. Age, gender, antidepressant use, frequency of vomiting, and other characteristics made up the data. This model is designed to assist with clinical decision-making and psychological assessment by predicting the levels of hope in patients undergoing palliative care. We utilized the dataset described in [43] for polyp-type classification and bounding-box prediction. This dataset is an annotated combination of three publicly available sources: the CVC-ColonDB dataset [44], the GLRC dataset [45], and the KUMC dataset, which comprised 80 colonoscopy video sequences from the University of Kansas Medical Center.

Center.

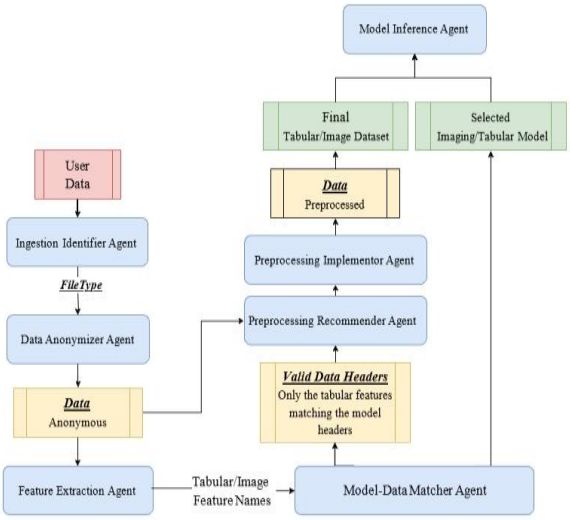

Fig. 1: Complete architecture

Section IV. The Agent Pipeline

We provide AI bots that can work independently inside a clinical data pipeline, handling tasks such as enrichment and inference. The purpose of these agents is to aid medical professionals in making sense of complex multimodal clinical data. There is alignment, contextual awareness, and clinical reasoning in every agent. At each step—from intake to model selection to preprocessing to modeling to explainability—agents work. In Figure 1 we can see the whole architecture with all the agents communicating with one another and the data. Table I provides a brief overview of all the AI agents and the roles they play. The whole process is coordinated by means of the Agent Development Kit (ADK) from Google [46]. All of our agents that carry out the aforementioned duties are listed below. At the very beginning of the clinical data pipeline is the Ingestion Identifier Agent, whose job it is to convert incoming files into a standard format so that subsequent processing may take context into account. A number of formats are compatible

with it, such as CSV, XLSX, JSON, and The ZIP and Magika frameworks use deep learning to examine the structure and byte-level content of the provided data in order to recognize the file types.

Fig. 2: Model Database

Performs automatic identification and masking of personally identifiable information (PII) to guarantee compliance with data privacy and governance standards. For entity-level anonymization, this agent uses the Google Cloud Data Loss Prevention (DLP) API and supports both structured and unstructured data modalities [48]. When dealing with structured data formats like CSV and Excel, the agent use DLP's inspection engine to search for several forms of personally identifiable information (PII). These include, but are not limited to, names, email addresses, phone numbers, medical record numbers, credit card numbers, IP addresses, and dates of birth. To conceal the identified elements while keeping the schema intact and the content obfuscated, we use placeholder tokens of defined length (e.g., "****"). In order to identify textual identifiers and other forms of personally identifiable information (PII) that may be encoded in images, the agent use DLP's visual inspection capabilities. To protect sensitive information, certain areas have been redacted using opaque overlays, such as black rectangles. To facilitate transformations across diverse clinical data that preserve privacy, the agent incorporates visual and textual anonymization into the intake process. Further data preparation and modeling activities may be safely built upon this approach, which also guarantees compatibility with legal frameworks like GDPR and HIPAA.

Model-Data Matcher Agent:

By picking the best model from a curated repository according to the semantic alignment between input features and model requirements, the Model-Data Feature Matcher agent closes the gap between raw data import and model deployment. The agent checks the user-uploaded dataset for structured (tabular) data by comparing the column headers to the necessary headers of each potential model in the model database. Illustration 2. The SapBERT model, which is trained on biomedical literature and is very good at encoding medical language, is used to convert each user data header into a fixed-length embedding vector (size=768) before the candidate models and user data can be compared semantically. In order to determine how close the user-uploaded dataset and the stored model headers are semantically, the comparison procedure use cosine similarity. To be considered eligible, a model must be able to match all of the relevant features to a column in the dataset with a similarity score higher than a threshold, which is often set at 0.6. The agent uses an iterative greedy selection strategy to choose the best available match for each necessary field, ensuring that each user-uploaded dataset column is allocated to no more than one needed heading. This prevents repeated matches. The agent will choose a model as a contender if all of its necessary headers have been matched. Following this, the LLM-enabled agent selects the optimal model according to the model's description. The best model name and a filtered dataset with just the headers needed by the chosen model are part of the final output. This ensures compatibility and prevents downstream schema mismatches. When dealing with unstructured picture data, however, the agent makes use of the MedGemma vision-language model [49] to choose the most appropriate model based on the image's "Modality" and "Disease Type" in order to identify the illness. Feature Extraction, MedGemma (for picture inputs), and Model Matching are the three agents that make up this architecture, as seen in Figure 4. Agents for Preprocessing Recommendation and Supplementation: Based on the uploaded dataset's structure and the chosen machine learning model's criteria, the Preprocessing Recommender Agent will autonomously suggest the best preprocessing procedures. When the dataset size goes over a certain threshold, such 50 MB, the user-specified preparation procedures are removed and are instead chosen automatically according to the model needs. This agent supports both completely automated and user-guided modes. The agent starts by collecting information for each header in tabular data. This metadata includes things like column names, data types, the amount of null and unique values, minimum and maximum values, and the number of attributes. "Binary," "Categorical," "Numerical," and "Textual" are the inferred column types from this information. A rule-

based heuristic is used to classify columns: columns with two unique values are considered binary, columns with few unique values and short string lengths are considered categorical, columns with a data type and distribution of values that exceed 0.8*(total rows of the dataset) are considered numerical, and columns that do not meet these criteria are labeled as text. The choice of header-specific preprocessing procedures is dictated by this labeling. The agent suggests common preprocessing techniques to the user based on the kinds of columns that have been inferred. mThe user does not have to explicitly specify preprocessing when dealing with picture data. A model-specific preprocessing pipeline, closely associated with the chosen picture model, is instead used by the system. A rising tendency in current vision architectures, as shown in models like DEtection TRansformer (DETR) and variations, is to co-train preprocessing routines with the model during its original construction rather than using generic ones. This design decision follows this trend. Image tokenization techniques that are best suited to the model's attention mechanism and feature extraction layers are one example of a pipeline. Other examples include learnt resizing strategies and unique normalization approaches. By enabling each model to encapsulate its own preprocessing logic, this technique improves modularity and eliminates mismatch problems that are common in manual preprocessing. As a result, deployment becomes much easier for a wide variety of image-based applications.

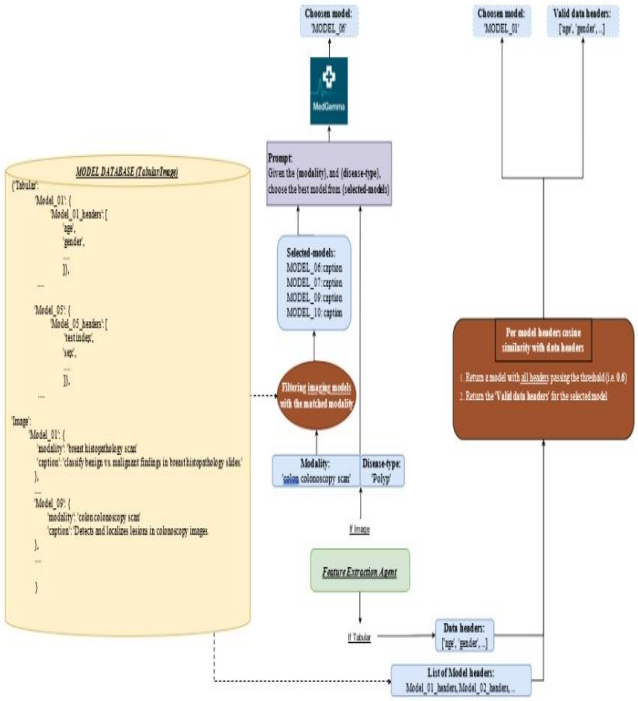

Fig. 4: Ingestion Feature Matcher framework

Complementing the Preprocessing Recommender Agent, the Preprocessing Implementor Agent is responsible for applying the selected preprocessing steps to the dataset. It takes the anonymized, feature-matched data from the previous steps and executes each preprocessing step given by the preprocessing recommender.

Conclusion

Automated, modality-aware transformations ustomized to particular model needs were madepossible by the construction of a feature extraction and model-data matcher agent, as well as a pre-processing recommender. Our new embedding-based feature matching and model selection technique for tabular data uses semantic similarity to fundamentally align the headers of user-uploaded data with the headers of the models. First, we infer the imaging modality and illness category using MedGemma-based multi-stage pipelines. Then, we use vision-language reasoning to pick the best suitable model for image data. The integration of these parts allows for automated inferences on both structured and unstructured clinical activities by systematically aligning incoming data with different stored models. This paradigm is great for places with little data science resources since these components make expert assistance unnecessary. The preparedness for implementation in sensitive healthcare situations is further enhanced by built-in compliance mechanisms, such as HIPAA-aligned anonymization via Google DLP. All things considered, this study lays the groundwork for healthcare AI systems that are scalable, smart, and morally sound. The agentic architecture shows potential for expediting the adoption of safe, interpretable, and cost-effective clinical AI by incorporating autonomous reasoning into every step of the pipeline.

Reference:

[1] A. Schmetz and A. Kampker, “Inside production data science: Exploring the main tasks of data scientists in production environments,” *AI*, vol. 5, no. 2, pp. 873–886, 2024.

[2] M. Nair, P. Svedberg, I. Larsson, and J. M. Nygren, “A comprehensive overview of barriers and strategies for ai implementation in healthcare: Mixed-method design,” *PLoS One*, vol. 19, no. 8, p. e0305949, 2024.

[3] S. Lund, J. Manyika, L. H. Segel, A. Dua, S. Rutherford, B. Hancock, and B. Macon, *The Future of Work in America: People and Places, Today and Tomorrow*. McKinsey Global Institute, 2019.

[4] P. P. Stuti Dhruv, *The Cost of Implementing AI in Healthcare*. AALPHA Information Systems India PVT LTD, 2025.

[5] D. Tahir, *AI was meant to cut health care costs. It turns out to need expensive human support*. San Francisco Chronicle, 2025.

[6] M. Hassan, A. Kushniruk, and E. Borycki, “Barriers to and facilitators of artificial intelligence adoption in health care: scoping review,” *JMIR Human Factors*, vol. 11, p. e48633, 2024.

[7] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, ..., and G. Kaissis, “The future of digital health with federated learning,” *npj Digital Medicine*, vol. 3, no. 1, p. 119, 2020.

[8] A. Subbaswamy and S. Saria, “Preventing failures due to dataset shift: A review,” *arXiv preprint arXiv:2010.02094*, 2020.

[9] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.

[10] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[11] U.S. Congress, “Health insurance portability and accountability act of 1996 (hipaa).” https://www.hhs.gov/hipaa/for-

professionals/privacy/ laws-regulations/index.html, 2025.

[12] European Parliament and Council of the European Union, "General data protection regulation (gdpr)." https://gdpr-info.eu/, 2016. Regulation (EU) 2016/679.

[13] M. Marks and C. E. Haupt, "Ai chatbots, health privacy, and challenges to hipaa compliance," *Jama*, vol. 330, no. 4, pp. 309–310, 2023.

[14] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art,"

Knowledge-based systems, vol. 212, p. 106622, 2020