DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp.958-969, 2026
International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Mapping Open-Source Health AI: A Graph-Analytics Study Across Imaging, EHR, Genomics, and General Clinical Applications

Ralph Baddour
Badhouse Ventures, Vancouver, Canada
research@badhouse.ca

ABSTRACT

Open-source software has become a core driver of progress in health-focused artificial intelligence, yet the structure of its developer and project ecosystem remains poorly understood. This paper presents a graph analytics study of public GitHub repositories related to health AI, revealing how collaborative communities are organised across subfields. A dataset of repositories was assembled using a keyword-based search that combines health-related terms from different subdomains, such as imaging, genetics, and electronic health records, with AI-related terms. Automated filters retained active, substantive projects, while removing duplicates and off-topic entries. The resulting corpus formed a manageable slice of the health-AI landscape. The ecosystem was modelled as a repository-level collaboration network, where nodes represent repositories and edges encode shared contributors and parent–fork relationships, with edge weights reflecting collaboration strength. The study identified mature, densely connected communities alongside emerging or weakly connected areas, and outlines implications for researchers, funders, public code repository maintainers, and industry stakeholders seeking tools, collaborators, or underexplored niches.

Keywords: *open-source, software, health, AI*

1. Introduction

Artificial intelligence (AI) has become deeply embedded in modern healthcare research and development, particularly in medical imaging, bioinformatics, and clinical decision support. Reviews of AI for medical image segmentation, for example, document rapid advances in model architectures and increasing concerns around explainability and trustworthiness in clinical use [1]. Parallel growth is visible in genomics and electronic health records (EHRs), where machine learning now underpins tasks ranging from variant calling to risk prediction. Open-source software occupies a central position in this transformation. Frameworks such as MONAI for medical imaging [2], MONAI Label for AI-assisted annotation [3], and nnU-Net for self-configuring segmentation [4] provide reusable components for researchers and industry teams. Benchmarking efforts such as MedPerf demonstrate how open ecosystems can support federated evaluation of medical AI models across institutions [5].

Despite these developments, relatively little is known about the structure of the broader open-source health-AI ecosystem. GitHub, the dominant platform for public code hosting, has been analysed as a global-scale collaborative social network [6], and social network analysis (SNA) of open-source software (OSS) communities has been systematically surveyed [7]. Recent work even characterises GitHub collaboration as exhibiting small-world properties [8]. However, these studies treat projects in aggregate and do not focus on domain-specific ecosystems such as health AI, where regulatory constraints, data privacy, and clinical workflows create distinctive collaboration patterns. A better understanding of how health-AI repositories and their contributors are organized could benefit several audiences. Researchers and students might discover mature toolchains or well-maintained benchmark suites more efficiently. Funders and maintainers of public repositories could allocate resources to critical infrastructure projects or to fragile but important subcommunities. Industry teams could identify influential open-source projects and bridge contributors when forming collaborations or recruiting.

This paper addresses this gap by constructing and analyzing a collaboration network of health-AI repositories on GitHub, with an emphasis on four subdomains: (i) imaging, (ii) EHR and clinical natural language processing, (iii) genomics and bioinformatics, and (iv) general health-AI applications that do not fit neatly into the other categories. The aim is to assemble a curated dataset of public repositories that combine health-related terminology from these subdomains with AI- and machine-learning-related terms, to model the resulting ecosystem as a repository-level collaboration graph based on shared human contributors and parent–fork relationships, and to use standard network analysis and community detection methods to characterize how projects are organized. Particular attention is given to distinguishing mature, densely connected communities from more recent or less connected areas, and to identifying observations that may be useful to researchers, funders, maintainers of public code repositories, and industry stakeholders seeking tools, collaborators, or underexplored niches within open-source health AI.

2. Research Methodology: Data Collection

The analysis focuses on public repositories hosted on GitHub. Public availability does not guarantee that a repository is licensed under an OSI-approved open-source license, but it does indicate that source code and metadata are freely inspectable. For the purposes of this study, the term "open-source health-AI project" refers to such public repositories, regardless of specific license family. License fields are retained but not used as exclusion criteria. All data collection and analysis steps described in this work were implemented in Python, and the corresponding ingestion and analysis scripts were made available as open-source code at https://github.com/rbad/mapping-health-ai-repositories [10].

The analysis time window spans from the start of 2018 to late 2025. This period covers the surge of deep-learning-based medical imaging frameworks, growth in large-scale genomics and single-cell projects, and the emergence of foundation models and large language models (LLMs) for clinical text. Four major health-AI subdomains were defined for the study:

a. **Imaging** – Medical imaging (e.g., radiology, digital pathology).
b. **EHR** – EHR/EMR/clinical text (structured health records, clinical notes, medical coding, interoperability standards).
c. **Genetics** – Genomics/bioinformatics (genomics, transcriptomics, proteomics).
d. **General** – General health AI (clinical decision support, patient monitoring, outcome prediction, public health analytics).

For each subdomain, a vocabulary of health-related phrases ("health terms") was constructed through an iterative, empirically calibrated procedure. An initial pool of candidate terms was assembled from domain literature (titles and author keywords in recent survey and methodology papers), widely used standards (for example, FHIR and ICD in the EHR setting), and terminology appearing in the names or documentation of prominent open-source toolkits and benchmarks. Each candidate term was then explored using GitHub's repository search to gauge both retrieval volume and topical precision, and terms that consistently produced predominantly generic or off-topic results were discarded or replaced with more specific alternatives. The number of retained terms per subdomain was chosen to keep queries within API limits while still yielding hundreds to thousands of candidate repositories after downstream filtering, acknowledging that terminological breadth differs across imaging, EHR, genomics, and general health AI. The resulting lists of health terms used in this study were as follows:

- Imaging: ["medical imaging", "radiology", "radiological", "xray", "x-ray", "computed tomography", "ct scan", "mri", "ultrasound", "digital pathology"]
- EHR: ["electronic health record", "EHR", "electronic medical record", "EMR", "clinical documentation", "clinical notes", "clinical text", "discharge summary", "ICD coding", "FHIR", "HL7", "medical coding", "medical billing", "claims data"]
- Genetics: ["genomics", "genetic sequencing", "DNA sequencing", "bioinformatics", "transcriptomics", "RNA sequencing", "variant calling", "GWAS", "PCR", "single cell", "proteomics", "CRISPR", "nucleic acid transcription"]
- General: ["clinical decision support", "healthcare analytics", "patient monitoring", "medical time series", "vital signs", "hospital readmission", "ICU mortality", "ER mortality", "patient mortality", "sepsis", "disease risk", "triage system", "public health"]

Each health term was combined with AI- and ML-related terms such as “machine learning”, “deep learning”, “neural network”, and “artificial intelligence” to form GitHub search queries. Keywords were designed to be mutually exclusive across subdomains where possible, in order to minimize overlap created purely by search syntax. The GitHub REST API search/repositories endpoint was used to retrieve candidate repositories. For each combination of health term, AI term, and calendar year, a query of the following form was issued:

> "<health term>" "<AI term>" in:name,description,readme pushed:YYYY-01-01..YYYY-12-31 stars:>=10 archived:false

Year-based segmentation avoided the GitHub limit that restricts search results to the first 1000 hits per query. For each combination (health term, AI term, year), paginated results were collected up to either exhaustion. For every repository returned, the script stored standardized metadata, including repository name, full name (owner/name), description, primary language, star/watch/fork counts, open issue count, creation and last push timestamps, topics, license identifiers, and an approximate README text. A separate step queried each repository’s contributors list to record unique human contributor logins, excluding known bot accounts. Each repository record was stored in the following type of Python data structure:

```
class RepoRecord:
  repo_id: int,  full_name: str,  name: str,  license_name: Optional[str],
  description: Optional[str],  html_url: str,  language: Optional[str],  topics: List[str],
  stars_count: int,  watchers_count: int,  open_issues_count: int,  forks_count: int,
  is_fork: bool,  parent_full_name: Optional[str],  created_at: str,  updated_at: str,
  pushed_at: str,  health_subdomain: str,  health_subdomains: List[str],
  matched_health_terms: List[str],  matched_ai_terms: List[str]
  contributors: List[str],   readme_text: Optional[str]
```

An initial set of filters was then removed:

- repositories with fewer than 10 stars,
- archived projects,
- repositories with fewer than 5 commits in the last year (approximate activity filter), and
- repositories considered off-topic based on a combined health-term and AI-term presence in the concatenated name, description, and README text.

This automated filtering approach was intentionally conservative to retain borderline cases that might be relevant in cross-subdomain analysis.

Repositories often matched multiple queries across health terms, AI terms, or years. De-duplication used the numeric GitHub repository ID as the primary key and the full name as a fallback. For each unique repository, a list of all matched health subdomains was created, allowing the field health_subdomains to be filled (e.g., ["imaging", "genetics"]). To support

simple per-subdomain analyses, a single primary subdomain label was also assigned. When a repository appeared in only one subdomain, that label was used directly. For multi-labelled repositories, a simple content-based tie-breaking procedure examined the description and README text and counted occurrences of subdomain-specific keywords. The subdomain with the highest keyword count was designated as primary. The resulting merged dataset contained 7166 unique repositories. Based on the primary subdomain label:

- imaging: 3657 repositories (51 %)
- genetics: 2349 repositories (≈33 %)
- ehr: 795 repositories (≈11 %)
- general: 365 repositories (≈5 %)

Approximately 6.8% of repositories were multi-labelled (i.e., appearing in both imaging and genetics); these entries were important for analysing cross-subdomain collaboration. Across the corpus, Python and Jupyter Notebook accounted for roughly two-thirds of the primary languages, followed by R, MATLAB, C++, JavaScript, and others, consistent with Python's dominance in contemporary machine learning workflows. Median star counts ranged between 23 and 34 across subdomains, with EHR-related repositories showing slightly higher median popularity metrics, possibly reflecting fewer but more widely used toolkits.

3. Theory and Calculation: Graph Construction and Network Analysis Methods

The collaboration ecosystem was modelled as an undirected, weighted graph $G = (V, E)$. Each node $v \in V$corresponded to a repository. Two types of edges were considered:

a) **Shared-contributor edges.** If at least one non-bot contributor appeared in the contributor lists of two repositories, an edge was added between the corresponding nodes. The edge weight equalled the number of distinct contributors shared by the pair, capturing collaboration intensity.

b) **Parent–fork edges.** If one repository in the dataset was listed as another repository's GitHub parent (i.e., was created by forking), an additional edge was added. If a shared-contributor edge already existed between the pair, the parent_fork flag was set and the weight incremented by one; otherwise, a new edge with weight 1 and parent_fork = True was created.

In this representation, isolated nodes corresponded to repositories whose contributors did not work on any other repository in the dataset and that did not participate in parent–fork relationships within the corpus. Standard network measures were computed, including:

- Node degree (number of adjacent repositories),
- Degree distribution,
- Size and count of connected components,
- Average clustering coefficient,
- Average shortest path length within the largest connected component (LCC).

These metrics provided a high-level characterization of collaboration density and fragmentation, and allowed comparison with prior GitHub-wide studies that reported small-world properties [8]. To identify influential or structurally important repositories, betweenness centrality was calculated on the LCC. Betweenness centrality measures the fraction of shortest paths between all pairs of nodes in the graph that pass through a given node. Nodes with high betweenness could act as “bridges” connecting otherwise separate regions of the network, such as different subdomains or methodological communities. Because the LCC contained 1561 nodes and 6124 edges, exact betweenness centrality was computationally feasible. Repository names and primary subdomains of the highest-centrality nodes were examined to qualitatively interpret their roles. Community structure within the LCC was detected using the Louvain

modularity maximization algorithm, as implemented in the NetworkX Python library [11]. Each resulting community (cluster of nodes with dense internal connections and sparser external links) was treated as a candidate collaborative ecosystem.

For each community, summary statistics were computed:

- number of repositories,
- number of edges and average internal degree,
- edge density,
- distribution of primary subdomains,
- median repository creation date.

Community labels were then derived through a lightweight textual analysis. For each cluster, the concatenated repository descriptions and README snippets were processed using simple term-frequency statistics (bag-of-words with stopword removal). High-frequency domain terms and the identities of high-degree nodes guided manual labeling, for example "general-purpose imaging frameworks", "bioinformatics pipelines for single-cell and omics", or "LLM-based assistants and tooling". This semi-automatic labelling approach balanced scalability with interpretability and is consistent with prior work that combines algorithmic community detection with human interpretation in OSS network studies [7].

4. Results

4.1. Global Network Structure

The repository-level graph contained 7166 nodes and 7673 edges. Degree values ranged from 0 to 61, with an average degree of approximately 2.20 and a median degree of 0, reflecting a large number of isolated projects.

Connected components analysis showed 4616 components, of which 4025 consisted of a single repository with no observed collaboration. In total, 71.5% of repositories belonged to components of size 3 or smaller. At the other extreme, the largest connected component contained 1561 repositories and 6124 internal edges. Figure 1 visualizes this largest connected component using a force-directed layout, revealing a dense central core of repositories surrounded by many short spokes and small peripheral clusters. Since most substantive collaboration occurs within the LCC, the remainder of the analysis focuses on this component. Within the LCC, the average degree rose to 7.89, the median degree to 4, the average clustering coefficient to 0.60, and the average shortest path length to 5.55. These values indicate a densely interconnected core where repositories sharing contributors form overlapping local clusters and are connected via short paths, echoing the small-world behavior observed in broader GitHub studies [8].

4.2. Community Structure: Mature Ecosystems

Louvain community detection on the LCC produced **34 communities**, ranging in size from a handful of repositories to a large cluster with 167 nodes. Several communities clearly corresponded to mature, densely connected ecosystems. Figure 2 summarizes these communities by plotting each cluster's median repository creation year against its average internal degree, with bubble size proportional to the number of repositories. Larger, older communities cluster to the left of the plot, while smaller, more recent, and sometimes highly connected communities appear to the right.

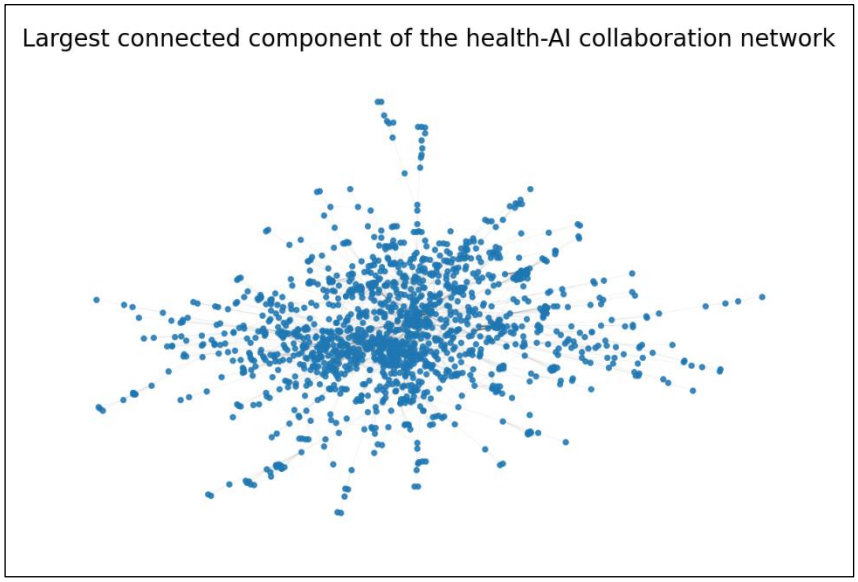

Figure 1: Largest connected component of the health-AI repository collaboration network.

Nodes represent repositories, and edges indicate shared contributors or parent–fork relationships. A force-directed layout reveals a dense central core of highly interconnected projects surrounded by many short spokes and small peripheral clusters. Isolated repositories and multi-repository components of size two or three are omitted from the visualization for clarity.

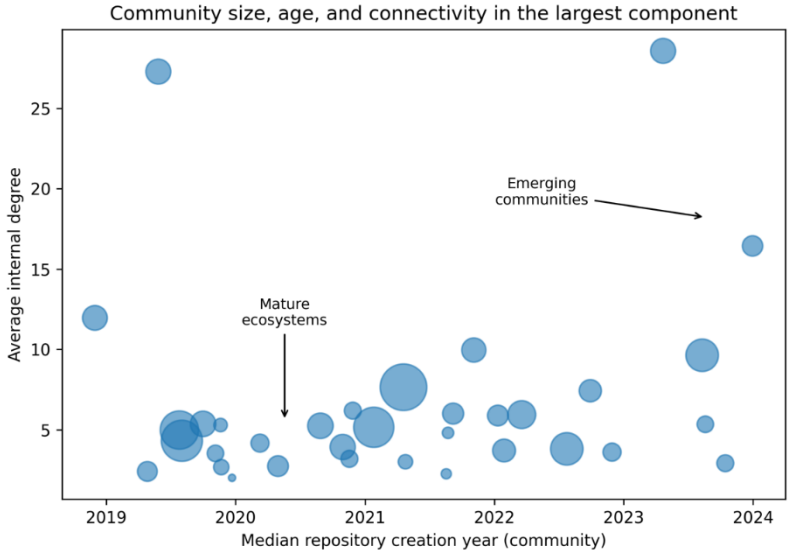

Figure 2: Communities in the largest connected component of the health-AI collaboration network.

Each bubble represents one Louvain community; the horizontal axis shows the community's median repository creation year and the vertical axis its average internal degree (mean number of collaborators per repository within the community). The bubble area is proportional to the number of repositories. Larger, older communities on the left correspond to mature core ecosystems (e.g., imaging frameworks, segmentation pipelines, and genomics toolchains), whereas smaller, more recent, and often highly connected communities on the right represent emerging areas such as LLM tooling and cloud-native platform solutions.

The most mature ecosystems found included:

a. **General-purpose imaging frameworks and toolkits.** The largest community (167 repositories) was dominated by imaging projects (≈86 % of nodes) but also included EHR and genetics repositories. High-degree nodes included widely used frameworks such as the core MONAI library, MONAI Label, TotalSegmentator, tutorials and reference pipelines, DICOM conversion utilities, and histopathology toolkits. Many of these projects are cited in the scientific literature as foundational infrastructure [2]–[4]. The community exhibited moderate density (≈0.046) and a median creation date around 2021, indicating a multi-year maturation period.
b. **Benchmarking, challenge infrastructure, and federated evaluation.** Another large community (120 repositories) centered on resources such as the GaNDLF framework, MedPerf benchmarking platform, challenge codebases for brain tumor and prostate imaging, and medical journal submission templates. The presence of MedPerf aligns with its role as a federated benchmarking hub for medical AI [5]. This community combined imaging, genetics, and EHR repositories, suggesting cross-domain interest in shared evaluation tooling.
c. **Segmentation pipelines and challenge ecosystems.** A distinct cluster (81 repositories) revolved around nnU-Net and related segmentation pipelines for neuroimaging and digital pathology. The nnU-Net repository itself had both high degree and high betweenness centrality, reflecting its use as a base for downstream segmentation challenges and derivative projects [4].
d. **Imaging datasets and COVID-19 chest X-ray resources.** Another mature community centered on curated lists and datasets for chest imaging, including large COVID-19 chest-xray collections and educational resources. This cluster showed a slightly older median creation date, corresponding to the 2020–2021 rush of pandemic-related open-data efforts.

These mature communities share several characteristics: relatively large size (≥80 nodes), multi-institutional contributor overlap, and star count that place several nodes among the most popular repositories in the dataset. They represent “infrastructure” layers of the health-AI ecosystem.

4.3. Emerging and Weakly Connected Areas

At the opposite end of the spectrum, many communities were small (≤10 nodes) and exhibited low internal density. These emerging communities correspond to the smaller, more recent bubbles with relatively high average degree on the right-hand side of Figure 2. These clusters often corresponded to:

- prototype implementations around a single dataset or institution,
- forks and thin wrappers of larger frameworks with minimal independent contributor activity,
- narrowly scoped educational repositories or course materials.

In between, a set of medium-sized but relatively young communities appeared to correspond to emerging areas:

a) **LLM-based assistants and RAG-style tooling for health data.** A 32-node community showed high internal density (≈0.53) and a median creation date in early 2024. Repositories included LLM-based assistants for biomedical literature review, retrieval-augmented generation (RAG) pipelines for clinical documents, and general-purpose agent frameworks adapted to health settings. Subdomain labels were mixed, with imaging, EHR, genetics, and general applications represented in similar proportions.

b) **Cross-subdomain foundation-model and evaluation tooling.** Another relatively dense community of 81 repositories, with a median creation date in mid-2023, brought together imaging, EHR, and genetics projects focusing on model zoos, leaderboard aggregation, and multi-modal representations. Repositories in this cluster included topic-agnostic "awesome" lists of foundation models, health-specific evaluation tools, and plug-ins that integrate LLMs with structured or imaging data.

c) **Tabular EHR frameworks and clinical benchmarking datasets.** A 25-node community with a significant EHR and "general" presence contained toolkits for modeling event-based medical records, repositories around the MEDS dataset family, and AutoML pipelines for tabular healthcare data. Many of these repositories appeared after 2022 and showed moderate internal density, suggesting the gradual consolidation of reusable EHR infrastructures.

These emerging communities differ from mature ones in both age and structure: they are newer, smaller, and often show more balanced cross-subdomain composition.

4.4. Cross-Subdomain Collaboration and Bridge Repositories

Betweenness centrality analysis highlighted repositories that act as bridges between subdomains or communities. To complement this node-level view, cross-subdomain collaboration intensity was summarized for each repository with at least one neighbor as the fraction of its collaborators whose primary subdomain differs from its own. Figure 3 shows the distribution of this ratio by primary subdomain. Imaging and genetics repositories have low median cross-subdomain ratios (medians near zero), indicating that most observed collaborations remain within subdomain boundaries, although genetics exhibits a somewhat broader tail. In contrast, EHR repositories have a median cross-subdomain ratio of 0.5, and general health-AI repositories have a median of 1.0, indicating that many of their collaborative ties connect to other subdomains. This pattern is consistent with the role of EHR and general health-AI projects as integrative layers that connect imaging, genomics, and other application areas.

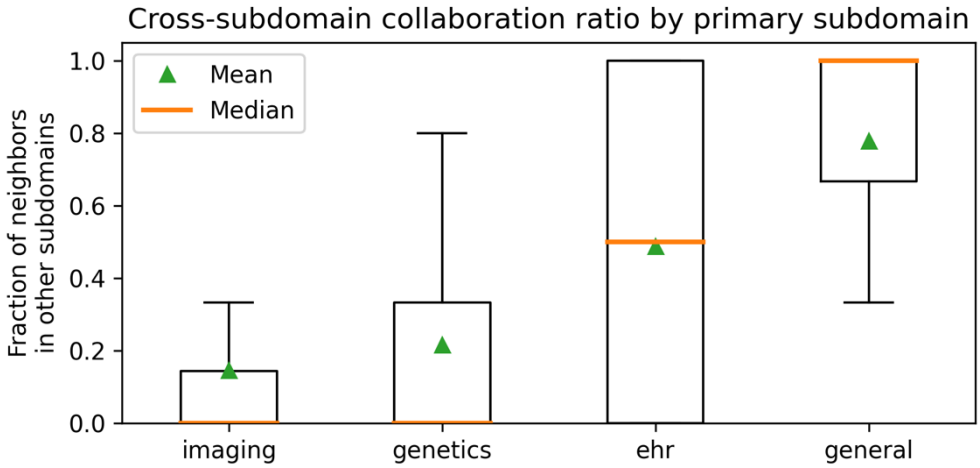

Figure 3: Cross-subdomain collaboration ratio by primary subdomain

For each repository with at least one collaborator, the plotted value is the fraction of its neighbors in the collaboration graph whose primary subdomain differs from its own. Boxplots summarize the distribution of this ratio for repositories labeled imaging, genetics, EHR, and general health AI; the green triangles mark the mean values. Imaging and genetics projects collaborate predominantly within their own subdomains, whereas EHR and general

health-AI repositories exhibit much higher levels of cross-subdomain collaboration, reflecting their integrative roles. As in Figure 3. The top central nodes included:

- Domain-spanning literature surveys and curated resource lists for biomedical machine learning,
- General-purpose medical data collections and utilities used across imaging and non-imaging tasks,
- Key framework repositories such as nnU-Net, MONAI, MONAI Label, broad "awesome-healthcare" and "awesome-bioinformatics" lists, and model explainability toolkits used in both clinical and non-clinical contexts.

High betweenness often coincided with multi-subdomain labels and broad topic coverage, suggesting that such repositories play a cross-cutting role in the health-AI ecosystem. They may serve as strategic targets for investment (e.g., maintenance, documentation, governance) because improvements propagate across many downstream users and subfields.

5. Discussion

5.1. Interpreting Mature Communities

The presence of large, dense communities around imaging frameworks and segmentation pipelines is consistent with the heavy emphasis on imaging in the health-AI literature [1],[2]. Imaging tasks align naturally with deep learning methods and have benefited from decades of public datasets and a culture of challenges. The observed communities contain many repositories that implement or extend widely cited methods, suggesting that GitHub has become the de facto coordination layer for imaging AI innovation. The benchmarking and challenge community indicates increasing attention to reproducibility, federated evaluation, and standardized metrics. MedPerf's central position illustrates how open initiatives can bridge academic, industrial, and standards-oriented stakeholders [5]. As more regulatory bodies and clinical adopters demand robust evidence of generalization, these communities are likely to grow in importance.

5.2. Emerging Trends and Opportunities

The LLM-related communities, despite their relatively small size, appear to be growing rapidly and show substantial cross-subdomain mixing. This pattern reflects the general shift in AI practice toward foundation models and shared text or code representations, as well as the proliferation of RAG architectures for clinical notes, guidelines, and biomedical literature. From a network perspective, these communities represent emerging bridges connecting previously separate ecosystems, such as genomics and EHRs. EHR and general-health communities remain smaller than imaging and genetics in absolute terms, but show competitive median star and fork counts. This suggests that even limited numbers of reusable toolkits, particularly those providing standardized representations of longitudinal health data, can have outsized impact. Strengthening these communities could yield disproportionate benefits for downstream clinical decision support and population health analytics.

5.3. Implications for Stakeholders

a. **Researchers and students.** The identified communities provide a starting point for navigating the health-AI open-source landscape. For new projects, selecting dependencies from mature, high-degree communities (e.g., established frameworks and benchmark suites) may reduce duplication of effort. Conversely, targeting emerging communities allows early involvement in rapidly evolving areas such as LLM-based tooling.
b. **Funders and maintainers.** High-betweenness repositories, particularly those that bridge subdomains, represent attractive targets for long-term support, including maintenance

grants and dedicated engineering time. Investments in documentation, testing, and governance for these projects could strengthen the entire ecosystem.

c. **Industry and startups.** Companies seeking to engage with open-source health AI can use the network to identify influential projects and active contributors. Communities centered on benchmarking and federated evaluation align directly with industry needs for robust model comparison and regulatory-grade evidence generation. Participation in these ecosystems may accelerate technology transfer.
d. **Platform operators and code-hosting services.** The prevalence of isolated repositories highlights opportunities for GitHub-like platforms to improve discovery and recommendation mechanisms, for example, by promoting repositories that connect isolated pockets of activity or by encouraging shared contributor credit for downstream forks.

6. Limitations and Future Work

Several limitations affect the present analysis:

- **Platform coverage.** Only GitHub was considered. Important health-AI projects hosted on GitLab, Bitbucket, institutional repositories, or model hubs such as Hugging Face were excluded. Incorporating multiple platforms would provide a more complete picture of the ecosystem.
- **Keyword-based sampling.** The dataset relies on manually crafted health and AI keywords, selected through an empirical calibration process applied to GitHub search results. Although this procedure reduces obvious noise and ensures manageable query sizes, some relevant projects may be missed due to atypical terminology, and some included projects may be only tangentially related to health AI. Refinement using semi-supervised classification or embeddings could improve precision.
- **License interpretation.** Public availability on GitHub does not guarantee an OSI-approved open-source license. Some repositories may be source-available but restrict certain kinds of reuse. A more nuanced future analysis could stratify results by license family and examine whether collaboration patterns differ between permissive and restrictive licenses.
- **Static snapshot.** The network reflects a snapshot aggregated across several years. Temporal analysis, tracking community formation, growth, and decay—could reveal more about the dynamics of health-AI collaboration, including how crises such as the COVID-19 pandemic alter patterns of open-source contribution.
- **Simplified contributor modeling.** All non-bot contributors were treated equally, regardless of commit volume, organizational affiliation, or role. Incorporating contribution weights or metadata about institutions could refine the identification of key maintainers and cross-organizational collaborations.

Future work could address these limitations by extending the dataset, applying more sophisticated topic modeling to label communities, and exploring longitudinal network evolution. Linking repositories to scientific publications, clinical trials, and regulatory submissions would further illuminate the path from open-source contribution to real-world impact.

7. Conclusions

This study mapped a large slice of the open-source health-AI landscape on GitHub by constructing a repository-level collaboration network and examining how projects and contributors are organized across imaging, EHR and clinical text, genomics, and more general clinical applications. The analysis showed that collaboration is highly uneven: a dense core of mature communities coexists with a long tail of isolated or weakly connected projects. Within the core, large communities centered on medical imaging frameworks, segmentation pipelines,

and benchmarking infrastructures emerged as critical hubs, with many shared contributors and strong internal connectivity. These ecosystems appear to function as de facto infrastructure layers that support a wide range of downstream work. At the same time, smaller but rapidly growing communities were identified around emerging themes such as large-language-model tooling, cross-subdomain foundation models, and cloud-native platforms for healthcare and omics. These newer clusters are structurally distinct, mixing repositories from multiple subdomains and often acting as bridges between previously separate areas of practice. The findings suggest several practical implications. For researchers and practitioners, orienting new work around established infrastructural communities can lower barriers to entry, while targeted engagement with emerging clusters offers opportunities to help shape new methodological directions. For funders, maintainers, and industry stakeholders, high-centrality repositories and cross-subdomain communities represent promising candidates for support, as improvements there are likely to propagate widely. At the same time, the prevalence of isolated and very small components highlights ongoing fragmentation and suggests that many potentially useful tools remain marginal or under-maintained. The approach taken here has clear limitations, including reliance on keyword-based sampling, focus on a single hosting platform, and relatively simple treatment of contributor roles and licensing. Nonetheless, it provides an initial, scalable framework for quantifying structure in the health-AI open-source ecosystem. Future work can extend this foundation by incorporating additional platforms and model hubs, refining topic and subdomain assignments, linking repositories to publications and clinical deployments, and analysing temporal dynamics to understand how communities emerge, stabilise, and sometimes fade over time.

Funding source

No funding was received for this study.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Teng, Z., Li, L., Xin, Z., Xiang, D., Huang, J., Zhou, H., ... & Chen, X. (2024). A literature review of artificial intelligence (AI) for medical image segmentation: from AI and explainable AI to trustworthy AI. *Quantitative imaging in medicine and surgery*, *14*(12), 9620-9652.

[2] Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., ... & Feng, A. (2022). Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.

[3] Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., ... & Cardoso, M. J. (2024). Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *Medical Image Analysis*, *95*, 103207.

[4] Isensee, F., Jäger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2019). Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*.

[5] Karargyris, A., Umeton, R., Sheller, M. J., Aristizabal, A., George, J., Bala, S., ... & Mattson, P. (2021). Medperf: open benchmarking platform for medical artificial intelligence using federated evaluation. *arXiv preprint arXiv:2110.01406*.

[6] Lima, A., Rossi, L., & Musolesi, M. (2014, May). Coding together at scale: GitHub as a collaborative social network. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 295-304).

[7] McClean, K., Greer, D., & Jurek-Loughrey, A. (2021). Social network analysis of open source software: A review and categorisation. *Information and software technology*, *130*, 106442.

[8] Zhang, G., Schuessler, J. H., & Shao, C. Y. (2025). Small-World Phenomenon of Global Open-Source Software Collaboration on Github: A Social Network Analysis. *Journal of Global Information Management (JGIM)*, *33*(1), 1-24.

[9] Muhammad, D., & Bendechache, M. (2024). Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Computational and structural biotechnology journal*, *24*, 542-560.

[10] R. Baddour, mapping-health-ai-repositories, GitHub repository, 2025. [Online]. Available: https://github.com/rbad/mapping-health-ai-repositories. Accessed: Dec. 13, 2025.

[11] NetworkX, Software for complex networks," NetworkX Project. [Online]. Available: https://networkx.org. Accessed: Dec. 1, 2025.