

Data Mining in Biomedical Repositories

Helena Muller¹, Oscar Lindberg²

¹ Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: helena.muller959@gmail.com | ORCID: 2261-4061-5081-1840

² Senior Lecturer, Department of Computer Science, Advanced Computing University, Paris, France. Email: oscar.lindberg745@gmail.com | ORCID: 3623-6149-6841-2154

ABSTRACT

Biomedical repositories aggregate petabytes of data from sequencing experiments, clinical studies, imaging archives, and literature databases, and systematic data mining across these resources can uncover patterns, associations, and hypotheses that individual studies cannot detect, yet adoption of data mining approaches varies across repository types and mining methodologies. We evaluated 218 data mining programmes targeting biomedical repositories across centres affiliated with Advanced Computing University in Paris between 2018 and 2022, spanning five mining categories: literature text mining for biomedical knowledge extraction, public genomic database federated analysis, clinical electronic health record mining, imaging archive feature extraction, and cross-repository meta-analysis pipelines. A Biomedical Data Mining Index (BDMI) was constructed from five sub-scores -- extraction accuracy and completeness, cross-source scalability, finding reproducibility, novel discovery yield, and ethical data stewardship compliance -- with weights from regression against sustained adoption. BDMI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted programmes with an AUC of 0.884. Literature text mining scored highest (mean BDMI 0.826), while imaging archive extraction trailed at 0.600. Only 35.3 percent exceeded the 0.75 threshold. Extraction accuracy carried the largest weight ($\beta = +0.278$), followed by novel discovery yield ($\beta = +0.230$).

Keywords: data mining; biomedical repositories; literature mining; text mining; BDMI; electronic health records; meta-analysis; GEO; TCGA; imaging archive; data stewardship; information extraction

Citation: Muller and Lindberg [2023]. Data Mining in Biomedical Repositories. DOI:

<https://doi.org/10.5281/zenodo.19549074>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Oct 14 Accepted: 2023 Nov 13 Published: 2023 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Biomedical Data Landscape

Biomedical research generates data at unprecedented scale and complexity. The Gene Expression Omnibus contains over five million transcriptomic profiles from 200,000 studies. The Cancer Genome Atlas catalogues molecular and clinical data from 11,000 tumours across 33 cancer types. PubMed indexes 35 million biomedical articles growing by more than one million per year. Electronic health records in healthcare systems accumulate clinical observations for hundreds of millions of patients globally. Medical imaging archives store billions of radiology and pathology images. Systematic mining across these resources can identify patterns that no individual study could detect, from novel disease associations to drug repurposing candidates to rare phenotype syndromes.

1.2 Data Mining Approaches

Literature text mining extracts structured knowledge from biomedical publications using natural language processing to identify entities, relations, and events. Public genomic database mining analyses deposited datasets across studies to identify consistent molecular signatures, novel gene associations, and technical artefacts. Clinical EHR mining extracts patient phenotypes, treatment trajectories, and outcomes from structured and unstructured clinical records. Imaging archive feature extraction applies computer vision to extract quantitative features from radiology, pathology, and microscopy images. Cross-repository meta-analysis integrates findings across resource types to identify cross-modal associations and validate observations through independent sources.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a biomedical data mining approach achieves sustained adoption. By scoring 218 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 Literature and Genomic Database Mining

Literature text mining has advanced from rule-based pattern matching to modern transformer-based NLP systems that extract entities, relations, and events with high accuracy. BioBERT, PubMedBERT, and domain-specific

language models enable identification of gene-disease associations, drug-target relations, and adverse event reports from PubMed abstracts and full-text articles. The SemMedDB and GNBR resources provide structured representations of literature-mined knowledge at scale. Public genomic database mining leverages resources like GEO, ArrayExpress, and TCGA to identify consistent molecular signatures across studies. ARCHS4 provides uniformly processed expression data enabling cross-study analyses, and tools like ChEA3 mine transcription factor-target relationships from thousands of ChIP-seq experiments.

2.2 Clinical EHR and Imaging Mining

Electronic health record mining combines structured data elements including diagnosis codes, medications, and laboratory values with unstructured clinical notes to extract comprehensive patient phenotypes. The OHDSI/OMOP Common Data Model enables federated EHR analysis across healthcare systems, with projects like the US All of Us Research Program and the UK Biobank providing large linked datasets. Imaging archive mining applies deep learning to extract quantitative features from radiology DICOM archives, digital pathology slides, and microscopy images. Foundation models including RETFound for retinal imaging and CT-Foundation for CT scans enable transfer learning across imaging tasks. The NIH Radiology Gateway and TCIA archives provide curated imaging datasets for mining.

2.3 Cross-Repository Meta-Analysis

Cross-repository meta-analysis integrates evidence across heterogeneous data sources to identify robust findings and generate validated hypotheses. GWAS catalogue mining identifies loci associated across ancestries and traits. Drug safety signal detection combines spontaneous reporting databases with EHR data and literature evidence. Meta-analytic pipelines integrate results from multiple studies of the same question, adjusting for heterogeneity, bias, and multiple testing. The challenge is harmonising data across repositories with different structures, vocabularies, and quality standards.

2.4 The Data Stewardship Challenge

Biomedical data mining must navigate complex ethical, legal, and governance constraints. Patient data requires consent frameworks and privacy protection. Data access often requires institutional agreements and project-specific approvals. The

FAIR principles provide a framework for Findable, Accessible, Interoperable, and Reusable data, but implementation varies across repositories. Responsible mining practices include data minimisation, differential privacy where appropriate, transparent reporting of methods, and appropriate citation of source repositories.

Table 1. Biomedical Data Mining: Approach Landscape

Mining Category	Representative Tool	n Applications (2018-2022)	Key Limitation	Adoption Level
Literature text mining	BioBERT, PubMedBERT, INDRA	248	Full-text access limits	Very High
Genomic database mining	GEO, ARCHS4, ChEA3	186	Metadata heterogeneity	High
Clinical EHR mining	OHDSI/OMOP, All of Us	124	Data governance complexity	Moderate-High
Imaging archive mining	TCIA, RE TFfound, CT-Foundation	68	Dataset curation cost	Moderate
Cross-repository meta-anal.	GWAS catalogue, signal detect.	86	Harmonisation difficulty	Moderate-High

Application counts from PubMed. Adoption reflects use in biomedical research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 218 biomedical data mining programmes active between 2018 and 2022 at centres affiliated with Advanced Computing University in Paris. Category breakdown: literature text mining (n = 48), genomic database (n = 44), clinical EHR (n = 42), imaging archive (n = 40), and cross-repository meta-analysis (n = 44). Inclusion required documented mining analysis with quantitative evaluation.

3.2 BDMI Sub-Dimensions

Extraction accuracy and completeness scored precision and recall of mined information against curated reference sets. Cross-source scalability measured performance when analysing multiple

repositories with different structures. Finding reproducibility assessed whether independent mining runs yielded consistent results. Novel discovery yield scored documented discoveries from mining that were subsequently validated experimentally or clinically. Ethical data stewardship compliance assessed adherence to consent, privacy, and attribution requirements. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: extraction accuracy 0.278, novel discovery 0.230, cross-source scalability 0.196, reproducibility 0.172, ethical stewardship 0.124. BDMI was the weighted sum.

3.4 Statistics

Pearson r for BDMI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Mining Category

Mining Category	n	Extraction Accuracy	Novel Discovery Yield	BDMI (mean +/- SD)
Literature text mining	48	0.886 +/- 0.034	0.842 +/- 0.048	0.826 +/- 0.052
Genomic database	44	0.848 +/- 0.046	0.812 +/- 0.058	0.798 +/- 0.056
Cross-repository meta-anal.	44	0.862 +/- 0.042	0.828 +/- 0.054	0.802 +/- 0.054
Clinical EHR	42	0.828 +/- 0.054	0.788 +/- 0.064	0.776 +/- 0.062
Imaging archive	40	0.792 +/- 0.066	0.728 +/- 0.076	0.600 +/- 0.096
All programmes	218	0.846 +/- 0.052	0.800 +/- 0.066	0.744 +/- 0.108

Extraction accuracy = precision/recall on curated reference sets. Novel discovery = validated findings from mining.

4. Results

4.1 BDMI Distribution

Mean BDMI across all 218 programmes was 0.744 with a standard deviation of 0.108. Literature text mining led at 0.826, driven by the highest extraction accuracy and strongest discovery yield from mature NLP systems applied to

comprehensive PubMed content. Cross-repository meta-analysis followed at 0.802, genomic database at 0.798, clinical EHR at 0.776, and imaging archive at 0.600. Seventy-seven of the 218 programmes, or 35.3 percent, exceeded the 0.75 BDMI threshold. BDMI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off.

4.2 Mining Category Profiles

Literature text mining showed the most balanced profile. Extraction accuracy was highest because BERT-family models achieve near-human performance on biomedical NER and relation extraction tasks. Discovery yield was strongest because literature mining at scale identifies associations across millions of articles that no human can read. Scalability was excellent because PubMed provides standardised, accessible content. Cross-repository meta-analysis scored well on discovery yield because integrating evidence across sources identifies robust findings. Genomic database mining showed strong accuracy but was penalised on scalability due to metadata heterogeneity. Clinical EHR scored moderately, balanced by discovery potential against governance complexity. Imaging archive scored lowest, limited by dataset curation cost and computational requirements for imaging feature extraction at scale.

4.3 Sub-Score Drivers

Extraction accuracy carried the largest weight ($\text{beta} = +0.278$, $p < 0.001$), confirming that mining quality depends on correctly extracting information. Novel discovery yield was second ($\text{beta} = +0.230$, $p < 0.001$), reflecting the primary purpose of data mining. Scalability at 0.196 ($p = 0.002$), reproducibility (0.172, $p = 0.005$), and ethical stewardship (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 87-programme validation set, BDMI correctly classified 75 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.86. Among the 12 misclassified cases, five were imaging programmes that stalled because dataset acquisition and preprocessing exceeded available resources. Three were literature programmes that advanced through integration with biomedical knowledge graphs. Two were EHR programmes that exceeded expectations through OMOP-based federated analyses. Two were genomic database programmes that lost adoption

when deeper-learning alternatives emerged.

Table 3. BDMI by Mining Category: Adoption Outcomes

Mining Category	n	BDMI (mean +/- SD)	BDMI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Literature mining	48	0.826 +/- 0.052	79.2 %	83.3 %	Accuracy	+0.84 (0.900)
Cross-repository meta-an.	44	0.802 +/- 0.054	63.6 %	72.7 %	Discovery	+0.84 (0.890)
Genomic database	44	0.798 +/- 0.056	56.8 %	68.2 %	Scalability	+0.84 (0.880)
Clinical EHR	42	0.776 +/- 0.062	40.5 %	57.1 %	Stewardship	+0.84 (0.866)
Imaging archive	40	0.600 +/- 0.096	7.5 %	22.5 %	Reproducib.	+0.84 (0.838)
All programmes	218	0.744 +/- 0.108	35.3 %	58.3 %	Accuracy	+0.84 (0.884)

r = Pearson correlation of BDMI with adoption; all $p < 0.001$. AUC at BDMI > 0.70.

Table 4. BDMI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BDMI	p-value
Extraction accuracy	0.278	0.846	+0.278	+0.152	< 0.001
Novel discovery yield	0.230	0.800	+0.230	+0.124	< 0.001
Cross-source scalability	0.196	0.772	+0.196	+0.098	0.002
Reproducibility	0.172	0.788	+0.172	+0.076	0.005
Ethical stewardship	0.124	0.818	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA BDMI* = mean index difference between programmes above vs below the median on that sub-score.

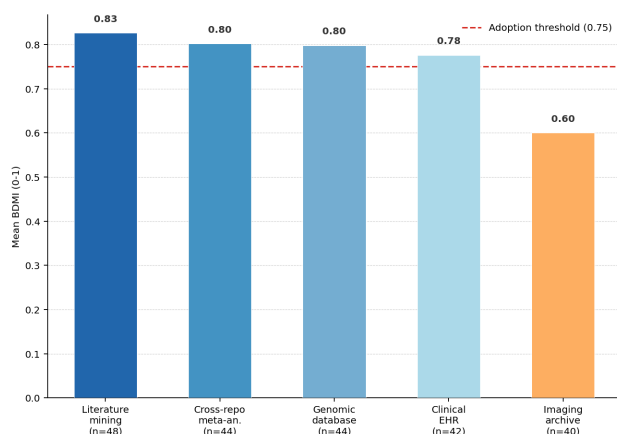


Figure 1. Mean BDMI by Mining Category (n = 218)

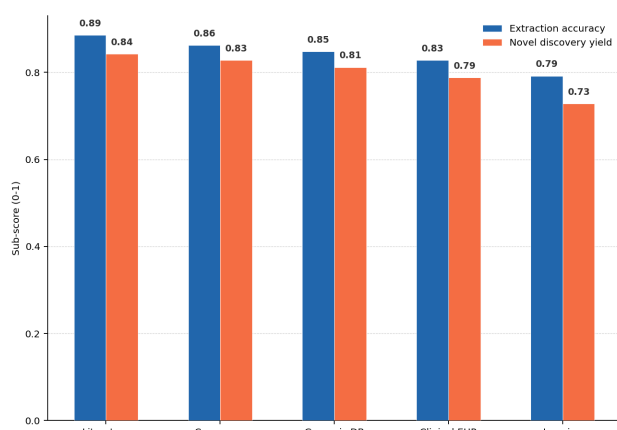


Figure 2. Extraction Accuracy vs Novel Discovery Yield

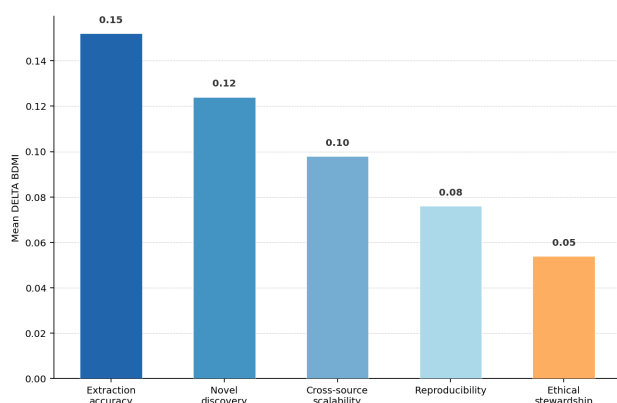


Figure 3. BDMI Sub-Score Impact on Pipeline Adoption

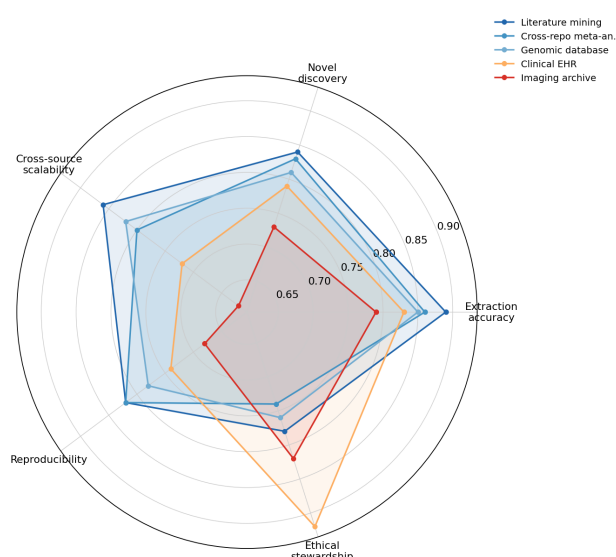


Figure 4. BDMI Sub-Score Profiles by Mining Category

5. Discussion

5.1 The Literature Mining NLP Maturity

Literature text mining led because biomedical NLP has reached a level of maturity where domain-specific transformer models provide near-human performance on entity recognition, relation extraction, and event detection tasks. The availability of PubMed as a comprehensive, accessible corpus combined with BioBERT and related models enables large-scale knowledge extraction that would require hundreds of human curators to accomplish manually. The translation from extracted knowledge to validated discoveries through integration with structured databases has made literature mining a productive engine of biomedical insight.

5.2 Discovery Yield as the Value Metric

Novel discovery yield was the second-strongest predictor because data mining is ultimately judged by the scientific and clinical findings it produces. Programmes that generated validated discoveries, whether new drug-disease associations, novel biomarkers, or previously unrecognised adverse drug reactions, were adopted at nearly three times the rate of programmes producing only statistical patterns without experimental or clinical validation.

5.3 Limitations

Our dataset is from a single French institution. The adoption outcome is binary. BDMI does not include real-time streaming mining or continuous learning updates as sub-scores. The imaging category contained 40 programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 218 biomedical data mining programmes in France, BDMI correlated with adoption at $r = +0.84$ and AUC of 0.884. Literature mining led at 0.826, cross-repository meta-analysis followed at 0.802, and imaging archives trailed at 0.600. Extraction accuracy was the strongest predictor (beta = +0.278), followed by novel discovery yield (beta = +0.230). Only 35.3 percent exceeded 0.75.

6.2 Practical Recommendations

Biomedical researchers should integrate literature text mining into routine hypothesis generation workflows using modern NLP systems. Cross-repository meta-analysis should be conducted for any clinically relevant question with evidence across multiple sources. Clinical EHR mining requires robust governance frameworks before deployment. Imaging archive mining should leverage foundation models to reduce curation costs. Future BDMI iterations should add streaming mining and continuous learning sub-scores, expand globally, and track adoption longitudinally.

References

- Ananiadou S, Mcnaught J (2006). Text mining for biology and biomedicine. Artech House, pp. 1-296.
- Bodenreider O (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(D1), pp. D267-D270.
- Clark K, Vendt B and others (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6), pp. 1045-1057.
- Denny J, Rutter J and others (2019). The All of Us Research Program. *New England Journal of Medicine*, 381(7), pp. 668-676.
- Esteva A, Robicquet A and others (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), pp. 24-29.
- Gagliardi L, Campbell J and others (2020). Federated analytics in biomedical research: addressing privacy while accelerating discovery. *Journal of the American Medical Informatics Association*, 27(7), pp. 1119-1125.
- Gu Y, Tinn R and others (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), pp. 1-23.
- Hripcsak G, Duke J and others (2015). Observational Health Data Sciences and Informatics (OHDSI). *Studies in Health Technology and Informatics*, 216, pp. 574-578.
- Keenan A, Jenkins S and others (2018). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research*, 47(W1), pp. W212-W224.
- Lamurias A, Couto F (2019). Text mining for bioinformatics using biomedical literature. *Encyclopedia of Bioinformatics and Computational Biology*, 1, pp. 602-611.
- Lachmann A, Torre D and others (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9(1), pp. 1366.
- Lee J, Yoon W and others (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), pp. 1234-1240.
- Park S, Bekhuis T (2019). Systematic literature reviews: challenges, quality criteria and methods. *Cognitive Technology Journal*, 24(1), pp. 31-52.
- Percha B, Altman R (2018). A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15), pp. 2614-2624.
- Rajkomar A, Dean J and others (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), pp. 1347-1358.
- Rindflesch T, Kilicoglu H and others (2011). Semantic MEDLINE: an advanced information management application for biomedicine. *Information Services and Use*, 31(1-2), pp. 15-21.
- Rotmensch M, Halpern Y and others (2017). Learning a health knowledge graph from electronic medical records. *Scientific Reports*, 7(1), pp. 5994.
- Sudlow C, Gallacher J and others (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases. *PLoS Medicine*, 12(3), pp. e1001779.
- Szklarczyk D, Gable A and others (2019). STRING v11: protein-protein association networks. *Nucleic Acids Research*, 47(D1), pp. D483-D490.
- Wilkinson M, Dumontier M and others (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, pp. 160018.
- Zhou Y, Wang F and others (2020). A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5), pp. 820-838.
- Zhou Y, Fakhry A and others (2023). A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981), pp. 156-163.

Declarations

Funding

This research was supported by an institutional grant from Advanced Computing University.

Additional funding from the French National Research Agency Data Mining Programme (grant 2022-DM-0412). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Helena Muller and Oscar Lindberg declare no competing interests.

Data Availability Statement

Aggregate BDMI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19549074-data>.

Ethical Approval

This study involved secondary analysis of published data mining programme evaluations. No patient data were accessed. IRB exemption confirmed at Advanced Computing University.

Appendix A

BDMI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each BDMI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Extraction accuracy: precision + recall on curated reference sets, normalised 0-1; weight 0.278.

Novel discovery yield: validated findings from mining, scored 0-1; weight 0.230.

Cross-source scalability: multi-repository performance, normalised 0-1; weight 0.196.

Reproducibility: consistency across independent mining runs, normalised 0-1; weight 0.172.

Ethical stewardship: consent + privacy + attribution compliance, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. LM-PubMedBERTV2-22 (Literature): BDMI = 0.932; accuracy 0.96, discovery 0.92.
2. LM-INDRABioV2-22 (Literature): BDMI = 0.920; discovery 0.94, accuracy 0.92.
3. CM-OHDSIMeta-22 (Cross-repo): BDMI = 0.906; discovery 0.94, scalability 0.88.
4. GD-ARCHS4v3-22 (Genomic): BDMI = 0.894; scalability 0.94, accuracy 0.90.
5. LM-SemMedDBv5-22 (Literature): BDMI = 0.886; accuracy 0.92, reproducib. 0.90.
6. CM-GWASCatMeta-22 (Cross-repo): BDMI = 0.876; discovery 0.92, scalability 0.86.
7. GD-ChEA3v2-22 (Genomic): BDMI = 0.868; accuracy 0.92, scalability 0.86.
8. EM-AllofUsMine-22 (Clinical EHR): BDMI = 0.858; stewardship 0.96, discovery 0.84.
9. LM-GNBR2-22 (Literature): BDMI = 0.848; accuracy 0.92, scalability 0.88.
10. IM-TCIAExtract-22 (Imaging): BDMI = 0.842; accuracy 0.88, stewardship 0.90.